Chainer Documentation

Release 1.23.0

Preferred Networks, inc. and Preferred Infrastructure, inc.

Apr 24, 2017

Contents

1 Install Guide 3

2 Chainer Tutorial 9

3 Chainer Reference Manual 41

4 CuPy Reference Manual 193

5 Chainer Contribution Guide 271

6 API Compatibility Policy 277

7 Tips and FAQs 283

8 Comparison with Other Frameworks 285

9 License 287

10 Indices and tables 289

Bibliography 291

Python Module Index 293

This is the Chainer documentation.

CHAPTER 1

Install Guide

Before installing Chainer

We recommend these platforms.

- Ubuntu 14.04 LTS 64bit
- CentOS 7 64bit

Chainer is supported on Python 2.7.6+, 3.4.3+, 3.5.1+. Chainer uses C++ compiler such as g++. You need to install it before installing Chainer. This is typical installation method for each platform:

```
# Ubuntu 14.04
$ apt-get install g++

# CentOS 7
$ yum install gcc-c++
```

If you use old `setuptools`, upgrade it:

```
$ pip install -U setuptools
```

Install Chainer

Chainer depends on these Python packages:

- NumPy 1.9, 1.10, 1.11
- Six 1.9

CUDA support

- CUDA 6.5, 7.0, 7.5, 8.0
- filelock

cuDNN support

- cuDNN v2, v3, v4, v5, v5.1, v6

Caffe model support

- Protocol Buffers
- protobuf>=3.0.0 is required for Py3

All these libraries are automatically installed with `pip` or `setup.py`.

Image dataset is optional

- Pillow

HDF5 serialization is optional

- h5py 2.5.0

Install Chainer via pip

We recommend to install Chainer via pip:

```
$ pip install chainer
```

Install Chainer from source

You can use `setup.py` to install Chainer from source:

```
$ tar zxf chainer-x.x.x.tar.gz
$ cd chainer-x.x.x
$ python setup.py install
```

When an error occurs...

Use `-vvvv` option with `pip` command. That shows all logs of installation. It may helps you:

```
$ pip install chainer -vvvv
```

Install Chainer with CUDA

You need to install CUDA Toolkit before installing Chainer. If you have CUDA in a default directory or set `CUDA_PATH` correctly, Chainer installer finds CUDA automatically:

```
$ pip install chainer
```

Note: Chainer installer looks up `CUDA_PATH` environment variable first. If it is empty, the installer looks for `nvcc` command from `PATH` environment variable and use its parent directory as the root directory of CUDA installation. If `nvcc` command is also not found, the installer tries to use the default directory for Ubuntu `/usr/local/cuda`.

If you installed CUDA into a non-default directory, you need to specify the directory with `CUDA_PATH` environment variable:

```
$ CUDA_PATH=/opt/nvidia/cuda pip install chainer
```

Warning: If you want to use `sudo` to install Chainer, note that `sudo` command initializes all environment variables. Please specify `CUDA_PATH` environment variable inside `sudo` like this:

```
$ sudo CUDA_PATH=/opt/nvidia/cuda pip install chainer
```

Install Chainer with CUDA and cuDNN

cuDNN is a library for Deep Neural Networks that NVIDIA provides. Chainer can use cuDNN. If you want to enable cuDNN, install cuDNN and CUDA before installing Chainer. We recommend you to install developer library of deb package of cuDNN.

If you want to install tar-gz version, we recommend you to install it to CUDA directory. For example if you uses Ubuntu Linux, copy `.h` files to `include` directory and `.so` files to `lib64` directory:

```
$ cp /path/to/cudnn.h $CUDA_PATH/include
$ cp /path/to/libcudnn.so* $CUDA_PATH/lib64
```

The destination directories depend on your environment.

If you want to use cuDNN installed in other directory, please use `CFLAGS`, `LDFLAGS` and `LD_LIBRARY_PATH` environment variables before installing Chainer:

```
export CFLAGS=-I/path/to/cudnn/include
export LDFLAGS=-L/path/to/cudnn/lib
export LD_LIBRARY_PATH=/path/to/cudnn/lib:$LD_LIBRARY_PATH
```

Install Chainer for developers

Chainer uses Cython (>=0.24). Developers need to use Cython to regenerate C++ sources from `pyx` files. We recommend to use `pip` with `-e` option for editable mode:

```
$ pip install -U cython
$ cd /path/to/chainer/source
$ pip install -e .
```

Users need not to install Cython as a distribution package of Chainer only contains generated sources.

Support image dataset

Install Pillow manually to activate image dataset. This feature is optional:

```
$ pip install pillow
```

Support HDF5 serialization

Install h5py manually to activate HDF5 serialization. This feature is optional:

```
$ pip install h5py
```

Before installing h5py, you need to install libhdf5. It depends on your environment:

```
# Ubuntu 14.04
$ apt-get install libhdf5-dev

# CentOS 7
$ yum -y install epel-release
$ yum install hdf5-devel
```

Uninstall Chainer

Use pip to uninstall Chainer:

```
$ pip uninstall chainer
```

Note: When you upgrade Chainer, `pip` sometimes installed various version of Chainer in `site-packages`. Please uninstall it repeatedly until `pip` returns an error.

Upgrade Chainer

Just use `pip` with `-U` option:

```
$ pip install -U chainer
```

Reinstall Chainer

If you want to reinstall Chainer, please uninstall Chainer and then install it. We recommend to use `--no-cache-dir` option as `pip` sometimes uses cache:

```
$ pip uninstall chainer
$ pip install chainer --no-cache-dir
```

When you install Chainer without CUDA, and after that you want to use CUDA, please reinstall Chainer. You need to reinstall Chainer when you want to upgrade CUDA.

Run Chainer with Docker

We provide the official Docker image. Use nvidia-docker command to run Chainer image with GPU. You can login to the environment with bash, and run the Python interpreter:

```
$ nvidia-docker run -it chainer/chainer /bin/bash
```

Or, run the interpreter directly:

```
$ nvidia-docker run -it chainer/chainer /usr/bin/python
```

What "recommend" means?

We tests Chainer automatically with Jenkins. All supported environments are tested in this environment. We cannot guarantee that Chainer works on other environments.

FAQ

The installer says "hdf5.h is not found"

You don't have libhdf5. Please install hdf5. See *Before installing Chainer*.

MemoryError happens

You maybe failed to install Cython. Please install it manually. See *When an error occurs...*.

Examples says "cuDNN is not enabled"

You failed to build Chainer with cuDNN. If you don't need cuDNN, ignore this message. Otherwise, retry to install Chainer with cuDNN. `-vvvv` option helps you. See *Install Chainer with CUDA and cuDNN*.

CHAPTER 2

Chainer Tutorial

Introduction to Chainer

This is the first section of the Chainer Tutorial. In this section, you will learn about the following things:

- Pros and cons of existing frameworks and why we are developing Chainer
- Simple example of forward and backward computation
- Usage of links and their gradient computation
- Construction of chains (a.k.a. "model" in most frameworks)
- Parameter optimization
- Serialization of links and optimizers

After reading this section, you will be able to:

- Compute gradients of some arithmetics
- Write a multi-layer perceptron with Chainer

Core Concept

As mentioned on the front page, Chainer is a flexible framework for neural networks. One major goal is flexibility, so it must enable us to write complex architectures simply and intuitively.

Most existing deep learning frameworks are based on the **"Define-and-Run"** scheme. That is, first a network is defined and fixed, and then the user periodically feeds it with mini-batches. Since the network is statically defined before any forward/backward computation, all the logic must be embedded into the network architecture as *data*. Consequently, defining a network architecture in such systems (e.g. Caffe) follows a declarative approach. Note that one can still produce such a static network definition using imperative languages (e.g. torch.nn, Theano-based frameworks, and TensorFlow).

In contrast, Chainer adopts a **"Define-by-Run"** scheme, i.e., the network is defined on-the-fly via the actual forward computation. More precisely, Chainer stores the history of computation instead of programming logic. This strategy

enables us to fully leverage the power of programming logic in Python. For example, Chainer does not need any magic to introduce conditionals and loops into the network definitions. The Define-by-Run scheme is the core concept of Chainer. We will show in this tutorial how to define networks dynamically.

This strategy also makes it easy to write multi-GPU parallelization, since logic comes closer to network manipulation. We will review such amenities in later sections of this tutorial.

Note: In the example code of this tutorial, we assume for simplicity that the following symbols are already imported:

```
import numpy as np
import chainer
from chainer import cuda, Function, gradient_check, report, training, utils, Variable
from chainer import datasets, iterators, optimizers, serializers
from chainer import Link, Chain, ChainList
import chainer.functions as F
import chainer.links as L
from chainer.training import extensions
```

These imports appear widely in Chainer code and examples. For simplicity, we omit these imports in this tutorial.

Forward/Backward Computation

As described above, Chainer uses the "Define-by-Run" scheme, so forward computation itself *defines* the network. In order to start forward computation, we have to set the input array to a *Variable* object. Here we start with a simple ndarray with only one element:

```
>>> x_data = np.array([5], dtype=np.float32)
>>> x = Variable(x_data)
```

A Variable object has basic arithmetic operators. In order to compute $y = x^2 - 2x + 1$, just write:

```
>>> y = x**2 - 2 * x + 1
```

The resulting y is also a Variable object, whose value can be extracted by accessing the data attribute:

```
>>> y.data
array([ 16.], dtype=float32)
```

What y holds is not only the result value. It also holds the history of computation (or computational graph), which enables us to compute its differentiation. This is done by calling its *backward()* method:

```
>>> y.backward()
```

This runs *error backpropagation* (a.k.a. *backprop* or *reverse-mode automatic differentiation*). Then, the gradient is computed and stored in the grad attribute of the input variable x:

```
>>> x.grad
array([ 8.], dtype=float32)
```

Also we can compute gradients of intermediate variables. Note that Chainer, by default, releases the gradient arrays of intermediate variables for memory efficiency. In order to preserve gradient information, pass the retain_grad argument to the backward method:

```
>>> z = 2*x
>>> y = x**2 - z + 1
>>> y.backward(retain_grad=True)
>>> z.grad
array([-1.], dtype=float32)
```

All these computations are easily generalized to a multi-element array input. Note that if we want to start backward computation from a variable holding a multi-element array, we must set the *initial error* manually. This is done simply by setting the `grad` attribute of the output variable:

```
>>> x = Variable(np.array([[1, 2, 3], [4, 5, 6]], dtype=np.float32))
>>> y = x**2 - 2*x + 1
>>> y.grad = np.ones((2, 3), dtype=np.float32)
>>> y.backward()
>>> x.grad
array([[  0.,   2.,   4.],
       [  6.,   8.,  10.]], dtype=float32)
```

Note: Many functions taking *`Variable`* object(s) are defined in the *`functions`* module. You can combine them to realize complicated functions with automatic backward computation.

Links

In order to write neural networks, we have to combine functions with *parameters* and optimize the parameters. You can use **links** to do this. A link is an object that holds parameters (i.e. optimization targets).

The most fundamental ones are links that behave like regular functions while replacing some arguments by their parameters. We will introduce higher level links, but here think of links as simply functions with parameters.

One of the most frequently used links is the `Linear` link (a.k.a. *fully-connected layer* or *affine transformation*). It represents a mathematical function $f(x) = Wx + b$, where the matrix W and the vector b are parameters. This link corresponds to its pure counterpart *`linear()`*, which accepts x, W, b as arguments. A linear link from three-dimensional space to two-dimensional space is defined by the following line:

```
>>> f = L.Linear(3, 2)
```

Note: Most functions and links only accept mini-batch input, where the first dimension of the input array is considered as the *batch dimension*. In the above Linear link case, input must have shape of (N, 3), where N is the mini-batch size.

The parameters of a link are stored as attributes. Each parameter is an instance of *`Variable`*. In the case of the Linear link, two parameters, `W` and `b`, are stored. By default, the matrix `W` is initialized randomly, while the vector `b` is initialized with zeros.

```
>>> f.W.data
array([[ 1.01847613,  0.23103087,  0.56507462],
       [ 1.29378033,  1.07823515, -0.56423163]], dtype=float32)
>>> f.b.data
array([ 0.,  0.], dtype=float32)
```

An instance of the Linear link acts like a usual function:

```
>>> x = Variable(np.array([[1, 2, 3], [4, 5, 6]], dtype=np.float32))
>>> y = f(x)
>>> y.data
array([[ 3.1757617 ,  1.75755572],
       [ 8.61950684,  7.18090773]], dtype=float32)
```

Gradients of parameters are computed by the *backward()* method. Note that gradients are **accumulated** by the method rather than overwritten. So first you must clear gradients to renew the computation. It can be done by calling the *cleargrads()* method.

```
>>> f.cleargrads()
```

Note: *cleargrads()* is introduced in v1.15 to replace *zerograds()* for efficiency. *zerograds()* is left only for backward compatibility.

Now we can compute the gradients of parameters by simply calling the backward method.

```
>>> y.grad = np.ones((2, 2), dtype=np.float32)
>>> y.backward()
>>> f.W.grad
array([[ 5.,  7.,  9.],
       [ 5.,  7.,  9.]], dtype=float32)
>>> f.b.grad
array([ 2.,  2.], dtype=float32)
```

Write a model as a chain

Most neural network architectures contain multiple links. For example, a multi-layer perceptron consists of multiple linear layers. We can write complex procedures with parameters by combining multiple links like this:

```
>>> l1 = L.Linear(4, 3)
>>> l2 = L.Linear(3, 2)
>>> def my_forward(x):
...     h = l1(x)
...     return l2(h)
```

Here the `L` indicates the *links* module. A procedure with parameters defined in this way is hard to reuse. More Pythonic way is combining the links and procedures into a class:

```
>>> class MyProc(object):
...     def __init__(self):
...         self.l1 = L.Linear(4, 3)
...         self.l2 = L.Linear(3, 2)
...
...     def forward(self, x):
...         h = self.l1(x)
...         return self.l2(h)
```

In order to make it more reusable, we want to support parameter management, CPU/GPU migration, robust and flexible save/load features, etc. These features are all supported by the *Chain* class in Chainer. Then, what we have to do here is just define the above class as a subclass of Chain:

```
>>> class MyChain(Chain):
...     def __init__(self):
```

```
...         super(MyChain, self).__init__(
...             l1=L.Linear(4, 3),
...             l2=L.Linear(3, 2),
...         )
...
...     def __call__(self, x):
...         h = self.l1(x)
...         return self.l2(h)
```

Note: We often define a single forward method of a link by `__call__` operator. Such links and chains are callable and behave like regular functions of Variables.

It shows how a complex chain is constructed by simpler links. Links like `l1` and `l2` are called *child links* of MyChain. **Note that Chain itself inherits Link**. It means we can define more complex chains that hold MyChain objects as their child links.

Another way to define a chain is using the *ChainList* class, which behaves like a list of links:

```
>>> class MyChain2(ChainList):
...     def __init__(self):
...         super(MyChain2, self).__init__(
...             L.Linear(4, 3),
...             L.Linear(3, 2),
...         )
...
...     def __call__(self, x):
...         h = self[0](x)
...         return self[1](h)
```

ChainList can conveniently use an arbitrary number of links, however if the number of links is fixed like in the above case, the Chain class is recommended as a base class.

Optimizer

In order to get good values for parameters, we have to optimize them by the *Optimizer* class. It runs a numerical optimization algorithm on a given link. Many algorithms are implemented in the `optimizers` module. Here we use the simplest one, called Stochastic Gradient Descent (SGD):

```
>>> model = MyChain()
>>> optimizer = optimizers.SGD()
>>> optimizer.use_cleargrads()
>>> optimizer.setup(model)
```

The method *use_cleargrads()* is for efficiency. See *use_cleargrads()* for detail.

The method *setup()* prepares for the optimization given a link.

Some parameter/gradient manipulations, e.g. weight decay and gradient clipping, can be done by setting *hook functions* to the optimizer. Hook functions are called after the gradient computation and right before the actual update of parameters. For example, we can set weight decay regularization by running the next line beforehand:

```
>>> optimizer.add_hook(chainer.optimizer.WeightDecay(0.0005))
```

Of course, you can write your own hook functions. It should be a function or a callable object, taking the optimizer as the argument.

There are two ways to use the optimizer. One is using it via *Trainer*, which we will see in the following sections. The other way is using it directly. We here review the latter case. *If you are interested in getting able to use the optimizer in a simple way, skip this section and go to the next one.*

There are two further ways to use the optimizer directly. One is manually computing gradients and then calling the *update()* method with no arguments. Do not forget to clear the gradients beforehand!

```
>>> x = np.random.uniform(-1, 1, (2, 4)).astype('f')
>>> model.cleargrads()
>>> # compute gradient here...
>>> loss = F.sum(model(chainer.Variable(x)))
>>> loss.backward()
>>> optimizer.update()
```

The other way is just passing a loss function to the *update()* method. In this case, *cleargrads()* is automatically called by the update method, so the user does not have to call it manually.

```
>>> def lossfun(arg1, arg2):
...     # calculate loss
...     loss = F.sum(model(arg1 - arg2))
...     return loss
>>> arg1 = np.random.uniform(-1, 1, (2, 4)).astype('f')
>>> arg2 = np.random.uniform(-1, 1, (2, 4)).astype('f')
>>> optimizer.update(lossfun, chainer.Variable(arg1), chainer.Variable(arg2))
```

See *Optimizer.update()* for the full specification.

Trainer

When we want to train neural networks, we have to run *training loops* that update the parameters many times. A typical training loop consists of the following procedures:

1. Iterations over training datasets
2. Preprocessing of extracted mini-batches
3. Forward/backward computations of the neural networks
4. Parameter updates
5. Evaluations of the current parameters on validation datasets
6. Logging and printing of the intermediate results

Chainer provides a simple yet powerful way to make it easy to write such training processes. The training loop abstraction mainly consists of two components:

- **Dataset abstraction**. It implements 1 and 2 in the above list. The core components are defined in the `dataset` module. There are also many implementations of datasets and iterators in *datasets* and *iterators* modules, respectively.
- **Trainer**. It implements 3, 4, 5, and 6 in the above list. The whole procedure is implemented by *Trainer*. The way to update parameters (3 and 4) is defined by *Updater*, which can be freely customized. 5 and 6 are implemented by instances of *Extension*, which appends an extra procedure to the training loop. Users can freely customize the training procedure by adding extensions. Users can also implement their own extensions.

We will see how to use Trainer in the example section below.

Serializer

Before proceeding to the first example, we introduce Serializer, which is the last core feature described in this page. Serializer is a simple interface to serialize or deserialize an object. *Link*, *Optimizer*, and *Trainer* supports serialization.

Concrete serializers are defined in the `serializers` module. It supports NumPy NPZ and HDF5 formats.

For example, we can serialize a link object into NPZ file by the *serializers.save_npz()* function:

```
>>> serializers.save_npz('my.model', model)
```

It saves the parameters of `model` into the file `'my.model'` in NPZ format. The saved model can be read by the *serializers.load_npz()* function:

```
>>> serializers.load_npz('my.model', model)
```

Note: Note that only the parameters and the *persistent values* are serialized by this serialization code. Other attributes are not saved automatically. You can register arrays, scalars, or any serializable objects as persistent values by the *Link.add_persistent()* method. The registered values can be accessed by attributes of the name passed to the add_persistent method.

The state of an optimizer can also be saved by the same functions:

```
>>> serializers.save_npz('my.state', optimizer)
>>> serializers.load_npz('my.state', optimizer)
```

Note: Note that serialization of optimizer only saves its internal states including number of iterations, momentum vectors of MomentumSGD, etc. It does not save the parameters and persistent values of the target link. We have to explicitly save the target link with the optimizer to resume the optimization from saved states.

Support of the HDF5 format is enabled if the h5py package is installed. Serialization and deserialization with the HDF5 format are almost identical to those with the NPZ format; just replace *save_npz()* and *load_npz()* by *save_hdf5()* and *load_hdf5()*, respectively.

Example: Multi-layer Perceptron on MNIST

Now you can solve a multiclass classification task using a multi-layer perceptron (MLP). We use a hand-written digits dataset called MNIST, which is one of the long-standing de facto "hello world" examples used in machine learning. This MNIST example is also found in the examples/mnist directory of the official repository. We show how to use *Trainer* to construct and run the training loop in this section.

We first have to prepare the MNIST dataset. The MNIST dataset consists of 70,000 greyscale images of size 28x28 (i.e. 784 pixels) and corresponding digit labels. The dataset is divided into 60,000 training images and 10,000 test images by default. We can obtain the vectorized version (i.e., a set of 784 dimensional vectors) by *datasets.get_mnist()*.

```
>>> train, test = datasets.get_mnist()
...
```

This code automatically downloads the MNIST dataset and saves the NumPy arrays to the `$(HOME)/.chainer` directory. The returned `train` and `test` can be seen as lists of image-label pairs (strictly speaking, they are instances of *TupleDataset*).

We also have to define how to iterate over these datasets. We want to shuffle the training dataset for every *epoch*, i.e. at the beginning of every sweep over the dataset. In this case, we can use *iterators.SerialIterator*.

```
>>> train_iter = iterators.SerialIterator(train, batch_size=100, shuffle=True)
```

On the other hand, we do not have to shuffle the test dataset. In this case, we can pass `shuffle=False` argument to disable the shuffling. It makes the iteration faster when the underlying dataset supports fast slicing.

```
>>> test_iter = iterators.SerialIterator(test, batch_size=100, repeat=False,␣
↪shuffle=False)
```

We also pass `repeat=False`, which means we stop iteration when all examples are visited. This option is usually required for the test/validation datasets; without this option, the iteration enters an infinite loop.

Next, we define the architecture. We use a simple three-layer rectifier network with 100 units per layer as an example.

```
>>> class MLP(Chain):
...     def __init__(self, n_units, n_out):
...         super(MLP, self).__init__(
...             # the size of the inputs to each layer will be inferred
...             l1=L.Linear(None, n_units),  # n_in -> n_units
...             l2=L.Linear(None, n_units),  # n_units -> n_units
...             l3=L.Linear(None, n_out),    # n_units -> n_out
...         )
...
...     def __call__(self, x):
...         h1 = F.relu(self.l1(x))
...         h2 = F.relu(self.l2(h1))
...         y = self.l3(h2)
...         return y
```

This link uses *relu()* as an activation function. Note that the `'l3'` link is the final linear layer whose output corresponds to scores for the ten digits.

In order to compute loss values or evaluate the accuracy of the predictions, we define a classifier chain on top of the above MLP chain:

```
>>> class Classifier(Chain):
...     def __init__(self, predictor):
...         super(Classifier, self).__init__(predictor=predictor)
...
...     def __call__(self, x, t):
...         y = self.predictor(x)
...         loss = F.softmax_cross_entropy(y, t)
...         accuracy = F.accuracy(y, t)
...         report({'loss': loss, 'accuracy': accuracy}, self)
...         return loss
```

This Classifier class computes accuracy and loss, and returns the loss value. The pair of arguments `x` and `t` corresponds to each example in the datasets (a tuple of an image and a label). *softmax_cross_entropy()* computes the loss value given prediction and ground truth labels. *accuracy()* computes the prediction accuracy. We can set an arbitrary predictor link to an instance of the classifier.

The *report()* function reports the loss and accuracy values to the trainer. For the detailed mechanism of collecting training statistics, see *Reporter*. You can also collect other types of observations like activation statistics in a similar ways.

Note that a class similar to the Classifier above is defined as *chainer.links.Classifier*. So instead of using the above example, we will use this predefined Classifier chain.

```
>>> model = L.Classifier(MLP(100, 10))  # the input size, 784, is inferred
>>> optimizer = optimizers.SGD()
>>> optimizer.setup(model)
```

Now we can build a trainer object.

```
>>> updater = training.StandardUpdater(train_iter, optimizer)
>>> trainer = training.Trainer(updater, (20, 'epoch'), out='result')
```

The second argument `(20, 'epoch')` represents the duration of training. We can use either `epoch` or `iteration` as the unit. In this case, we train the multi-layer perceptron by iterating over the training set 20 times.

In order to invoke the training loop, we just call the *`run()`* method.

```
>>> trainer.run()
```

This method executes the whole training sequence.

The above code just optimizes the parameters. In most cases, we want to see how the training proceeds, where we can use extensions inserted before calling the `run` method.

```
>>> trainer.extend(extensions.Evaluator(test_iter, model))
>>> trainer.extend(extensions.LogReport())
>>> trainer.extend(extensions.PrintReport(['epoch', 'main/accuracy', 'validation/main/
↪accuracy']))
>>> trainer.extend(extensions.ProgressBar())
>>> trainer.run()
```

These extensions perform the following tasks:

`Evaluator` Evaluates the current model on the test dataset at the end of every epoch.

`LogReport` Accumulates the reported values and emits them to the log file in the output directory.

`PrintReport` Prints the selected items in the LogReport.

`ProgressBar` Shows the progress bar.

There are many extensions implemented in the *`chainer.training.extensions`* module. The most important one that is not included above is *`snapshot()`*, which saves the snapshot of the training procedure (i.e., the Trainer object) to a file in the output directory.

The example code in the *examples/mnist* directory additionally contains GPU support, though the essential part is the same as the code in this tutorial. We will review in later sections how to use GPU(s).

Recurrent Nets and their Computational Graph

In this section, you will learn how to write

- recurrent nets with full backprop,
- recurrent nets with truncated backprop,
- evaluation of networks with few memory.

After reading this section, you will be able to:

- Handle input sequences of variable length
- Truncate upper stream of the network during forward computation

- Use volatile variables to prevent network construction

Recurrent Nets

Recurrent nets are neural networks with loops. They are often used to learn from sequential input/output. Given an input stream $x_1, x_2, \ldots, x_t, \ldots$ and the initial state h_0, a recurrent net iteratively updates its state by $h_t = f(x_t, h_{t-1})$, and at some or every point in time t, it outputs $y_t = g(h_t)$. If we expand the procedure along the time axis, it looks like a regular feed-forward network except that same parameters are repeatedly used within the network.

Here we learn how to write a simple one-layer recurrent net. The task is language modeling: given a finite sequence of words, we want to predict the next word at each position without peeking the successive words. Suppose there are 1,000 different word types, and that we use 100 dimensional real vectors to represent each word (a.k.a. word embedding).

Let's start from defining the recurrent neural net language model (RNNLM) as a chain. We can use the *chainer.links.LSTM* link that implements a fully-connected stateful LSTM layer. This link looks like an ordinary fully-connected layer. On construction, you pass the input and output size to the constructor:

```
>>> l = L.LSTM(100, 50)
```

Then, call on this instance `l(x)` executes *one step of LSTM layer*:

```
>>> l.reset_state()
>>> x = Variable(np.random.randn(10, 100).astype(np.float32))
>>> y = l(x)
```

Do not forget to reset the internal state of the LSTM layer before the forward computation! Every recurrent layer holds its internal state (i.e. the output of the previous call). At the first application of the recurrent layer, you must reset the internal state. Then, the next input can be directly fed to the LSTM instance:

```
>>> x2 = Variable(np.random.randn(10, 100).astype(np.float32))
>>> y2 = l(x2)
```

Based on this LSTM link, let's write our recurrent network as a new chain:

```
class RNN(Chain):
    def __init__(self):
        super(RNN, self).__init__(
            embed=L.EmbedID(1000, 100),  # word embedding
            mid=L.LSTM(100, 50),  # the first LSTM layer
            out=L.Linear(50, 1000),  # the feed-forward output layer
        )

    def reset_state(self):
        self.mid.reset_state()

    def __call__(self, cur_word):
        # Given the current word ID, predict the next word.
        x = self.embed(cur_word)
        h = self.mid(x)
        y = self.out(h)
        return y

rnn = RNN()
model = L.Classifier(rnn)
optimizer = optimizers.SGD()
optimizer.setup(model)
```

Here *EmbedID* is a link for word embedding. It converts input integers into corresponding fixed-dimensional embedding vectors. The last linear link `out` represents the feed-forward output layer.

The `RNN` chain implements a *one-step-forward computation*. It does not handle sequences by itself, but we can use it to process sequences by just feeding items in a sequence straight to the chain.

Suppose we have a list of word variables `x_list`. Then, we can compute loss values for the word sequence by simple `for` loop.

```
def compute_loss(x_list):
    loss = 0
    for cur_word, next_word in zip(x_list, x_list[1:]):
        loss += model(cur_word, next_word)
    return loss
```

Of course, the accumulated loss is a Variable object with the full history of computation. So we can just call its *backward()* method to compute gradients of the total loss according to the model parameters:

```
# Suppose we have a list of word variables x_list.
rnn.reset_state()
model.cleargrads()
loss = compute_loss(x_list)
loss.backward()
optimizer.update()
```

Or equivalently we can use the `compute_loss` as a loss function:

```
rnn.reset_state()
optimizer.update(compute_loss, x_list)
```

Truncate the Graph by Unchaining

Learning from very long sequences is also a typical use case of recurrent nets. Suppose the input and state sequence is too long to fit into memory. In such cases, we often truncate the backpropagation into a short time range. This technique is called *truncated backprop*. It is heuristic, and it makes the gradients biased. However, this technique works well in practice if the time range is long enough.

How to implement truncated backprop in Chainer? Chainer has a smart mechanism to achieve truncation, called **backward unchaining**. It is implemented in the *Variable.unchain_backward()* method. Backward unchaining starts from the Variable object, and it chops the computation history backwards from the variable. The chopped variables are disposed automatically (if they are not referenced explicitly from any other user object). As a result, they are no longer a part of computation history, and are not involved in backprop anymore.

Let's write an example of truncated backprop. Here we use the same network as the one used in the previous subsection. Suppose we are given a very long sequence, and we want to run backprop truncated at every 30 time steps. We can write truncated backprop using the model defined above:

```
loss = 0
count = 0
seqlen = len(x_list[1:])

rnn.reset_state()
for cur_word, next_word in zip(x_list, x_list[1:]):
    loss += model(cur_word, next_word)
    count += 1
    if count % 30 == 0 or count == seqlen:
        model.cleargrads()
```

```
        loss.backward()
        loss.unchain_backward()
        optimizer.update()
```

State is updated at `model()`, and the losses are accumulated to `loss` variable. At each 30 steps, backprop takes place at the accumulated loss. Then, the *`unchain_backward()`* method is called, which deletes the computation history backward from the accumulated loss. Note that the last state of `model` is not lost, since the RNN instance holds a reference to it.

The implementation of truncated backprop is simple, and since there is no complicated trick on it, we can generalize this method to different situations. For example, we can easily extend the above code to use different schedules between backprop timing and truncation length.

Network Evaluation without Storing the Computation History

On evaluation of recurrent nets, there is typically no need to store the computation history. While unchaining enables us to walk through unlimited length of sequences with limited memory, it is a bit of a work-around.

As an alternative, Chainer provides an evaluation mode of forward computation which does not store the computation history. This is enabled by just passing `volatile` flag to all input variables. Such variables are called *volatile variables*.

Volatile variable is created by passing `volatile='on'` at the construction:

```
x_list = [Variable(..., volatile='on') for _ in range(100)]  # list of 100 words
loss = compute_loss(x_list)
```

Note that we cannot call `loss.backward()` to compute the gradient here, since the volatile variable does not remember the computation history.

Volatile variables are also useful to evaluate feed-forward networks to reduce the memory footprint.

Variable's volatility can be changed directly by setting the `Variable.volatile` attribute. This enables us to combine a fixed feature extractor network and a trainable predictor network. For example, suppose we want to train a feed-forward network `predictor_func`, which is located on top of another fixed pre-trained network `fixed_func`. We want to train `predictor_func` without storing the computation history for `fixed_func`. This is simply done by following code snippets (suppose `x_data` and `y_data` indicate input data and label, respectively):

```
x = Variable(x_data, volatile='on')
feat = fixed_func(x)
feat.volatile = 'off'
y = predictor_func(feat)
y.backward()
```

At first, the input variable `x` is volatile, so `fixed_func` is executed in volatile mode, i.e. without memorizing the computation history. Then the intermediate variable `feat` is manually set to non-volatile, so `predictor_func` is executed in non-volatile mode, i.e., with memorizing the history of computation. Since the history of computation is only memorized between variables `feat` and `y`, the backward computation stops at the `feat` variable.

Warning: It is not allowed to mix volatile and non-volatile variables as arguments to same function. If you want to create a variable that behaves like a non-volatile variable while can be mixed with volatile ones, use `'auto'` flag instead of `'off'` flag.

Making it with Trainer

The above codes are written with plain Function/Variable APIs. When we write a training loop, it is better to use Trainer, since we can then easily add functionalities by extensions.

Before implementing it on Trainer, let's clarify the training settings. We here use Penn Tree Bank dataset as a set of sentences. Each sentence is represented as a word sequence. We concatenate all sentences into one long word sequence, in which each sentence is separated by a special word <eos>, which stands for "End of Sequence". This dataset is easily obtained by *chainer.datasets.get_ptb_words()*. This function returns train, validation, and test dataset, each of which is represented as a long array of integers. Each integer represents a word ID.

Our task is to learn a recurrent neural net language model from the long word sequence. We use words in different locations to form mini-batches. It means we maintain B indices pointing to different locations in the sequence, read from these indices at each iteration, and increment all indices after the read. Of course, when one index reaches the end of the whole sequence, we turn the index back to 0.

In order to implement this training procedure, we have to customize the following components of Trainer:

- Iterator. Built-in iterators do not support reading from different locations and aggregating them into a mini-batch.
- Update function. The default update function does not support truncated BPTT.

When we write a dataset iterator dedicated to the dataset, the dataset implementation can be arbitrary; even the interface is not fixed. On the other hand, the iterator must support the *Iterator* interface. The important methods and attributes to implement are `batch_size`, `epoch`, `epoch_detail`, `is_new_epoch`, `iteration`, `__next__`, and `serialize`. Following is a code from the official example in the `examples/ptb` directory.

```
from __future__ import division

class ParallelSequentialIterator(chainer.dataset.Iterator):
    def __init__(self, dataset, batch_size, repeat=True):
        self.dataset = dataset
        self.batch_size = batch_size
        self.epoch = 0
        self.is_new_epoch = False
        self.repeat = repeat
        self.offsets = [i * len(dataset) // batch_size for i in range(batch_size)]
        self.iteration = 0

    def __next__(self):
        length = len(self.dataset)
        if not self.repeat and self.iteration * self.batch_size >= length:
            raise StopIteration
        cur_words = self.get_words()
        self.iteration += 1
        next_words = self.get_words()

        epoch = self.iteration * self.batch_size // length
        self.is_new_epoch = self.epoch < epoch
        if self.is_new_epoch:
            self.epoch = epoch

        return list(zip(cur_words, next_words))

    @property
    def epoch_detail(self):
        return self.iteration * self.batch_size / len(self.dataset)

    def get_words(self):
```

```
        return [self.dataset[(offset + self.iteration) % len(self.dataset)]
                for offset in self.offsets]

    def serialize(self, serializer):
        self.iteration = serializer('iteration', self.iteration)
        self.epoch = serializer('epoch', self.epoch)

train_iter = ParallelSequentialIterator(train, 20)
val_iter = ParallelSequentialIterator(val, 1, repeat=False)
```

Although the code is slightly long, the idea is simple. First, this iterator creates `offsets` pointing to positions equally spaced within the whole sequence. The i-th examples of mini-batches refer the sequence with the i-th offset. The iterator returns a list of tuples of the current words and the next words. Each mini-batch is converted to a tuple of integer arrays by the `concat_examples` function in the standard updater (see the previous tutorial).

Backprop Through Time is implemented as follows.

```
def update_bptt(updater):
    loss = 0
    for i in range(35):
        batch = train_iter.__next__()
        x, t = chainer.dataset.concat_examples(batch)
        loss += model(chainer.Variable(x), chainer.Variable(t))

    model.cleargrads()
    loss.backward()
    loss.unchain_backward()  # truncate
    optimizer.update()

updater = training.StandardUpdater(train_iter, optimizer, update_bptt)
```

In this case, we update the parameters on every 35 consecutive words. The call of `unchain_backward` cuts the history of computation accumulated to the LSTM links. The rest of the code for setting up Trainer is almost same as one given in the previous tutorial.

In this section we have demonstrated how to write recurrent nets in Chainer and some fundamental techniques to manage the history of computation (a.k.a. computational graph). The example in the `examples/ptb` directory implements truncated backprop learning of a LSTM language model from the Penn Treebank corpus. In the next section, we will review how to use GPU(s) in Chainer.

Using GPU(s) in Chainer

In this section, you will learn about the following things:

- Relationship between Chainer and CuPy
- Basics of CuPy
- Single-GPU usage of Chainer
- Multi-GPU usage of model-parallel computing
- Multi-GPU usage of data-parallel computing

After reading this section, you will be able to:

- Use Chainer on a CUDA-enabled GPU

- Write model-parallel computing in Chainer
- Write data-parallel computing in Chainer

Relationship between Chainer and CuPy

Note: As of the release of v1.3.0, Chainer changes its GPU backend from PyCUDA to CuPy. CuPy covers all features of PyCUDA used by Chainer, though their interfaces are not compatible.

Chainer uses *CuPy* as its backend for GPU computation. In particular, the *cupy.ndarray* class is the GPU array implementation for Chainer. CuPy supports a subset of features of NumPy with a compatible interface. It enables us to write a common code for CPU and GPU. It also supports PyCUDA-like user-defined kernel generation, which enables us to write fast implementations dedicated to GPU.

Note: The *chainer.cuda* module imports many important symbols from CuPy. For example, the cupy namespace is referred as `cuda.cupy` in the Chainer code. Note that the *chainer.cuda* module can be imported even if CUDA is not installed.

Chainer uses a memory pool for GPU memory allocation. As shown in the previous sections, Chainer constructs and destructs many arrays during learning and evaluating iterations. It is not well suited for CUDA architecture, since memory allocation and release in CUDA (i.e. `cudaMalloc` and `cudaFree` functions) synchronize CPU and GPU computations, which hurts performance. In order to avoid memory allocation and deallocation during the computation, Chainer uses CuPy's memory pool as the standard memory allocator. Chainer changes the default allocator of CuPy to the memory pool, so user can use functions of CuPy directly without dealing with the memory allocator.

Basics of `cupy.ndarray`

Note: CuPy does not require explicit initialization, so `cuda.init()` function is deprecated.

CuPy is a GPU array backend that implements a subset of NumPy interface. The *cupy.ndarray* class is in its core, which is a compatible GPU alternative of `numpy.ndarray`. CuPy implements many functions on *cupy.ndarray* objects. *See the reference for the supported subset of NumPy API*. Understanding NumPy might help utilizing most features of CuPy. See the NumPy documentation for learning it.

The main difference of *cupy.ndarray* from `numpy.ndarray` is that the content is allocated on the device memory. The allocation takes place on the current device by default. The current device can be changed by *cupy.cuda.Device* object as follows:

```
with cupy.cuda.Device(1):
    x_on_gpu1 = cupy.array([1, 2, 3, 4, 5])
```

Most operations of CuPy is done on the current device. Be careful that it causes an error to process an array on a non-current device.

Chainer provides some convenient functions to automatically switch and choose the device. For example, the *chainer.cuda.to_gpu()* function copies a `numpy.ndarray` object to a specified device:

```
x_cpu = np.ones((5, 4, 3), dtype=np.float32)
x_gpu = cuda.to_gpu(x_cpu, device=1)
```

It is equivalent to the following code using CuPy:

```
x_cpu = np.ones((5, 4, 3), dtype=np.float32)
with cupy.cuda.Device(1):
    x_gpu = cupy.array(x_cpu)
```

Moving a device array to the host can be done by *chainer.cuda.to_cpu()* as follows:

```
x_cpu = cuda.to_cpu(x_gpu)
```

It is equivalent to the following code using CuPy:

```
with x_gpu.device:
    x_cpu = x_gpu.get()
```

Note: The *with* statements in these codes are required to select the appropriate CUDA device. If user uses only one device, these device switching is not needed. *chainer.cuda.to_cpu()* and *chainer.cuda.to_gpu()* functions automatically switch the current device correctly.

Chainer also provides a convenient function *chainer.cuda.get_device()* to select a device. It accepts an integer, CuPy array, NumPy array, or None (indicating the current device), and returns an appropriate device object. If the argument is a NumPy array, then *a dummy device object* is returned. The dummy device object supports *with* statements like above which does nothing. Here are some examples:

```
cuda.get_device(1).use()
x_gpu1 = cupy.empty((4, 3), dtype='f')  # 'f' indicates float32

with cuda.get_device(1):
    x_gpu1 = cuda.empty((4, 3), dtype='f')

with cuda.get_device(x_gpu1):
    y_gpu1 = x_gpu + 1
```

Since it accepts NumPy arrays, we can write a function that accepts both NumPy and CuPy arrays with correct device switching:

```
def add1(x):
    with cuda.get_device(x):
        return x + 1
```

The compatibility of CuPy with NumPy enables us to write CPU/GPU generic code. It can be made easy by the *chainer.cuda.get_array_module()* function. This function returns the `numpy` or *cupy* module based on arguments. A CPU/GPU generic function is defined using it like follows:

```
# Stable implementation of log(1 + exp(x))
def softplus(x):
    xp = cuda.get_array_module(x)
    return xp.maximum(0, x) + xp.log1p(xp.exp(-abs(x)))
```

Run Neural Networks on a Single GPU

Single-GPU usage is very simple. What you have to do is transferring *Link* and input arrays to the GPU beforehand. In this subsection, the code is based on *our first MNIST example in this tutorial*.

A *Link* object can be transferred to the specified GPU using the *to_gpu()* method.

This time, we make the number of input, hidden, and output units configurable. The *to_gpu()* method also accepts a device ID like `model.to_gpu(0)`. In this case, the link object is transferred to the appropriate GPU device. The current device is used by default.

If we use *chainer.training.Trainer*, what we have to do is just let the updater know the device ID to send each mini-batch.

```
updater = training.StandardUpdater(train_iter, optimizer, device=0)
trainer = training.Trainer(updater, (20, 'epoch'), out='result')
```

We also have to specify the device ID for an evaluator extension as well.

```
trainer.extend(extensions.Evaluator(test_iter, model, device=0))
```

When we write down the training loop by hand, we have to transfer each mini-batch to the GPU manually:

```
model.to_gpu()
batchsize = 100
datasize = len(x_train)
for epoch in range(20):
    print('epoch %d' % epoch)
    indexes = np.random.permutation(datasize)
    for i in range(0, datasize, batchsize):
        x = Variable(cuda.to_gpu(x_train[indexes[i : i + batchsize]]))
        t = Variable(cuda.to_gpu(y_train[indexes[i : i + batchsize]]))
        optimizer.update(model, x, t)
```

Model-parallel Computation on Multiple GPUs

Parallelization of machine learning is roughly classified into two types called "model-parallel" and "data-parallel". Model-parallel means parallelizations of the computations inside the model. In contrast, data-parallel means parallelizations using data sharding. In this subsection, we show how to use the model-parallel approach on multiple GPUs in Chainer.

Recall the MNIST example. Now suppose that we want to modify this example by expanding the network to 6 layers with 2000 units each using two GPUs. In order to make multi-GPU computation efficient, we only make the two GPUs communicate at the third and sixth layer. The overall architecture looks like the following diagram:

```
(GPU0) input --+--> l1 --> l2 --> l3 --+--> l4 --> l5 --> l6 --+--> output
               |                       |                       |
(GPU1)         +--> l1 --> l2 --> l3 --+--> l4 --> l5 --> l6 --+
```

We can use the above MLP chain as following diagram:

```
(GPU0) input --+--> mlp1 --+--> mlp2 --+--> output
               |           |           |
(GPU1)         +--> mlp1 --+--> mlp2 --+
```

Let's write a link for the whole network.

```
class ParallelMLP(Chain):
    def __init__(self):
        super(ParallelMLP, self).__init__(
            # the input size, 784, is inferred
            mlp1_gpu0=MLP(1000, 2000).to_gpu(0),
            mlp1_gpu1=MLP(1000, 2000).to_gpu(1),
```

```
            # the input size, 2000, is inferred
            mlp2_gpu0=MLP(1000, 10).to_gpu(0),
            mlp2_gpu1=MLP(1000, 10).to_gpu(1),
        )

    def __call__(self, x):
        # assume x is on GPU 0
        z0 = self.mlp1_gpu0(x)
        z1 = self.mlp1_gpu1(F.copy(x, 1))

        # sync
        h0 = F.relu(z0 + F.copy(z1, 0))
        h1 = F.relu(z1 + F.copy(z0, 1))

        y0 = self.mlp2_gpu0(h0)
        y1 = self.mlp2_gpu1(h1)

        # sync
        y = y0 + F.copy(y1, 0)
        return y  # output is on GPU0
```

Recall that the *Link.to_gpu()* method returns the link itself. The *copy()* function copies an input variable to specified GPU device and returns a new variable on the device. The copy supports backprop, which just reversely transfers an output gradient to the input device.

Note: Above code is not parallelized on CPU, but is parallelized on GPU. This is because all the functions in the above code run asynchronously to the host CPU.

An almost identical example code can be found at examples/mnist/train_mnist_model_parallel.py.

Data-parallel Computation on Multiple GPUs with Trainer

Data-parallel computation is another strategy to parallelize online processing. In the context of neural networks, it means that a different device does computation on a different subset of the input data. In this subsection, we review the way to achieve data-parallel learning on two GPUs.

Suppose again our task is *the MNIST example*. This time we want to directly parallelize the three-layer network. The most simple form of data-parallelization is parallelizing the gradient computation for a distinct set of data. First, define a model and optimizer instances:

```
model = L.Classifier(MLP(1000, 10))  # the input size, 784, is inferred
optimizer = optimizers.SGD()
optimizer.setup(model)
```

Recall that the `MLP` link implements the multi-layer perceptron, and the *Classifier* link wraps it to provide a classifier interface. We used *StandardUpdater* in the previous example. In order to enable data-parallel computation with multiple GPUs, we only have to replace it with *ParallelUpdater*.

```
updater = training.ParallelUpdater(train_iter, optimizer,
                                   devices={'main': 0, 'second': 1})
```

The `devices` option specifies which devices to use in data-parallel learning. The device with name `'main'` is used as the main device. The original model is sent to this device, so the optimization runs on the main device. In the above example, the model is also cloned and sent to GPU 1. Half of each mini-batch is fed to this cloned model. After every

backward computation, the gradient is accumulated into the main device, the parameter update runs on it, and then the updated parameters are sent to GPU 1 again.

See also the example code in examples/mnist/train_mnist_data_parallel.py.

Data-parallel Computation on Multiple GPUs without Trainer

We here introduce a way to write data-parallel computation without the help of *Trainer*. Most users can skip this section. If you are interested in how to write a data-parallel computation by yourself, this section should be informative. It is also helpful to, e.g., customize the *ParallelUpdater* class.

We again start from the MNIST example. At this time, we use a suffix like `_0` and `_1` to distinguish objects on each device. First, we define a model.

```
model_0 = L.Classifier(MLP(1000, 10))  # the input size, 784, is inferred
```

We want to make two copies of this instance on different GPUs. The *Link.to_gpu()* method runs in place, so we cannot use it to make a copy. In order to make a copy, we can use *Link.copy()* method.

```
import copy
model_1 = copy.deepcopy(model_0)
model_0.to_gpu(0)
model_1.to_gpu(1)
```

The *Link.copy()* method copies the link into another instance. *It just copies the link hierarchy*, and does not copy the arrays it holds.

Then, set up an optimizer:

```
optimizer = optimizers.SGD()
optimizer.setup(model_0)
```

Here we use the first copy of the model as *the master model*. Before its update, gradients of `model_1` must be aggregated to those of `model_0`.

Then, we can write a data-parallel learning loop as follows:

```
batchsize = 100
datasize = len(x_train)
for epoch in range(20):
    print('epoch %d' % epoch)
    indexes = np.random.permutation(datasize)
    for i in range(0, datasize, batchsize):
        x_batch = x_train[indexes[i : i + batchsize]]
        y_batch = y_train[indexes[i : i + batchsize]]

        x0 = Variable(cuda.to_gpu(x_batch[:batchsize//2], 0))
        t0 = Variable(cuda.to_gpu(y_batch[:batchsize//2], 0))
        x1 = Variable(cuda.to_gpu(x_batch[batchsize//2:], 1))
        t1 = Variable(cuda.to_gpu(y_batch[batchsize//2:], 1))

        loss_0 = model_0(x0, t0)
        loss_1 = model_1(x1, t1)

        model_0.cleargrads()
        model_1.cleargrads()

        loss_0.backward()
```

```
        loss_1.backward()

        model_0.addgrads(model_1)
        optimizer.update()

        model_1.copyparams(model_0)
```

Do not forget to clear the gradients of both model copies! One half of the mini-batch is forwarded to GPU 0, the other half to GPU 1. Then the gradients are accumulated by the *Link.addgrads()* method. This method adds the gradients of a given link to those of the self. After the gradients are prepared, we can update the optimizer in usual way. Note that the update only modifies the parameters of `model_0`. So we must manually copy them to `model_1` using *Link.copyparams()* method.

Note: If the batch size used in one model remain the same, the scale of the gradient is roughly proportional to the number of models, when we aggregate gradients from all models by *chainer.Link.addgrads()*. So you need to adjust the batch size and/or learning rate of the optimizer accordingly.

Now you can use Chainer with GPUs. All examples in the `examples` directory support GPU computation, so please refer to them if you want to know more practices on using GPUs. In the next section, we will show how to define a differentiable (i.e. *backpropable*) function on Variable objects. We will also show there how to write a simple (elementwise) CUDA kernel using Chainer's CUDA utilities.

Define your own function

In this section, you will learn about the following things:

- How to define a function on variables
- Useful tools to write a function using a GPU
- How to test the function definition

After reading this section, you will be able to:

- Write your own functions
- Define simple kernels in the function definition

Differentiable Functions

Chainer provides a collection of functions in the *functions* module. It covers typical use cases in deep learning, so many existing works can be implemented with them. On the other hand, deep learning is evolving rapidly and we cannot cover all possible functions to define unseen architectures. So it is important to learn how to define your own functions.

First, suppose we want to define an elementwise function $f(x, y, z) = x * y + z$. While it is possible to implement this equation using a combination of the $*$ and $+$ functions, defining it as a single function may reduce memory consumption, so it is not *only* a toy example. Here we call this function *MulAdd*.

Let's start with defining MulAdd working on the CPU. Any function must inherit the *Function* class. The skeleton of a function looks like:

```
class MulAdd(Function):
    def forward_cpu(self, inputs):
        # do forward computation on CPU
        return some_tuple

    def backward_cpu(self, inputs, grad_outputs):
        # do backward computation on CPU
        return some_tuple
```

We must implement *forward_cpu()* and *backward_cpu()* methods. The non-self arguments of these functions are tuples of array(s), and these functions must return a tuple of array(s).

Warning: Be careful to return a tuple of arrays even if you have just one array to return.

MulAdd is simple and implemented as follows

```
class MulAdd(Function):
    def forward_cpu(self, inputs):
        x, y, z = inputs
        w = x * y + z
        return w,

    def backward_cpu(self, inputs, grad_outputs):
        x, y, z = inputs
        gw, = grad_outputs

        gx = y * gw
        gy = x * gw
        gz = gw
        return gx, gy, gz
```

As per the warning above, the `forward_cpu` method returns a tuple of single element. Note that all arrays appearing in CPU functions are `numpy.ndarray`. The forward function is straightforward: It unpacks the input tuple, computes the output, and packs it into a tuple. The backward function is a bit more complicated. Recall the rule of differentiation of multiplication. This example just implements the rule. Look at the return values, the function just packs the gradient of each input in same order and returns them.

By just defining the core computation of forward and backward, Function class provides a chaining logic on it (i.e. storing the history of computation, etc.).

Note: Assuming we implement a (forward) function $y = f(x)$ which takes as input the vector $x \in \mathbb{R}^n$ and produces as output a vector $y \in \mathbb{R}^m$. Then the `backward` method has to compute

$$\lambda_i = \sum_{j=1}^{m} \frac{\partial y_j}{\partial x_i} \gamma_j \text{ for } i = 1 \dots n$$

where γ is the `grad_outputs`. Note, that the resulting vector λ must have the same shape as the arguments of the `forward` method.

Now let's define the corresponding GPU methods. You can easily predict that the methods we have to write are named *forward_gpu()* and *backward_gpu()*:

```
class MulAdd(Function):
    def forward_cpu(self, inputs):
```

```
        ...

    def backward_cpu(self, inputs, grad_outputs):
        ...

    def forward_gpu(self, inputs):
        x, y, z = inputs
        w = x * y + z
        return w,

    def backward_gpu(self, inputs, grad_outputs):
        x, y, z = inputs
        gw, = grad_outputs

        gx = y * gw
        gy = x * gw
        gz = gw
        return gx, gy, gz
```

In GPU methods, arrays are of type *cupy.ndarray*. We use arithmetic operators defined for this class. These operators implement the basic elementwise arithmetics.

You may find that the definitions of GPU methods are exactly same as those of CPU methods. In that case, we can reduce them to *forward()* and *backward()* methods

```
class MulAdd(Function):
    def forward(self, inputs):
        x, y, z = inputs
        w = x * y + z
        return w,

    def backward(self, inputs, grad_outputs):
        x, y, z = inputs
        gw, = grad_outputs

        gx = y * gw
        gy = x * gw
        gz = gw
        return gx, gy, gz
```

Since the *cupy.ndarray* class implements many methods of numpy.ndarray, we can write these unified methods in most cases.

The MulAdd function is used as follows:

```
x = Variable(np.random.uniform(-1, 1, (3, 2)).astype(np.float32))
y = Variable(np.random.uniform(-1, 1, (3, 2)).astype(np.float32))
z = Variable(np.random.uniform(-1, 1, (3, 2)).astype(np.float32))
w = MulAdd()(x, y, z)
```

It looks a bit ugly: we have to explicitly instantiate MulAdd before applying it to variables. We also have to be careful that one instance of MulAdd must not be used multiple times, since it acts as a node in the computational graph. In Chainer, we often define a thin wrapper Python function that hide the instantiation:

```
def muladd(x, y, z):
    return MulAdd()(x, y, z)

w = muladd(x, y, z)
```

Unified forward/backward methods with NumPy/CuPy functions

CuPy also implements many functions that are compatible to those of NumPy. We can write unified forward/backward methods with them. Consider that we want to write a backprop-able function $f(x, y) = \exp(x) + \exp(y)$. We name it *ExpAdd* here. It can be written straight-forward as follows

```
class ExpAdd(Function):
    def forward_cpu(self, inputs):
        x, y = inputs
        z = np.exp(x) + np.exp(y)
        return z,

    def backward_cpu(self, inputs, grad_outputs):
        x, y = inputs
        gz, = grad_outputs

        gx = gz * np.exp(x)
        gy = gz * np.exp(y)
        return gx, gy

    def forward_gpu(self, inputs):
        cupy = cuda.cupy
        x, y = inputs
        z = cupy.exp(x) + cupy.exp(y)
        return z,

    def backward_gpu(self, inputs, grad_outputs):
        cupy = cuda.cupy
        x, y = inputs
        gz, = grad_outputs

        gx = gz * cupy.exp(x)
        gy = gz * cupy.exp(y)
        return gx, gy

def expadd(x, y):
    return ExpAdd()(x, y)
```

Note: Here we used `cuda.cupy` instead of directly accessing *`cupy`*. This is because the `cupy` module cannot be imported if the CUDA is not installed. In order to keep the implementation valid in non-CUDA environment, we have to defer the access to the `cupy` module. Note that the *`chainer.cuda`* module can be imported even if the CUDA is not installed. Of course, the module in such environment is almost useless, but if the interpreter does not run through the code accessing CUDA-dedicated functions, the code is still valid.

The CPU and GPU implementations are almost same, except that `numpy` is replaced by *`cupy`* in GPU methods. We can unify these functions using the *`cuda.get_array_module()`* function. This function accepts arbitrary number of arrays, and returns an appropriate module for them. See the following code

```
class ExpAdd(Function):
    def forward(self, inputs):
        xp = cuda.get_array_module(*inputs)
        x, y = inputs
        z = xp.exp(x) + xp.exp(y)
        return z,

    def backward(self, inputs, grad_outputs):
```

```
        xp = cuda.get_array_module(*inputs)
        x, y = inputs
        gz, = grad_outputs

        gx = gz * xp.exp(x)
        gy = gz * xp.exp(y)
        return gx, gy

def expadd(x, y):
    return ExpAdd()(x, y)
```

Note that this code works correctly even if CUDA is not installed in the environment. If CUDA is not found, get_array_module function always returns `numpy`. We often use the name `xp` for the variadic module name, which is analogous to the abbreviation `np` for NumPy and `cp` for CuPy.

Write an Elementwise Kernel Function

Let's turn back to the MulAdd example.

The GPU implementation of MulAdd as shown above is already fast and parallelized on GPU cores. However, it invokes two kernels during each of forward and backward computations. It might hurt performance, since the intermediate temporary arrays are read and written by possibly different GPU cores, which consumes much bandwidth. We can reduce the number of invocations by defining our own kernel. It also reduce the memory consumption.

Most functions only require elementwise operations like MulAdd. CuPy provides a useful tool to define elementwise kernels, the `cupy.elementwise.ElementwiseKernel` class, and Chainer wraps it by *`cuda.elementwise()`* function. Our MulAdd implementation can be improved as follows:

```
class MulAdd(Function):
    def forward_cpu(self, inputs):
        ...

    def backward_cpu(self, inputs, grad_outputs):
        ...

    def forward_gpu(self, inputs):
        cupy = cuda.cupy
        x, y, z = inputs
        w = cuda.elementwise(
            'float32 x, float32 y, float32 z',
            'float32 w',
            'w = x * y + z',
            'muladd_fwd')(x, y, z)
        return w,

    def backward_gpu(self, inputs, grad_outputs):
        x, y, z = inputs
        gw, = grad_outputs

        gx, gy = cuda.elementwise(
            'float32 x, float32 y, float32 gw',
            'float32 gx, float32 gy',
            '''
               gx = y * gw;
               gy = x * gw;
            ''',
            'muladd_bwd')(x, y, gw)
```

```
        gz = gw
        return gx, gy, gz
```

cuda.elementwise() function accepts the essential implementation of the kernel function, and returns a kernel invocation function (actually, it returns `ElementwiseKernel` object, which is callable). In typical usage, we pass four arguments to this function as follows:

1. Input argument list. This is a comma-separated string each entry of which consists of a type specification and an argument name.
2. Output argument list in the same format as the input argument list.
3. Body of *parallel loop*. We can use the input/output argument names as an element of these arrays.
4. Name of the kernel function, which is shown in debuggers and profilers.

Above code is not compiled on every forward/backward computation thanks to two caching mechanisms provided by *cuda.elementwise()*.

The first one is *binary caching*: *cuda.elementwise()* function caches the compiled binary in the `$(HOME)/.cupy/kernel_cache` directory with a hash value of the CUDA code, and reuses it if the given code matches the hash value. This caching mechanism is actually implemented in CuPy.

The second one is *upload caching*: Given a compiled binary code, we have to upload it to the current GPU in order to execute it. *cuda.elementwise()* function memoizes the arguments and the current device, and if it is called with the same arguments for the same device, it reuses the previously uploaded kernel code.

The above MulAdd code only works for float32 arrays. The `ElementwiseKernel` also supports the type-variadic kernel definition. In order to define variadic kernel functions, you can use *type placeholder* by placing a single character as type specifier:

```
class MulAdd(Function):
    def forward_cpu(self, inputs):
        ...

    def backward_cpu(self, inputs, grad_outputs):
        ...

    def forward_gpu(self, inputs):
        cupy = cuda.cupy
        x, y, z = inputs
        w = cuda.elementwise(
            'T x, T y, T z',
            'T w',
            'w = x * y + z',
            'muladd_fwd')(x, y, z)
        return w,

    def backward_gpu(self, inputs, grad_outputs):
        x, y, z = inputs
        gw, = grad_outputs

        gx, gy = cuda.elementwise(
            'T x, T y, T gw',
            'T gx, T gy',
            '''
               gx = y * gw;
               gy = x * gw;
            ''',
```

```
            'muladd_bwd')(x, y, gw)

        gz = gw
        return gx, gy, gz
```

The type placeholder `T` indicates an arbitrary data type that CuPy supports.

There are more functionalities on user-defined kernels in CuPy. *See the CuPy documentation on user-defined kernels for more details.*

Links that wrap functions

Some functions are meant to be combined with parameters. In such case, it is useful to write a small **link** that wraps the function. We have already seen how to define a chain that wraps other links (by inheriting *Chain* class). Here we study how to define a link that does not hold any other links.

As the first example, suppose that we want to implement elementwise product function between the input array and the parameter array. It can be defined as follows:

```
class EltwiseParamProduct(Link):
    def __init__(self, shape):
        # By passing a shape of the parameter, the initializer allocates a
        # parameter variable of the shape.
        super(EltwiseParamProduct, self).__init__(W=shape)
        self.W.data[...] = np.random.randn(*shape)

    def __call__(self, x):
        return self.W * x
```

We can also initialize the parameter after the initialization by the *Link.add_param()* method.

```
class EltwiseParamProduct(Link):
    def __init__(self, shape):
        super(EltwiseParamProduct, self).__init__()
        self.add_param('W', shape)
        self.W.data[...] = np.random.randn(*shape)

    def __call__(self, x):
        return self.W * x
```

Note that the initializer and the *add_param()* method does not initialize elements of the parameter array. We have to manually initialize the elements by random values, zeros, etc.

For another example, assume we want to define a simple linear layer. It is already defined as *Linear*, so this is an educational example. The linear layer is divided into two parts: a function and its wrapper link. First, we have to define a function on variables:

```
class LinearFunction(Function):
    def forward(self, inputs):
        x, W, b = inputs
        return x.dot(W.T) + b,

    def backward(self, inputs, grad_outputs):
        x, W, b = inputs
        gy, = grad_outputs

        gx = gy.dot(W)
```

```
        gW = gy.T.dot(x)
        gb = gy.sum(axis=0)
        return gx, gW, gb

def linear(x, W, b):
    return LinearFunction()(x, W, b)
```

This function takes three arguments: input, weight, and bias. It can be used as a part of model definition, though is inconvenient since the user have to manage the weight and bias parameters directly. In order to make a convenient module, let's wrap it into a link:

```
class Linear(Link):
    def __init__(self, in_size, out_size):
        super(Linear, self).__init__(W=(out_size, in_size), b=out_size)
        self.W.data[...] = np.random.randn(out_size, in_size) / math.sqrt(in_size)
        self.b.data.fill(0)

    def __call__(self, x):
        return linear(x, self.W, self.b)
```

This link hides the parameters of the linear layer.

Note: An advanced tip to implement functions: if you want to preserve some information between forward and backward computations (e.g. to cache some arrays), you can store it as attributes. Be careful that it might increase the memory consumption during the whole forward-backward computation. If you want to train very large networks on a GPU with limited memory, it is not recommended to cache arrays between forward and backward. There is one exception for this: caching the output arrays does not change the memory consumption, because they are also held by the output Variable objects.

Warning: You should not assume a one-to-one match of calls of forward and backward. Some users may call backward more than once after one forward call.

Testing Function

In order to isolate the cause of learning failure from implementation bugs, it is important to test function implementations. Chainer provides simple utilities to help writing unit tests. They are defined in the *gradient_check* module.

The most important test utility is the *numerical_grad()* function. This function computes the numerical gradient of given function using finite differences. It can be used as follows

```
x  = np.random.randn(4, 3).astype(np.float32)
gy = np.ones((4, 3), dtype=np.float32)
f  = lambda: (x * x,)
gx = gradient_check.numerical_grad(f, (x,), (gy,))
```

`f` is a closure that returns a tuple of array(s) computed from input arrays. The second and third arguments of *numerical_grad()* are tuples of input arrays and output gradient arrays, respectively. The code above computes the numerical gradients of `sum(f(x))`, where `sum` indicates the summation over all elements. The summation can be weighted by changing `gy`. *numerical_grad()* function also accepts additional `eps` argument, which indicates the quantization width of finite differences.

Note: *numerical_grad()* function accepts both CPU and GPU arrays. Note that we cannot mix CPU and GPU arrays.

Another utility is *chainer.testing.assert_allclose()* function. This is similar to numpy.testing.assert_allclose() function. The difference is that Chainer's version accepts CPU and GPU arrays as inputs. We can mix them in one invocation of *chainer.testing.assert_allclose()*. The default values of optional arguments are also different.

Here is a typical usage of gradient checking utilities. This is a test example of *functions.relu()* function

```
import unittest

from chainer import testing

class TestReLU(unittest.TestCase):
    def test_backward_cpu(self):
        x = Variable(np.random.randn(3, 2).astype(np.float32))
        y = F.relu(x)
        y.grad = np.random.randn(3, 2).astype(np.float32)
        y.backward()

        f = lambda: (F.relu(x).data,)
        gx, = gradient_check.numerical_grad(f, (x.data,), (y.grad,))

        testing.assert_allclose(gx, x.grad)
```

The first four lines of the test code are simple forward and backward computation of ReLU function. The next two lines compute numerical gradient using the same forward function without backward routine. And at last, we compare these two results elementwise. Note that the above test code can be easily modified to test GPU version just by replacing CPU arrays to GPU arrays.

You can find many examples of function tests under `tests/chainer_tests/function_tests` directory.

Type check

In this section, you will learn about the following things:

- Basic usage of type check
- Detail of type information
- Internal mechanism of type check
- More complicated cases
- Call functions
- Typical type check example

After reading this section, you will be able to:

- Write a code to check types of input arguments of your own functions

Basic usage of type check

When you call a function with an invalid type of array, you sometimes receive no error, but get an unexpected result by broadcasting. When you use CUDA with an illegal type of array, it causes memory corruption, and you get a serious error. These bugs are hard to fix. Chainer can check preconditions of each function, and helps to prevent such problems. These conditions may help a user to understand specification of functions.

Each implementation of *Function* has a method for type check, `check_type_forward()`. This function is called just before the `forward()` method of the *Function* class. You can override this method to check the condition on types and shapes of arguments.

`check_type_forward()` gets an argument `in_types`:

```
def check_type_forward(self, in_types):
  ...
```

`in_types` is an instance of *TypeInfoTuple*, which is a sub-class of `tuple`. To get type information about the first argument, use `in_types[0]`. If the function gets multiple arguments, we recommend to use new variables for readability:

```
x_type, y_type = in_types
```

In this case, `x_type` represents the type of the first argument, and `y_type` represents the second one.

We describe usage of `in_types` with an example. When you want to check if the number of dimension of `x_type` equals to `2`, write this code:

```
utils.type_check.expect(x_type.ndim == 2)
```

When this condition is true, nothing happens. Otherwise this code throws an exception, and the user gets a message like this:

```
Traceback (most recent call last):
...
InvalidType: Expect: in_types[0].ndim == 2
Actual: 3 != 2
```

This error message means that "`ndim` of the first argument expected to be `2`, but actually it is `3`".

Detail of type information

You can access three information of `x_type`.

- `.shape` is a tuple of ints. Each value is size of each dimension.
- `.ndim` is `int` value representing the number of dimensions. Note that `ndim == len(shape)`
- `.dtype` is `numpy.dtype` representing data type of the value.

You can check all members. For example, the size of the first dimension must be positive, you can write like this:

```
utils.type_check.expect(x_type.shape[0] > 0)
```

You can also check data types with `.dtype`:

```
utils.type_check.expect(x_type.dtype == np.float64)
```

And an error is like this:

```
Traceback (most recent call last):
...
InvalidType: Expect: in_types[0].dtype == <type 'numpy.float64'>
Actual: float32 != <type 'numpy.float64'>
```

You can also check `kind` of `dtype`. This code checks if the type is floating point

```
utils.type_check.expect(x_type.dtype.kind == 'f')
```

You can compare between variables. For example, the following code checks if the first argument and the second argument have the same length:

```
utils.type_check.expect(x_type.shape[1] == y_type.shape[1])
```

Internal mechanism of type check

How does it show an error message like "`in_types[0].ndim == 2`"? If `x_type` is an object containing `ndim` member variable, we cannot show such an error message because this equation is evaluated as a boolean value by Python interpreter.

Actually `x_type` is a *Expr* objects, and doesn't have a `ndim` member variable itself. *Expr* represents a syntax tree. `x_type.ndim` makes a *Expr* object representing `(getattr, x_type, 'ndim')`. `x_type.ndim == 2` makes an object like `(eq, (getattr, x_type, 'ndim'), 2)`. `type_check.expect()` gets a *Expr* object and evaluates it. When it is `True`, it causes no error and shows nothing. Otherwise, this method shows a readable error message.

If you want to evaluate a *Expr* object, call `eval()` method:

```
actual_type = x_type.eval()
```

`actual_type` is an instance of `TypeInfo`, while `x_type` is an instance of *Expr*. In the same way, `x_type.shape[0].eval()` returns an int value.

More powerful methods

Expr class is more powerful. It supports all mathematical operators such as `+` and `*`. You can write a condition that the first dimension of `x_type` is the first dimension of `y_type` times four:

```
utils.type_check.expect(x_type.shape[0] == y_type.shape[0] * 4)
```

When `x_type.shape[0] == 3` and `y_type.shape[0] == 1`, users can get the error message below:

```
Traceback (most recent call last):
...
InvalidType: Expect: in_types[0].shape[0] == in_types[1].shape[0] * 4
Actual: 3 != 4
```

To compare a member variable of your function, wrap a value with `Variable` to show readable error message:

```
x_type.shape[0] == utils.type_check.Variable(self.in_size, "in_size")
```

This code can check the equivalent condition below:

```
x_type.shape[0] == self.in_size
```

However, the latter condition doesn't know the meaning of this value. When this condition is not satisfied, the latter code shows unreadable error message:

```
InvalidType: Expect: in_types[0].shape[0] == 4  # what does '4' mean?
Actual: 3 != 4
```

Note that the second argument of `utils.type_check.Variable` is only for readability.

The former shows this message:

```
InvalidType: Expect: in_types[0].shape[0] == in_size  # OK, `in_size` is a value that␣
↪is given to the constructor
Actual: 3 != 4  # You can also check actual value here
```

Call functions

How to check summation of all values of shape? *Expr* also supports function call:

```
sum = utils.type_check.Variable(np.sum, 'sum')
utils.type_check.expect(sum(x_type.shape) == 10)
```

Why do we need to wrap the function `numpy.sum` with `utils.type_check.Variable`? `x_type.shape` is not a tuple but an object of *Expr* as we have seen before. Therefore, `numpy.sum(x_type.shape)` fails. We need to evaluate this function lazily.

The above example produces an error message like this:

```
Traceback (most recent call last):
...
InvalidType: Expect: sum(in_types[0].shape) == 10
Actual: 7 != 10
```

More complicated cases

How to write a more complicated condition that can't be written with these operators? You can evaluate *Expr* and get its result value with `eval()` method. Then check the condition and show warning message by hand:

```
x_shape = x_type.shape.eval()  # get actual shape (int tuple)
if not more_complicated_condition(x_shape):
    expect_msg = 'Shape is expected to be ...'
    actual_msg = 'Shape is ...'
    raise utils.type_check.InvalidType(expect_msg, actual_msg)
```

Please write a readable error message. This code generates the following error message:

```
Traceback (most recent call last):
...
InvalidType: Expect: Shape is expected to be ...
Actual: Shape is ...
```

Typical type check example

We show a typical type check for a function.

First check the number of arguments:

```
utils.type_check.expect(in_types.size() == 2)
```

`in_types.size()` returns a *Expr* object representing the number of arguments. You can check it in the same way.

And then, get each type:

```
x_type, y_type = in_types
```

Don't get each value before checking `in_types.size()`. When the number of argument is illegal, `type_check.expect` might output unuseful error messages. For example, this code doesn't work when the size of `in_types` is 0:

```
utils.type_check.expect(
  in_types.size() == 2,
  in_types[0].ndim == 3,
)
```

After that, check each type:

```
utils.type_check.expect(
  x_type.dtype == np.float32,
  x_type.ndim == 3,
  x_type.shape[1] == 2,
)
```

The above example works correctly even when `x_type.ndim == 0` as all conditions are evaluated lazily.

CHAPTER 3

Chainer Reference Manual

Core functionalities

Variable

class chainer.**Variable**(*data*, *volatile=OFF*, *name=None*, *grad=None*)

Array with a structure to keep track of computation.

Every variable holds a data array of type either numpy.ndarray or *cupy.ndarray*.

A Variable object may be constructed in two ways: by the user or by some function. When a variable is created by some function as one of its outputs, the variable holds a reference to that function. This reference is used in error backpropagation (a.k.a. backprop). It is also used in *backward unchaining*. A variable that does not hold a reference to its creator is called a *root* variable. A variable is root if it is created by the user, or if the reference is deleted by *unchain_backward()*.

Users can disable this chaining behavior by setting the volatile flag for the initial variables. When a function gets volatile variables as its inputs, the output variables do not hold references to the function. This acts like unchaining on every function application.

Parameters

- **data** (array) – Initial data array.
- **volatile** (Flag) – Volatility flag. String ('on', 'off', or 'auto') or boolean values can be used, too.
- **name** (*str*) – Name of the variable.
- **grad** (array) – Initial gradient array.

Variables

- **data** – Data array of type either numpy.ndarray or *cupy.ndarray*.
- **grad** – Gradient array.

- **creator** – The function who creates this variable. It is None if the variable is not created by any function.
- **volatile** – Ternary *Flag* object. If 'ON', the variable does not keep track of any function applications. See *Flag* for the detail of ternary flags.

addgrad(*var*)
Accumulates the gradient array from given source variable.

This method just runs self.grad += var.grad, except that the accumulation is even done across the host and different devices.

Parameters **var** (Variable) – Source variable.

backward(*retain_grad=False*)
Runs error backpropagation (a.k.a. backprop) from this variable.

On backprop, *Function.backward()* is called on each *Function* object appearing in the backward graph starting from this variable. The backward graph is represented by backward references from variables to their creators, and from functions to their inputs. The backprop stops at all root variables. Some functions set None as gradients of some inputs, where further backprop does not take place at such input variables.

This method uses grad as the initial error array. User can manually set a gradient array before calling this method. If data contains only one element (i.e., it is scalar) and grad is None, then this method automatically complements 1.0 as the initial error. This is useful on starting backprop from some scalar loss value.

Parameters **retain_grad** (*bool*) – If True, the gradient arrays of all intermediate variables are kept. Otherwise, grad of the intermediate variables are set to None on appropriate timing, which may reduce the maximum memory consumption.

In most cases of training some models, the purpose of backprop is to compute gradients of parameters, not of variables, so it is recommended to set this flag False.

cleargrad()
Clears the gradient array.

copydata(*var*)
Copies the data array from given source variable.

This method just copies the data attribute from given variable to this variable, except that the copy is even done across the host and different devices.

Parameters **var** (Variable) – Source variable.

debug_print()
Display a summary of the stored data and location of the Variable

label
Short text that represents the variable.

reshape(**shape*)
Returns a variable of a different shape and the same content.

See also:

chainer.functions.reshape() for full documentation,

set_creator(*gen_func*)
Notifies the variable that the given function is its creator.

Parameters **gen_func** (Function) – Function object that creates this variable as one of its outputs.

to_cpu()
Copies the data and gradient arrays to CPU.

to_gpu(*device=None*)
Copies the data and gradient arrays to specified GPU.

Parameters **device** – Target device specifier. If omitted, the current device is used.

transpose(**axes*)
Permute the dimensions of an input variable without copy.

See also:

chainer.functions.transpose() for full documentation.

unchain_backward()
Deletes references between variables and functions backward.

After this method completes, intermediate variables and functions that are not referenced from anywhere are deallocated by reference count GC. Also this variable itself deletes the reference to its creator function, i.e. this variable becomes root in the computation graph. It indicates that backprop after unchaining stops at this variable. This behavior is useful to implement truncated BPTT.

zerograd()
Initializes the gradient array by zeros.

Deprecated since version v1.15: Use *cleargrad()* instead.

Flag

class chainer.**Flag**
Ternary flag object for variables.

It takes three values: ON, OFF, and AUTO.

ON and OFF flag can be evaluated as a boolean value. These are converted to True and False, respectively. AUTO flag cannot be converted to boolean. In this case, ValueError is raised.

Parameters **name** (*str, bool, or None*) – Name of the flag. Following values are allowed:

- `'on'`, `'ON'`, or `True` for ON value
- `'off'`, `'OFF'`, or `False` for OFF value
- `'auto'`, `'AUTO'`, or `None` for AUTO value

chainer.**ON = ON**
Equivalent to Flag('on').

chainer.**OFF = OFF**
Equivalent to Flag('off').

chainer.**AUTO = AUTO**
Equivalent to Flag('auto').

chainer.flag.**aggregate_flags**(*flags*)
Returns an aggregated flag given a sequence of flags.

If both ON and OFF are found, this function raises an error. Otherwise, either of ON and OFF that appeared is returned. If all flags are AUTO, then it returns AUTO.

Parameters **flags** (*sequence of Flag*) – Input flags.

Returns The result of aggregation.

Return type *Flag*

Function

class chainer.**Function**

Function on variables with backpropagation ability.

All function implementations defined in *chainer.functions* inherit this class.

The main feature of this class is keeping track of function applications as a backward graph. When a function is applied to *Variable* objects, its *forward()* method is called on data fields of input variables, and at the same time it chains references from output variables to the function and from the function to its inputs.

Note: As of v1.5, a function instance cannot be used twice in any computational graphs. In order to reuse a function object multiple times, use copy.copy() before the function applications to make a copy of the instance.

This restriction also means that we cannot make a *stateful function* anymore. For example, it is now not allowed to let a function hold parameters. Define a function as a pure (stateless) procedure, and use *Link* to combine it with parameter variables.

Example

Let x an instance of *Variable* and f an instance of *Function* taking only one argument. Then a line

```
>>> import numpy, chainer, chainer.functions as F
>>> x = chainer.Variable(numpy.zeros(10))
>>> f = F.Identity()
>>> y = f(x)
```

computes a new variable y and creates backward references. Actually, backward references are set as per the following diagram:

```
x <--- f <--- y
```

If an application of another function g occurs as

```
>>> g = F.Identity()
>>> z = g(x)
```

then the graph grows with a branch:

```
    |--- f <--- y
x <-+
    |--- g <--- z
```

Note that the branching is correctly managed on backward computation, i.e. the gradients from f and g are accumulated to the gradient of x.

Every function implementation should provide *forward_cpu()*, *forward_gpu()*, *backward_cpu()* and *backward_gpu()*. Alternatively, one can provide *forward()* and *backward()* instead of separate methods. Backward methods have default implementations that just return None, which indicates that the function is non- differentiable.

Variables

- **inputs** – A tuple or list of input variables.
- **outputs** – A tuple or list of output variables.
- **type_check_enable** – When it is `True`, the function checks types of input arguments. Set `CHAINER_TYPE_CHECK` environment variable `0` to disable type check, or set the variable directly in your own program.

add_hook(*hook*, *name=None*)
Registers the function hook.

Parameters

- **hook** (`FunctionHook`) – Function hook to be registered.
- **name** (*str*) – Name of the function hook. name must be unique among function hooks registered to the function. If `None`, default name of the function hook is used.

backward(*inputs*, *grad_outputs*)
Applies backprop to output gradient arrays.

It delegates the procedure to *backward_cpu()* or *backward_gpu()* by default. Which it selects is determined by the type of input arrays and output gradient arrays. Implementations of *Function* must implement either CPU/GPU methods or this method, if the function is intended to be backprop-ed.

Parameters

- **inputs** – Tuple of input arrays.
- **grad_outputs** – Tuple of output gradient arrays.

Returns Tuple of input gradient arrays. Some or all of them can be `None`, if the function is not differentiable on inputs.

Return type tuple

Warning: Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

backward_cpu(*inputs*, *grad_outputs*)
Applies backprop to output gradient arrays on CPU.

Parameters

- **inputs** – Tuple of input `numpy.ndarray` object(s).
- **grad_outputs** – Tuple of output gradient `numpy.ndarray` object(s).

Returns Tuple of input gradient `numpy.ndarray` object(s). Some or all of them can be `None`, if the function is not differentiable on corresponding inputs.

Return type tuple

Warning: Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

backward_gpu(*inputs*, *grad_outputs*)
Applies backprop to output gradient arrays on GPU.

Parameters

- **inputs** – Tuple of input *cupy.ndarray* object(s).
- **grad_outputs** – Tuple of output gradient *cupy.ndarray* object(s).

Returns Tuple of input gradient *cupy.ndarray* object(s). Some or all of them can be `None`, if the function is not differentiable on corresponding inputs.

Return type tuple

> **Warning:** Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

check_type_forward(*in_types*)
Checks types of input data before forward propagation.

Before *forward()* is called, this function is called. You need to validate types of input data in this function using *the type checking utilities*.

Parameters **in_types** (`TypeInfoTuple`) – The type information of input data for *forward()*.

delete_hook(*name*)
Unregisters the function hook.

Parameters **name** (*str*) – the name of the function hook to be unregistered.

forward(*inputs*)
Applies forward propagation to input arrays.

It delegates the procedure to *forward_cpu()* or *forward_gpu()* by default. Which it selects is determined by the type of input arrays. Implementations of *Function* must implement either CPU/GPU methods or this method.

Parameters **inputs** – Tuple of input array(s).

Returns Tuple of output array(s).

> **Warning:** Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

forward_cpu(*inputs*)
Applies forward propagation to input arrays on CPU.

Parameters **inputs** – Tuple of `numpy.ndarray` object(s).

Returns Tuple of `numpy.ndarray` object(s).

Return type tuple

> **Warning:** Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

forward_gpu(*inputs*)
Applies forward propagation to input arrays on GPU.

Parameters **inputs** – Tuple of *cupy.ndarray* object(s).

Returns Tuple of *cupy.ndarray* object(s).

Return type tuple

> **Warning:** Implementations of *Function* must take care that the return value must be a tuple even if it returns only one array.

label
Short text that represents the function.

The default implementation returns its type name. Each function should override it to give more information.

local_function_hooks
Ordered Dictionary of registered function hooks.

Contrary to `chainer.thread_local.function_hooks`, which registers its elements to all functions, Function hooks in this property is specific to this function.

unchain()
Purges in/out variables and this function itself from the graph.

This method is called from *Variable.unchain_backward()* method.

chainer.**force_backprop_mode**(**args*, ***kwds*)
Enable back-propagation for Variable whose volatile is auto.

When you want to enable back-propagation in *no_backprop_mode()*, call this method. In this context, *Variable* object whose `volatile` attribute is `'auto'` behaves like a **volatile** variable. That means you can disable *no_backprop_mode()* in this context.

If you call this method outside of *no_backprop_mode()* context, it changes nothing. *Variable* object with `volatile='auto'` behaves like a volatile variable by default.

In this example, the volatility of `x` and `y` is `'auto'`. In *no_backprop_mode()* context, `y` does not have a computational graph but in *force_backprop_mode()* it has a graph.

```
>>> with chainer.no_backprop_mode():
...   # Variable with volatile='auto' behaves like volatile='on'
...   with chainer.force_backprop_mode():
...     # Variable with volatile='auto' behaves like volatile='off'
...     y = x + 1
```

See also:

See *no_backprop_mode()* for details of back-prop mode.

chainer.**no_backprop_mode**(**args*, ***kwds*)
Disable back-propagation for Variable whose volatile is auto.

In the default setting a *Variable* object whose `volatile` attribute is `'auto'` behaves like a **non-volatile** variable. That means such a *Variable* object builds a computational graph, consumes memory to store the graph, and you can execute back-propagation for it. With this context such a *Variable* object behaves like a **volatile** variable. So, you can easily switch training and evaluation.

In this example, the volatility of `x` and `y` is `'auto'`. So, `y` does not have a computational graph.

```
>>> x = chainer.Variable(numpy.array([1,], 'f'), volatile='auto')
>>> with chainer.no_backprop_mode():
...     y = x + 1
```

Link and Chain

class chainer.**Link**(**params*)
Building block of model definitions.

Link is a building block of neural network models that support various features like handling parameters, defining network fragments, serialization, etc.

Link is the primitive structure for the model definitions. It supports management of parameter variables and *persistent values* that should be incorporated to serialization. Parameters are variables registered via the *add_param()* method, or given to the initializer method. Persistent values are arrays, scalars, or any other serializable values registered via the *add_persistent()* method.

Note: Whereas arbitrary serializable objects can be registered as persistent values, it is strongly recommended to just register values that should be treated as results of learning. A typical example of persistent values is ones computed during training and required for testing, e.g. running statistics for batch normalization.

Parameters and persistent values are referred by their names. They can be accessed as attributes of the links. Link class itself manages the lists of names of parameters and persistent values to distinguish parameters and persistent values from other attributes.

Link can be composed into more complex models. This composition feature is supported by child classes like *Chain* and *ChainList*. One can create a chain by combining one or more links. See the documents for these classes for details.

As noted above, Link supports the serialization protocol of the *Serializer* class. **Note that only parameters and persistent values are saved and loaded.** Other attributes are considered as a part of user program (i.e. a part of network definition). In order to construct a link from saved file, other attributes must be identically reconstructed by user codes.

Example

This is a simple example of custom link definition. Chainer itself also provides many links defined under the *links* module. They might serve as examples, too.

Consider we want to define a simple primitive link that implements a fully-connected layer based on the *linear()* function. Note that this function takes input units, a weight variable, and a bias variable as arguments. Then, the fully-connected layer can be defined as follows:

```
import chainer
import chainer.functions as F
import numpy as np

class LinearLayer(chainer.Link):

    def __init__(self, n_in, n_out):
        # Parameters are initialized as a numpy array of given shape.
        super(LinearLayer, self).__init__(
            W=(n_out, n_in),
            b=(n_out,),
        )
        self.W.data[...] = np.random.randn(n_out, n_in)
        self.b.data.fill(0)

    def __call__(self, x):
        return F.linear(x, self.W, self.b)
```

This example shows that a user can define arbitrary parameters and use them in any methods. Links typically implement the __call__ operator.

Parameters **params** – Names, shapes, and optional dtypes of initial parameters. The keywords are used as the parameter names and the corresponding values consist either of the shape or a tuple of shape and a dtype *(shape, dtype)*. If only the shape is supplied, the default dtype will be used.

Variables **name** (*str*) – Name of this link, given by the parent chain (if exists).

add_param(*name*, *shape*, *dtype=<type 'numpy.float32'>*, *initializer=None*)
Registers a parameter to the link.

The registered parameter is saved and loaded on serialization and deserialization, and involved in the optimization. The data and gradient of the variable are initialized by NaN arrays. If `initializer` is not `None`, the data is initialized by `initializer`.

If the supplied `name` argument corresponds to an uninitialized parameter (that is, one that was added with the *add_uninitialized_param()* method), `name` will be removed from the set of uninitialized parameters.

The parameter is set to an attribute of the link with the given name.

Parameters

- **name** (*str*) – Name of the parameter. This name is also used as the attribute name. Any uninitialized parameters with the same name will be removed.
- **shape** (*int or tuple of ints*) – Shape of the parameter array.
- **dtype** – Data type of the parameter array.
- **initializer** (`chainer.initializer.Initializer`) – If it is not `None`, the data is initialized with the given initializer. Note that in this case `dtype` argument is ignored.

add_persistent(*name*, *value*)
Registers a persistent value to the link.

The registered value is saved and loaded on serialization and deserialization. The value is set to an attribute of the link.

Parameters

- **name** (*str*) – Name of the persistent value. This name is also used for the attribute name.
- **value** – Value to be registered.

add_uninitialized_param(*name*)
Registers an uninitialized parameter to the link.

An uninitialized parameter is defined as a parameter that has a name but that does not yet have a shape. If the shape of a parameter depends on the shape of the inputs to the __call__ operator, it can be useful to defer initialization (that is, setting the shape) until the first forward call of the link. Such parameters are intended to be defined as uninitialized parameters in the initializer and then initialized during the first forward call.

An uninitialized parameter is intended to be registered to a link by calling this method in the initializer method. Then, during the first forward call, the shape of the parameter will be determined from the size of the inputs and the parameter must be initialized by calling the *add_param()* method.

Parameters **name** – (str): Name of the uninitialized parameter.

addgrads(*link*)
Accumulates gradient values from given link.

This method adds each gradient array of the given link to corresponding gradient array of this link. The accumulation is even done across host and different devices.

Parameters **link** (Link) – Source link object.

children()
Returns a generator of all child links.

Returns A generator object that generates all child links.

cleargrads()
Clears all gradient arrays.

This method should be called before the backward computation at every iteration of the optimization.

copy()
Copies the link hierarchy to new one.

The whole hierarchy rooted by this link is copied. The copy is basically shallow, except that the parameter variables are also shallowly copied. It means that the parameter variables of copied one are different from ones of original link, while they share the data and gradient arrays.

The name of the link is reset on the copy, since the copied instance does not belong to the original parent chain (even if exists).

Returns Copied link object.

Return type *Link*

copyparams(*link*)
Copies all parameters from given link.

This method copies data arrays of all parameters in the hierarchy. The copy is even done across the host and devices. Note that this method does not copy the gradient arrays.

Parameters **link** (Link) – Source link object.

has_uninitialized_params
Check if the link has uninitialized parameters.

Returns `True` if the link has any uninitialized parameters. Otherwise returns `False`.

Return type bool

links(*skipself=False*)
Returns a generator of all links under the hierarchy.

Parameters **skipself** (*bool*) – If `True`, then the generator skips this link and starts with the first child link.

Returns A generator object that generates all links.

namedlinks(*skipself=False*)
Returns a generator of all (path, link) pairs under the hierarchy.

Parameters **skipself** (*bool*) – If `True`, then the generator skips this link and starts with the first child link.

Returns A generator object that generates all (path, link) pairs.

namedparams()
Returns a generator of all (path, param) pairs under the hierarchy.

Returns A generator object that generates all (path, parameter) pairs. The paths are relative from this link.

params()
Returns a generator of all parameters under the link hierarchy.

Returns A generator object that generates all parameters.

serialize(*serializer*)
Serializes the link object.

Parameters **serializer** (AbstractSerializer) – Serializer object.

to_cpu()
Copies parameter variables and persistent values to CPU.

This method does not handle non-registered attributes. If some of such attributes must be copied to CPU, the link implementation must override this method to do so.

Returns: self

to_gpu(*device=None*)
Copies parameter variables and persistent values to GPU.

This method does not handle non-registered attributes. If some of such attributes must be copied to GPU, the link implementation must override this method to do so.

Parameters **device** – Target device specifier. If omitted, the current device is used.

Returns: self

xp
Array module for this link.

Depending on which of CPU/GPU this link is on, this property returns numpy or *cupy*.

zerograds()
Initializes all gradient arrays by zero.

This method can be used for the same purpose of cleargrads, but less efficient. This method is left for backward compatibility.

Deprecated since version v1.15: Use *cleargrads()* instead.

class chainer.**Chain**(***links*)
Composable link with object-like interface.

Composability is one of the most important features of neural nets. Neural net models consist of many reusable fragments, and each model itself might be embedded into a larger learnable system. Chain enables us to write a neural net based on composition, without bothering about routine works like collecting parameters, serialization, copying the structure with parameters shared, etc.

This class actually provides a way to compose one or more links into one structure. A chain can contain one or more *child links*. Child link is a link registered to the chain with its own name. The child link is stored to an attribute of the chain with the name. User can write a whole model or a fragment of neural nets as a child class of Chain.

Each chain itself is also a link. Therefore, one can combine chains into higher-level chains. In this way, links and chains construct a *link hierarchy*. Link hierarchy forms a tree structure, where each node is identified by the path from the root. The path is represented by a string like a file path in UNIX, consisting of names of nodes on the path, joined by slashes /.

Example

This is a simple example of custom chain definition. Chainer itself also provides some chains defined under the *links* module. They might serve as examples, too.

Consider we want to define a multi-layer perceptron consisting of two hidden layers with rectifiers as activation functions. We can use the *Linear* link as a building block:

```
import chainer
import chainer.functions as F
import chainer.links as L

class MultiLayerPerceptron(chainer.Chain):

    def __init__(self, n_in, n_hidden, n_out):
        # Create and register three layers for this MLP
        super(MultiLayerPerceptron, self).__init__(
            layer1=L.Linear(n_in, n_hidden),
            layer2=L.Linear(n_hidden, n_hidden),
            layer3=L.Linear(n_hidden, n_out),
        )

    def __call__(self, x):
        # Forward propagation
        h1 = F.relu(self.layer1(x))
        h2 = F.relu(self.layer2(h1))
        return self.layer3(h2)
```

Child links are registered via the initializer method. They also can be registered by the *add_link()* method. The forward propagation is often implemented as The `__call__` operator as the above example, though it is not mandatory.

Parameters **links** – Child links. The keywords are used as their names. The names are also set to the links.

add_link(*name*, *link*)

Registers a child link to this chain.

The registered link is saved and loaded on serialization and deserialization, and involved in the optimization. The registered link is called a child. The child link is set to an attribute of the chain with the given name.

This method also sets the `name` attribute of the registered link. If the given link already has the name attribute set, then it raises an error.

Parameters

- **name** (*str*) – Name of the child link. This name is also used as the attribute name.
- **link** (Link) – The link object to be registered.

class chainer.**ChainList**(**links*)

Composable link with list-like interface.

This is another example of compositional link. Unlike *Chain*, this class can be used like a list of child links. Each child link is indexed by a non-negative integer, and it maintains the current number of registered child links. The *add_link()* method inserts a new link at the end of the list. It is useful to write a chain with arbitrary number of child links, e.g. an arbitrarily deep multi-layer perceptron.

Note that this class does not implement all methods of `list`.

Parameters **links** – Initial child links.

add_link(*link*)
Registers a child link to this chain.

The registered link is saved and loaded on serialization and deserialization, and involved in the optimization. The registered link is called a child. The child link is accessible via `children()` generator, which returns a generator running through the children in registered order.

This method also sets the `name` attribute of the registered link. If the given link already has the name attribute set, then it raises an error.

Parameters **link** (`Link`) – The link object to be registered.

Optimizer

class chainer.**Optimizer**
Base class of all numerical optimizers.

This class provides basic features for all optimization methods. It optimizes parameters of a *target link*. The target link is registered via the *setup()* method, and then the *update()* method updates its parameters based on a given loss function.

Each optimizer implementation must be defined as a child class of Optimizer. It must override *update()* method. An optimizer can use *internal states* each of which is tied to one of the parameters. State is a dictionary of serializable values (typically arrays of size same as the corresponding parameters). In order to use state dictionaries, the optimizer must override *init_state()* method (or its CPU/GPU versions, *init_state_cpu()* and *init_state_gpu()*).

If the optimizer is based on single gradient computation (like most first-order methods), then it should inherit *GradientMethod*, which adds some features dedicated for the first order methods.

Optimizer instance also supports *hook functions*. Hook function is registered by the *add_hook()* method. Each hook function is called in registration order in advance of the actual parameter update.

Variables

- **target** – Target link object. It is set by the *setup()* method.
- **t** – Number of update steps. It must be incremented by the *update()* method.
- **epoch** – Current epoch. It is incremented by the *new_epoch()* method.

accumulate_grads(*grads*)
Accumulates gradients from other source.

This method just adds given gradient arrays to gradients that this optimizer holds. It is typically used in data-parallel optimization, where gradients for different shards are computed in parallel and aggregated by this method. This method correctly treats multiple GPU devices.

Parameters **grads** (*Iterable*) – Iterable of gradient arrays to be accumulated.

Deprecated since version v1.5: Use the *chainer.Link.addgrads()* method of the target link instead.

add_hook(*hook*, *name=None*)
Registers a hook function.

Hook function is typically called right after the gradient computation, though the timing depends on the optimization method.

Parameters

- **hook** (`function`) – Hook function. It accepts the optimizer object.

- **name** (*str*) – Name of the registration. If omitted, `hook.name` is used by default.

call_hooks()
Invokes hook functions in registration order.

clip_grads(*maxnorm*)
Clips the norm of whole gradients up to the threshold.

Parameters **maxnorm** (*float*) – Threshold of gradient L2 norm.

Deprecated since version v1.5: Use the *GradientClipping* hook function instead.

compute_grads_norm()
Computes the norm of whole gradients.

Returns L2 norm of whole gradients, i.e. square root of sum of square of all gradient elements.

Return type float

Warning: This method returns a CPU-computed value, which means that this method synchronizes between CPU and GPU if at least one of the gradients reside on the GPU.

Deprecated since version v1.5.

init_state(*param*, *state*)
Initializes the optimizer state corresponding to the parameter.

This method should add needed items to the `state` dictionary. Each optimizer implementation that uses its own states should override this method or CPU/GPU dedicated versions (*init_state_cpu()* and *init_state_gpu()*).

Parameters

- **param** (*Variable*) – Parameter variable.
- **state** (*dict*) – State dictionary.

See also:

init_state_cpu(), *init_state_gpu()*

init_state_cpu(*param*, *state*)
Initializes the optimizer state on CPU.

This method is called from *init_state()* by default.

Parameters

- **param** (*Variable*) – Parameter variable. Its data array is of type `numpy.ndarray`.
- **state** (*dict*) – State dictionary.

See also:

init_state()

init_state_gpu(*param*, *state*)
Initializes the optimizer state on GPU.

This method is called from *init_state()* by default.

Parameters

- **param** (*Variable*) – Parameter variable. Its data array is of type *cupy.ndarray*.
- **state** (*dict*) – State dictionary.

See also:

init_state()

new_epoch()
Starts a new epoch.

This method increments the epoch count. Note that if the optimizer depends on the epoch count, then user should call this method appropriately at the beginning of each epoch.

prepare()
Prepares for an update.

This method initializes missing optimizer states (e.g. for newly added parameters after the set up), and copies arrays in each state dictionary to CPU or GPU according to the corresponding parameter array.

remove_hook(*name*)
Removes a hook function.

Parameters **name** (*str*) – Registered name of the hook function to remove.

serialize(*serializer*)
Serializes or deserializes the optimizer.

It only saves or loads the following things:

•Optimizer states

•Global states (t and epoch)

It does not saves nor loads the parameters of the target link. They should be separately saved or loaded.

Parameters **serializer** (AbstractSerializer) – Serializer or deserializer object.

setup(*link*)
Sets a target link and initializes the optimizer states.

Given link is set to the target attribute. It also prepares the optimizer state dictionaries corresponding to all parameters in the link hierarchy. The existing states are discarded.

Parameters **link** (Link) – Target link object.

update(*lossfun=None*, **args*, ***kwds*)
Updates the parameters and optimizer states.

This method updates the parameters of the target link and corresponding optimizer states. The behavior of this method is different for the cases either lossfun is given or not.

If lossfun is given, then this method initializes the gradients by zeros, calls it with given extra arguments, and calls the *backward()* method of its output to compute the gradients. The implementation might call lossfun more than once.

If lossfun is not given, then this method assumes that the gradients of all parameters are already computed. An implementation that requires multiple gradient computations might raise an error on this case.

In both cases, this method invokes the update procedure for all parameters.

Parameters

- **lossfun** (function) – Loss function. It accepts arbitrary arguments and returns one *Variable* object that represents the loss (or objective) value. This argument can be omitted for single gradient-based methods. In this case, this method assumes gradient arrays computed.
- **kwds** (*args,*) – Arguments for the loss function.

weight_decay(*decay*)
Applies weight decay to the parameter/gradient pairs.

Parameters **decay** (*float*) – Coefficient of weight decay.

Deprecated since version v1.5: Use the *WeightDecay* hook function instead.

zero_grads()
Fills all gradient arrays by zeros.

Deprecated since version v1.5: Use the *chainer.Link.cleargrads()* method for the target link instead.

class chainer.**GradientMethod**
Base class of all single gradient-based optimizers.

This is an extension of the *Optimizer* class. Typical gradient methods that just require the gradient at the current parameter vector on an update can be implemented as its child class.

An implementation of a gradient method must override the following methods:

- `init_state()` or both `init_state_cpu()` and `init_state_gpu()`
- *update_one()* or both *update_one_cpu()* and *update_one_gpu()*

Note: It is recommended to call *use_cleargrads()* after creating a *GradientMethod* object for efficiency.

---

call_hooks()
Invokes hook functions in registration order.

reallocate_cleared_grads()
Reallocate gradients cleared by *cleargrad()*.

This method allocates arrays for all gradients which have `None`. This method is called before and after every optimizer hook. If an inheriting optimizer does not require this allocation, the optimizer can override this method with a blank function.

update(*lossfun=None*, **args*, ***kwds*)
Updates parameters based on a loss function or computed gradients.

This method runs in two ways.

- If `lossfun` is given, then use it as a loss function to compute gradients.
- Otherwise, this method assumes that the gradients are already computed.

In both cases, the computed gradients are used to update parameters. The actual update routines are defined by the *update_one()* method (or its CPU/GPU versions, *update_one_cpu()* and *update_one_gpu()*).

update_one(*param*, *state*)
Updates a parameter based on the corresponding gradient and state.

This method calls appropriate one from `update_param_cpu()` or `update_param_gpu()`.

Parameters

- **param** (`Variable`) – Parameter variable.
- **state** (*dict*) – State dictionary.

update_one_cpu(*param*, *state*)
Updates a parameter on CPU.

Parameters

- **param** (Variable) – Parameter variable.
- **state** (*dict*) – State dictionary.

update_one_gpu(*param*, *state*)
Updates a parameter on GPU.

Parameters

- **param** (Variable) – Parameter variable.
- **state** (*dict*) – State dictionary.

use_cleargrads(*use=True*)
Enables or disables use of *cleargrads()* in *update*.

Parameters **use** (*bool*) – If True, this function enables use of *cleargrads*. If False, disables use of *cleargrads* (*zerograds* is used).

Note: Note that *update()* calls *zerograds()* by default for backward compatibility. It is recommended to call this method before first call of *update* because *cleargrads* is more efficient than *zerograds*.

Hook functions

class chainer.optimizer.**WeightDecay**(*rate*)
Optimizer hook function for weight decay regularization.

This hook function adds a scaled parameter to the corresponding gradient. It can be used as a regularization.

Parameters **rate** (*float*) – Coefficient for the weight decay.

Variables **rate** (*float*) – Coefficient for the weight decay.

class chainer.optimizer.**Lasso**(*rate*)
Optimizer hook function for Lasso regularization.

This hook function adds a scaled parameter to the sign of each weight. It can be used as a regularization.

Parameters **rate** (*float*) – Coefficient for the weight decay.

Variables **rate** (*float*) – Coefficient for the weight decay.

class chainer.optimizer.**GradientClipping**(*threshold*)
Optimizer hook function for gradient clipping.

This hook function scales all gradient arrays to fit to the defined L2 norm threshold.

Parameters **threshold** (*float*) – L2 norm threshold.

Variables **threshold** (*float*) – L2 norm threshold of gradient norm.

class chainer.optimizer.**GradientNoise**(*eta*, *noise_func=<function exponential_decay_noise>*)
Optimizer hook function for adding gradient noise.

This hook function simply adds noise generated by the noise_func to the gradient. By default it adds time-dependent annealed Gaussian noise to the gradient at every training step:

$$g_t \leftarrow g_t + N(0, \sigma_t^2)$$

where

$$\sigma_t^2 = \frac{\eta}{(1+t)^\gamma}$$

with η selected from $\{0.01, 0.3, 1.0\}$ and $\gamma = 0.55$.

Parameters

- **eta** (*float*) – Parameter that defines the scale of the noise, which for the default noise function is recommended to be either 0.01, 0.3 or 1.0.
- **noise_func** (function) – Noise generating function which by default is given by Adding Gradient Noise Improves Learning for Very Deep Networks.

Serializer

class chainer.**AbstractSerializer**
Abstract base class of all serializers and deserializers.

class chainer.**Serializer**
Base class of all serializers.

save(*obj*)
Saves an object by this serializer.

This is equivalent to `obj.serialize(self)`.

Parameters **obj** – Target object to be serialized.

class chainer.**Deserializer**
Base class of all deserializers.

load(*obj*)
Loads an object from this deserializer.

This is equivalent to `obj.serialize(self)`.

Parameters **obj** – Target object to be serialized.

Dataset abstraction

Chainer has a support of common interface of training and validation datasets. The dataset support consists of three components: datasets, iterators, and batch conversion functions.

Dataset represents a set of examples. The interface is only determined by combination with iterators you want to use on it. The built-in iterators of Chainer requires the dataset to support `__getitem__` and `__len__` method. In particular, the `__getitem__` method should support indexing by both an integer and a slice. We can easily support slice indexing by inheriting *DatasetMixin*, in which case users only have to implement *get_example()* method for indexing. Some iterators also restrict the type of each example. Basically, datasets are considered as *stateless* objects, so that we do not need to save the dataset as a checkpoint of the training procedure.

Iterator iterates over the dataset, and at each iteration, it yields a mini batch of examples as a list. Iterators should support the *Iterator* interface, which includes the standard iterator protocol of Python. Iterators manage where to read next, which means they are *stateful*.

Batch conversion function converts the mini batch into arrays to feed to the neural nets. They are also responsible to send each array to an appropriate device. Chainer currently provides *concat_examples()* as the only example of batch conversion functions.

These components are all customizable, and designed to have a minimum interface to restrict the types of datasets and ways to handle them. In most cases, though, implementations provided by Chainer itself are enough to cover the usages.

Chainer also has a light system to download, manage, and cache concrete examples of datasets. All datasets managed through the system are saved under *the dataset root directory*, which is determined by the `CHAINER_DATASET_ROOT` environment variable, and can also be set by the *`set_dataset_root()`* function.

Dataset representation

See *Dataset examples* for dataset implementations.

class `chainer.dataset.`**`DatasetMixin`**

Default implementation of dataset indexing.

DatasetMixin provides the `__getitem__()` operator. The default implementation uses *`get_example()`* to extract each example, and combines the results into a list. This mixin makes it easy to implement a new dataset that does not support efficient slicing.

Dataset implementation using DatasetMixin still has to provide the `__len__()` operator explicitly.

`get_example`(*i*)

Returns the i-th example.

Implementations should override it. It should raise `IndexError` if the index is invalid.

Parameters **`i`** (*int*) – The index of the example.

Returns The i-th example.

Iterator interface

See *Iterator examples* for dataset iterator implementations.

class `chainer.dataset.`**`Iterator`**

Base class of all dataset iterators.

Iterator iterates over the dataset, yielding a minibatch at each iteration. Minibatch is a list of examples. Each implementation should implement an iterator protocol (e.g., the `__next__()` method).

Note that, even if the iterator supports setting the batch size, it does not guarantee that each batch always contains the same number of examples. For example, if you let the iterator to stop at the end of the sweep, the last batch may contain a fewer number of examples.

The interface between the iterator and the underlying dataset is not fixed, and up to the implementation.

Each implementation should provide the following attributes (not needed to be writable).

- `batch_size`: Number of examples within each minibatch.
- `epoch`: Number of completed sweeps over the dataset.
- `epoch_detail`: Floating point number version of the epoch. For example, if the iterator is at the middle of the dataset at the third epoch, then this value is 2.5.
- `previous_epoch_detail`: The value of `epoch_detail` at the previous iteration. This value is `None` before the first iteration.
- `is_new_epoch`: `True` if the epoch count was incremented at the last update.

Each implementation should also support serialization to resume/suspend the iteration.

finalize()
Finalizes the iterator and possibly releases the resources.

This method does nothing by default. Implementation may override it to better handle the internal resources.

next()
Python2 alternative of `__next__`.

It calls `__next__()` by default.

serialize(*serializer*)
Serializes the internal state of the iterator.

This is a method to support serializer protocol of Chainer.

Note: It should only serialize the internal state that changes over the iteration. It should not serializes what is set manually by users such as the batch size.

Batch conversion function

chainer.dataset.**concat_examples**(*batch*, *device=None*, *padding=None*)
Concatenates a list of examples into array(s).

Dataset iterator yields a list of examples. If each example is an array, this function concatenates them along the newly-inserted first axis (called *batch dimension*) into one array. The basic behavior is same for examples consisting of multiple arrays, i.e., corresponding arrays of all examples are concatenated.

For instance, consider each example consists of two arrays `(x, y)`. Then, this function concatenates `x` 's into one array, and `y` 's into another array, and returns a tuple of these two arrays. Another example: consider each example is a dictionary of two entries whose keys are `'x'` and `'y'`, respectively, and values are arrays. Then, this function concatenates `x` 's into one array, and `y` 's into another array, and returns a dictionary with two entries `x` and `y` whose values are the concatenated arrays.

When the arrays to concatenate have different shapes, the behavior depends on the `padding` value. If `padding` is `None` (default), it raises an error. Otherwise, it builds an array of the minimum shape that the contents of all arrays can be substituted to. The padding value is then used to the extra elements of the resulting arrays.

TODO(beam2d): Add an example.

Parameters

- **batch** (*list*) – A list of examples. This is typically given by a dataset iterator.
- **device** (*int*) – Device ID to which each array is sent. Negative value indicates the host memory (CPU). If it is omitted, all arrays are left in the original device.
- **padding** – Scalar value for extra elements. If this is None (default), an error is raised on shape mismatch. Otherwise, an array of minimum dimensionalities that can accommodate all arrays is created, and elements outside of the examples are padded by this value.

Returns Array, a tuple of arrays, or a dictionary of arrays. The type depends on the type of each example in the batch.

chainer.dataset.**to_device**(*device*, *x*)
Send an array to a given device.

This method send a given array to a given device. This method is used in *concat_examples()*. You can also use this method in a custom converter method used in *Updater* and *Extension* such as *StandardUpdater* and *Evaluator*.

Parameters

- **device** (*int or None*) – Device ID to which an array is sent. If it is negative value, an array is sent to CPU. If it is positive, an array is sent to GPU with the given ID. If it is None, an array is left in the original device.
- **x** (*numpy.ndarray or cupy.ndarray*) – An array to send.

Returns Converted array.

Dataset management

chainer.dataset.**get_dataset_root**()
Gets the path to the root directory to download and cache datasets.

Returns The path to the dataset root directory.

Return type str

chainer.dataset.**set_dataset_root**(*path*)
Sets the root directory to download and cache datasets.

There are two ways to set the dataset root directory. One is by setting the environment variable CHAINER_DATASET_ROOT. The other is by using this function. If both are specified, one specified via this function is used. The default dataset root is $HOME/.chainer/dataset.

Parameters **path** (*str*) – Path to the new dataset root directory.

chainer.dataset.**cached_download**(*url*)
Downloads a file and caches it.

It downloads a file from the URL if there is no corresponding cache. After the download, this function stores a cache to the directory under the dataset root (see *set_dataset_root()*). If there is already a cache for the given URL, it just returns the path to the cache without downloading the same file.

Parameters **url** (*str*) – URL to download from.

Returns Path to the downloaded file.

Return type str

chainer.dataset.**cache_or_load_file**(*path*, *creator*, *loader*)
Caches a file if it does not exist, or loads it otherwise.

This is a utility function used in dataset loading routines. The creator creates the file to given path, and returns the content. If the file already exists, the loader is called instead, and it loads the file and returns the content.

Note that the path passed to the creator is temporary one, and not same as the path given to this function. This function safely renames the file created by the creator to a given path, even if this function is called simultaneously by multiple threads or processes.

Parameters

- **path** (*str*) – Path to save the cached file.
- **creator** – Function to create the file and returns the content. It takes a path to temporary place as the argument. Before calling the creator, there is no file at the temporary path.
- **loader** – Function to load the cached file and returns the content.

Returns It returns the returned values by the creator or the loader.

Training loop abstraction

Chainer provides a standard implementation of the training loops under the *chainer.training* module. It is built on top of many other core features of Chainer, including Variable and Function, Link/Chain/ChainList, Optimizer, Dataset, and Reporter/Summary. Compared to the training loop abstraction of other machine learning tool kits, Chainer's training framework aims at maximal flexibility, while keeps the simplicity for the typical usages. Most components are pluggable, and users can overwrite the definition.

The core of the training loop abstraction is *Trainer*, which implements the training loop itself. The training loop consists of two parts: one is *Updater*, which actually updates the parameters to train, and the other is *Extension* for arbitrary functionalities other than the parameter update.

Updater and some extensions use *dataset* and `Iterator` to scan the datasets and load mini batches. The trainer also uses *Reporter* to collect the observed values, and some extensions use *DictSummary* to accumulate them and computes the statistics.

You can find many examples for the usage of this training utilities from the official examples. You can also search the extension implementations from *Trainer extensions*.

Trainer

class `chainer.training.`**`Trainer`**(*updater*, *stop_trigger=None*, *out='result'*)

The standard training loop in Chainer.

Trainer is an implementation of a training loop. Users can invoke the training by calling the *run()* method.

Each iteration of the training loop proceeds as follows.

- Update of the parameters. It includes the mini-batch loading, forward and backward computations, and an execution of the update formula. These are all done by the update object held by the trainer.
- Invocation of trainer extensions in the descending order of their priorities. A trigger object is attached to each extension, and it decides at each iteration whether the extension should be executed. Trigger objects are callable objects that take the trainer object as the argument and return a boolean value indicating whether the extension should be called or not.

Extensions are callable objects that take the trainer object as the argument. There are three ways to define custom extensions: inheriting the *Extension* class, decorating functions by *make_extension()*, and defining any callable including lambda functions. See *Extension* for more details on custom extensions and how to configure them.

Users can register extensions to the trainer by calling the *extend()* method, where some configurations can be added.

- Trigger object, which is also explained above. In most cases, `IntervalTrigger` is used, in which case users can simply specify a tuple of the interval length and its unit, like `(1000, 'iteration')` or `(1, 'epoch')`.
- The order of execution of extensions is determined by their priorities. Extensions of higher priorities are invoked earlier. There are three standard values for the priorities:
  - `PRIORITY_WRITER`. This is the priority for extensions that write some records to the `observation` dictionary. It includes cases that the extension directly adds values to the observation dictionary, or the extension uses the *chainer.report()* function to report values to the observation dictionary.

- PRIORITY_EDITOR. This is the priority for extensions that edit the observation dictionary based on already reported values.
- PRIORITY_READER. This is the priority for extensions that only read records from the observation dictionary. This is also suitable for extensions that do not use the observation dictionary at all.

• Extensions with invoke_before_training flag on are also invoked at the beginning of the training loop. Extensions that update the training status (e.g., changing learning rates) should have this flag to be True to ensure that resume of the training loop correctly recovers the training status.

The current state of the trainer object and objects handled by the trainer can be serialized through the standard serialization protocol of Chainer. It enables us to easily suspend and resume the training loop.

Note: The serialization does not recover everything of the training loop. It only recovers the states which change over the training (e.g. parameters, optimizer states, the batch iterator state, extension states, etc.). You must initialize the objects correctly before deserializing the states.

On the other hand, it means that users can change the settings on deserialization. For example, the exit condition can be changed on the deserialization, so users can train the model for some iterations, suspend it, and then resume it with larger number of total iterations.

During the training, it also creates a *Reporter* object to store observed values on each update. For each iteration, it creates a fresh observation dictionary and stores it in the observation attribute.

Links of the target model of each optimizer are registered to the reporter object as observers, where the name of each observer is constructed as the format <optimizer name><link name>. The link name is given by the chainer.Link.namedlink() method, which represents the path to each link in the hierarchy. Other observers can be registered by accessing the reporter object via the reporter attribute.

The default trainer is *plain*, i.e., it does not contain any extensions.

Parameters

- **updater** (Updater) – Updater object. It defines how to update the models.
- **stop_trigger** – Trigger that determines when to stop the training loop. If it is not callable, it is passed to IntervalTrigger.

Variables

- **updater** – The updater object for this trainer.
- **stop_trigger** – Trigger that determines when to stop the training loop. The training loop stops at the iteration on which this trigger returns True.
- **observation** – Observation of values made at the last update. See the Reporter class for details.
- **out** – Output directory.
- **reporter** – Reporter object to report observed values.

elapsed_time

Total time used for the training.

The time is in seconds. If the training is resumed from snapshot, it includes the time of all the previous training to get the current state of the trainer.

extend(*extension*, *name=None*, *trigger=None*, *priority=None*, *invoke_before_training=None*)

Registers an extension to the trainer.

Extension is a callable object which is called after each update unless the corresponding trigger object decides to skip the iteration. The order of execution is determined by priorities: extensions with higher priorities are called earlier in each iteration. Extensions with the same priority are invoked in the order of registrations.

If two or more extensions with the same name are registered, suffixes are added to the names of the second to last extensions. The suffix is `_N` where N is the ordinal of the extensions.

See *Extension* for the interface of extensions.

Parameters

- **extension** – Extension to register.
- **name** (*str*) – Name of the extension. If it is omitted, the `default_name` attribute of the extension is used instead. Note that the name would be suffixed by an ordinal in case of duplicated names as explained above.
- **trigger** (*tuple or Trigger*) – Trigger object that determines when to invoke the extension. If it is `None`, `extension.trigger` is used instead. If it is `None` and the extension does not have the trigger attribute, the extension is triggered at every iteration by default. If the trigger is not callable, it is passed to `IntervalTrigger` to build an interval trigger.
- **priority** (*int*) – Invocation priority of the extension. Extensions are invoked in the descending order of priorities in each iteration. If this is `None`, `extension.priority` is used instead.
- **invoke_before_training** (*bool or None*) – If `True`, the extension is also invoked just before entering the training loop. If this is `None`, `extension.invoke_before_training` is used instead. This option is mainly used for extensions that alter the training configuration (e.g., learning rates); in such a case, resuming from snapshots require the call of extension to recover the configuration before any updates.

get_extension(*name*)
Returns the extension of a given name.

Parameters **name** (*str*) – Name of the extension.

Returns Extension.

run()
Executes the training loop.

This method is the core of `Trainer`. It executes the whole loop of training the models.

Note that this method cannot run multiple times for one trainer object.

Updater

class chainer.training.**Updater**
Interface of updater objects for trainers.

TODO(beam2d): document it.

connect_trainer(*trainer*)
Connects the updater to the trainer that will call it.

The typical usage of this method is to register additional links to the reporter of the trainer. This method is called at the end of the initialization of *Trainer*. The default implementation does nothing.

Parameters **trainer** (Trainer) – Trainer object to which the updater is registered.

finalize()
Finalizes the updater object.

This method is called at the end of training loops. It should finalize each dataset iterator used in this updater.

get_all_optimizers()
Gets a dictionary of all optimizers for this updater.

Returns Dictionary that maps names to optimizers.

Return type dict

get_optimizer(*name*)
Gets the optimizer of given name.

Updater holds one or more optimizers with names. They can be retrieved by this method.

Parameters **name** (*str*) – Name of the optimizer.

Returns Optimizer of the name.

Return type *Optimizer*

serialize(*serializer*)
Serializes the current state of the updater object.

update()
Updates the parameters of the target model.

This method implements an update formula for the training task, including data loading, forward/backward computations, and actual updates of parameters.

This method is called once at each iteration of the training loop.

class chainer.training.**StandardUpdater**(*iterator*, *optimizer*, *converter=<function concat_examples>*, *device=None*, *loss_func=None*)
Standard implementation of Updater.

This is the standard implementation of *Updater*. It accepts one or more training datasets and one or more optimizers. The default update routine assumes that there is only one training dataset and one optimizer. Users can override this update routine by inheriting this class and overriding the update_core() method. Each batch is converted to input arrays by concat_examples() by default, which can also be manually set by converter argument.

Parameters

- **iterator** – Dataset iterator for the training dataset. It can also be a dictionary of iterators. If this is just an iterator, then the iterator is registered by the name 'main'.
- **optimizer** – Optimizer to update parameters. It can also be a dictionary of optimizers. If this is just an optimizer, then the optimizer is registered by the name 'main'.
- **converter** – Converter function to build input arrays. Each batch extracted by the main iterator and the device option are passed to this function. *concat_examples()* is used by default.
- **device** – Device to which the training data is sent. Negative value indicates the host memory (CPU).
- **loss_func** – Loss function. The target link of the main optimizer is used by default.

Variables

- **converter** – Converter function.
- **loss_func** – Loss function. If it is None, the target link of the main optimizer is used instead.
- ***device*** – Device to which the training data is sent.
- **iteration** – Current number of completed updates.

get_iterator(*name*)
Gets the dataset iterator of given name.

Parameters **name** (*str*) – Name of the dataset iterator.

Returns Corresponding dataset iterator.

Return type *Iterator*

class chainer.training.**ParallelUpdater**(*iterator*, *optimizer*, *converter=<function concat_examples>*, *models=None*, *devices=None*, *loss_func=None*)
Implementation of a parallel GPU Updater.

This is an implementation of *Updater* that uses multiple GPUs. It behaves similarly to *StandardUpdater*. The update routine is modified to support data-parallel computation on multiple GPUs in one machine. It is based on synchronous parallel SGD: it parallelizes the gradient computation over a mini-batch, and updates the parameters only in the main device.

Parameters

- **iterator** – Dataset iterator for the training dataset. It can also be a dictionary of iterators. If this is just an iterator, then the iterator is registered by the name 'main'.
- **optimizer** – Optimizer to update parameters. It can also be a dictionary of optimizers. If this is just an optimizer, then the optimizer is registered by the name 'main'.
- **converter** – Converter function to build input arrays. Each batch extracted by the main iterator is split equally between the devices and then passed with corresponding device option to this function. *concat_examples()* is used by default.
- **models** – Dictionary of models. The main model should be the same model attached to the 'main' optimizer.
- **devices** – Dictionary of devices to which the training data is sent. The devices should be arranged in a dictionary with the same structure as models.
- **loss_func** – Loss function. The model is used as a loss function by default.

Extension

class chainer.training.**Extension**
Base class of trainer extensions.

Extension of *Trainer* is a callable object that takes the trainer object as the argument. It also provides some default configurations as its attributes, e.g. the default trigger and the default priority. This class provides a set of typical default values for these attributes.

There are three ways to define users' own extensions: inheriting this class, decorating closures by *make_extension()*, or using any callable including lambda functions as extensions. Decorator can slightly reduce the overhead and is much easier to use, while this class provides more flexibility (for example, it can have methods to configure the behavior). Using a lambda function allows one-line coding for simple purposes,

but users have to specify the configurations as arguments to *Trainer.extend()*. For a callable not inheriting this class, the default configurations of this class are used unless the user explicitly specifies them in *Trainer.extend()* method.

Variables

- **trigger** – Default value of trigger for this extension. It is set to `(1, 'iteration')` by default.
- **priority** – Default priority of the extension. It is set to `PRIORITY_READER` by default.
- **invoke_before_training** – Default flag to decide whether this extension should be invoked before the training starts. The default value is `False`.

default_name
Default name of the extension.

It is the name of the class by default. Implementation can override this property, or provide a class attribute to hide it.

finalize()
Finalizes the extension.

This method is called at the end of the training loop.

serialize(*serializer*)
Serializes the extension state.

It is called when a trainer that owns this extension is serialized. It serializes nothing by default.

chainer.training.**make_extension**(*trigger=None*, *default_name=None*, *priority=None*, *invoke_before_training=False*, *finalizer=None*)
Decorator to make given functions into trainer extensions.

This decorator just adds some attributes to a given function. The value of the attributes are given by the arguments of this decorator.

See *Extension* for details of trainer extensions. Most of the default values of arguments also follow those for this class.

Parameters

- **trigger** – Default trigger of the extension.
- **default_name** – Default name of the extension. The name of a given function is used by default.
- **priority** (*int*) – Default priority of the extension.
- **invoke_before_training** (*bool*) – Default flag to decide whether the extension should be invoked before any training.
- **finalizer** – Finalizer function of this extension. The finalizer is called at the end of the training loop.

Trigger

Trigger is a callable object to decide when to process some specific event within the training loop. It takes a Trainer object as the argument, and returns True if some event should be fired.

It is mainly used to determine when to call an extension. It is also used to determine when to quit the training loop.

chainer.training.**get_trigger**(*trigger*)
Gets a trigger object.

Trigger object is a callable that accepts a *Trainer* object as an argument and returns a boolean value. When it returns True, various kinds of events can occur depending on the context in which the trigger is used. For example, if the trigger is passed to the *Trainer* as the *stop trigger*, the training loop breaks when the trigger returns True. If the trigger is passed to the *extend()* method of a trainer, then the registered extension is invoked only when the trigger returns True.

This function returns a trigger object based on the argument. If `trigger` is already a callable, it just returns the trigger. If `trigger` is `None`, it returns a trigger that never fires. Otherwise, it passes the value to *IntervalTrigger*.

Parameters **trigger** – Trigger object. It can be either an already built trigger object (i.e., a callable object that accepts a trainer object and returns a bool value), or a tuple. In latter case, the tuple is passed to *IntervalTrigger*.

Returns `trigger` if it is a callable, otherwise a *IntervalTrigger* object made from `trigger`.

Debug mode

In debug mode, Chainer checks values of variables on runtime and shows more detailed error messages. It helps you to debug your programs. Instead it requires additional overhead time.

In debug mode, Chainer checks all results of forward and backward computation, and if it founds a NaN value, it raises `RuntimeError`. Some functions and links also check validity of input values.

chainer.**is_debug**()
Get the debug mode.

Returns Return `True` if Chainer is in debug mode.

Return type bool

chainer.**set_debug**(*debug*)
Set the debug mode.

Note: This method changes global state. When you use this method on multi-threading environment, it may affects other threads.

Parameters **debug** (*bool*) – New debug mode.

class chainer.**DebugMode**(*debug*)
Debug mode context.

This class provides a context manager for debug mode. When entering the context, it sets the debug mode to the value of *debug* parameter with memorizing its original value. When exiting the context, it sets the debug mode back to the original value.

Parameters **debug** (*bool*) – Debug mode used in the context.

FunctionSet (deprecated)

class chainer.**FunctionSet**(***links*)
Set of links (as "parameterized functions").

FunctionSet is a subclass of *Chain*. Function registration is done just by adding an attribute to object.

Deprecated since version v1.5: Use *Chain* instead.

Note: FunctionSet was used for manipulation of one or more parameterized functions. The concept of parameterized function is gone, and it has been replaced by *Link* and *Chain*.

---

collect_parameters()
Returns a tuple of parameters and gradients.

Returns Tuple (pair) of two tuples. The first element is a tuple of parameter arrays, and the second is a tuple of gradient arrays.

copy_parameters_from(*params*)
Copies parameters from another source without reallocation.

Parameters **params** (*Iterable*) – Iterable of parameter arrays.

gradients
Tuple of gradient arrays of all registered functions.

The order of gradients is consistent with *parameters()* property.

parameters
Tuple of parameter arrays of all registered functions.

The order of parameters is consistent with *parameters()* property.

Utilities

CUDA utilities

Device, context and memory management on CuPy.

Chainer uses CuPy (with very thin wrapper) to exploit the speed of GPU computation. Following modules and classes are imported to cuda module for convenience (refer to this table when reading chainer's source codes).

imported name	original name
chainer.cuda.cupy	*cupy*
chainer.cuda.ndarray	*cupy.ndarray*
chainer.cuda.cupy.cuda	cupy.cuda
chainer.cuda.Device	*cupy.cuda.Device*
chainer.cuda.Event	*cupy.cuda.Event*
chainer.cuda.Stream	*cupy.cuda.Stream*

Chainer replaces the default allocator of CuPy by its memory pool implementation. It enables us to reuse the device memory over multiple forward/backward computations, and temporary arrays for consecutive elementwise operations.

Devices

chainer.cuda.**get_device**(**args*)
Gets the device from a device object, an ID integer or an array object.

Note: This API is deprecated. Please use :method:‘cupy.cuda.get_device_from_id‘ or :method:‘cupy.cuda.get_device_from_array‘ instead.

This is a convenient utility to select a correct device if the type of `arg` is unknown (i.e., one can use this function on arrays that may be on CPU or GPU). The returned device object supports the context management protocol of Python for the *with* statement.

Parameters **args** – Values to specify a GPU device. The first device object, integer or *cupy.ndarray* object is used to select a device. If it is a device object, it is returned. If it is an integer, the corresponding device is returned. If it is a CuPy array, the device on which this array reside is returned. If any arguments are neither integers nor CuPy arrays, a dummy device object representing CPU is returned.

Returns Device object specified by given `args`.

See also:

See *cupy.cuda.Device* for the device selection not by arrays.

chainer.cuda.**get_device_from_id**(*device_id*)

Gets the device from an ID integer.

Parameters **device_id** (*int or None*) – The ID of the device which this function returns.

chainer.cuda.**get_device_from_array**(**arrays*)

Gets the device from a list of CuPy array or a single CuPy array.

The device on which the given CuPy array reside is returned.

Parameters **array** (*cupy.ndarray* or list of *cupy.ndarray*) – A CuPy array which this function returns the device corresponding to. If a list of :class:‘cupy.ndarray‘s are given, it returns the first device object of an array in the list.

CuPy array allocation and copy

Note: As of v1.3.0, the following array construction wrappers are marked as deprecated. Use the corresponding functions of the *cupy* module instead. The main difference of them is that the default dtype is changed from float32 to float64.

Deprecated functions	Recommended functions
chainer.cuda.empty	*cupy.empty()*
chainer.cuda.empty_like	*cupy.empty_like()*
chainer.cuda.zeros	*cupy.zeros()*
chainer.cuda.zeros_like	*cupy.zeros_like()*
chainer.cuda.ones	*cupy.ones()*
chainer.cuda.ones_like	*cupy.ones_like()*
chainer.cuda.full	*cupy.full()*
chainer.cuda.full_like	*cupy.full_like()*

chainer.cuda.**copy**(*array*, *out=None*, *out_device=None*, *stream=None*)

Copies a *cupy.ndarray* object using the default stream.

This function can copy the device array to the destination array on another device.

Parameters

- **array** (cupy.ndarray) – Array to be copied.

- **out** (cupy.ndarray) – Destination array. If it is not None, then out_device argument is ignored.
- **out_device** – Destination device specifier. Actual device object is obtained by passing this value to *get_device()*.
- **stream** (cupy.cuda.Stream) – CUDA stream.

Returns

Copied array.

If out is not specified, then the array is allocated on the device specified by out_device argument.

Return type *cupy.ndarray*

chainer.cuda.**to_cpu**(*array*, *stream=None*)

Copies the given GPU array to host CPU.

Parameters

- **array** – Array to be sent to CPU.
- **stream** (cupy.cuda.Stream) – CUDA stream.

Returns

Array on CPU.

If given array is already on CPU, then this function just returns array without performing any copy.

Return type numpy.ndarray

chainer.cuda.**to_gpu**(*array*, *device=None*, *stream=None*)

Copies the given CPU array to specified device.

Parameters

- **array** – Array to be sent to GPU.
- **device** – Device specifier.
- **stream** (cupy.cuda.Stream) – CUDA stream. If not None, the copy runs asynchronously.

Returns

Array on GPU.

If array is already on GPU, then this function just returns array without performing any copy. Note that this function does not copy *cupy.ndarray* into specified device.

Return type *cupy.ndarray*

Kernel definition utilities

chainer.cuda.**memoize**(*for_each_device=False*)

Makes a function memoizing the result for each argument and device.

This is a similar version of *cupy.memoize()*. The difference is that this function can be used in the global scope even if CUDA is not available. In such case, this function does nothing.

Note: This decorator acts as a dummy if CUDA is not available. It cannot be used for general purpose memoization even if `for_each_device` is set to False.

chainer.cuda.**clear_memo**()
Clears the memoized results for all functions decorated by memoize.

This function works like *cupy.clear_memo()* as a counterpart for *chainer.cuda.memoize()*. It can be used even if CUDA is not available. In such a case, this function does nothing.

chainer.cuda.**elementwise**()
Creates an elementwise kernel function.

This function uses *memoize()* to cache the kernel object, i.e. the resulting kernel object is cached for each argument combination and CUDA device.

The arguments are the same as those for *cupy.ElementwiseKernel*, except that the `name` argument is mandatory.

chainer.cuda.**reduce**()
Creates a global reduction kernel function.

This function uses *memoize()* to cache the resulting kernel object, i.e. the resulting kernel object is cached for each argument combination and CUDA device.

The arguments are the same as those for *cupy.ReductionKernel*, except that the `name` argument is mandatory.

CPU/GPU generic code support

chainer.cuda.**get_array_module**(**args*)
Gets an appropriate one from `numpy` or *cupy*.

This is almost equivalent to *cupy.get_array_module()*. The differences are that this function can be used even if CUDA is not available and that it will return their data arrays' array module for *Variable* arguments.

Parameters **args** – Values to determine whether NumPy or CuPy should be used.

Returns *cupy* or `numpy` is returned based on the types of the arguments.

Return type module

cuDNN support

chainer.cuda.**set_max_workspace_size**(*size*)
Sets the workspace size for cuDNN.

Check "cuDNN Library User Guide" for detail.

Parameters **size** – The workspace size for cuDNN.

chainer.cuda.**get_max_workspace_size**()
Gets the workspace size for cuDNN.

Check "cuDNN Library User Guide" for detail.

Returns The workspace size for cuDNN.

Return type int

Common algorithms

class chainer.utils.**WalkerAlias**(*probs*)
Implementation of Walker's alias method.

This method generates a random sample from given probabilities $p_1, \ldots, p_n$ in $O(1)$ time. It is more efficient than choice(). This class works on both CPU and GPU.

Parameters **probs** (*float list*) – Probabilities of entries. They are normalized with *sum(probs)*.

See: Wikipedia article

sample(*shape*)
Generates a random sample based on given probabilities.

Parameters **shape** (*tuple of int*) – Shape of a return value.

Returns Returns a generated array with the given shape. If a sampler is in CPU mode the return value is a numpy.ndarray object, and if it is in GPU mode the return value is a *cupy.ndarray* object.

to_gpu()
Make a sampler GPU mode.

Reporter

Reporter

class chainer.**Reporter**
Object to which observed values are reported.

Reporter is used to collect values that users want to watch. The reporter object holds a mapping from value names to the actually observed values. We call this mapping *observations*.

When a value is passed to the reporter, an object called *observer* can be optionally attached. In this case, the name of the observer is added as the prefix of the value name. The observer name should be registered beforehand.

See the following example:

```
>>> from chainer import Reporter, report, report_scope
>>>
>>> reporter = Reporter()
>>> observer = object()  # it can be an arbitrary (reference) object
>>> reporter.add_observer('my_observer:', observer)
>>> observation = {}
>>> with reporter.scope(observation):
...     reporter.report({'x': 1}, observer)
...
>>> observation
{'my_observer:x': 1}
```

There are also a global API to add values:

```
>>> observation = {}
>>> with report_scope(observation):
...     report({'x': 1}, observer)
...
```

```
>>> observation
{'my_observer:x': 1}
```

The most important application of Reporter is to report observed values from each link or chain in the training and validation procedures. *Trainer* and some extensions prepare their own Reporter object with the hierarchy of the target link registered as observers. We can use *report()* function inside any links and chains to report the observed values (e.g., training loss, accuracy, activation statistics, etc.).

Variables **observation** – Dictionary of observed values.

add_observer(*name*, *observer*)
Registers an observer of values.

Observer defines a scope of names for observed values. Values observed with the observer are registered with names prefixed by the observer name.

Parameters

- **name** (*str*) – Name of the observer.
- **observer** – The observer object. Note that the reporter distinguishes the observers by their object ids (i.e., `id(owner)`), rather than the object equality.

add_observers(*prefix*, *observers*)
Registers multiple observers at once.

This is a convenient method to register multiple objects at once.

Parameters

- **prefix** (*str*) – Prefix of each name of observers.
- **observers** – Iterator of name and observer pairs.

report(*values*, *observer=None*)
Reports observed values.

The values are written with the key, prefixed by the name of the observer object if given.

Parameters

- **values** (*dict*) – Dictionary of observed values.
- **observer** – Observer object. Its object ID is used to retrieve the observer name, which is used as the prefix of the registration name of the observed value.

scope(**args*, ***kwds*)
Creates a scope to report observed values to `observation`.

This is a context manager to be passed to `with` statements. In this scope, the observation dictionary is changed to the given one.

It also makes this reporter object current.

Parameters **observation** (*dict*) – Observation dictionary. All observations reported inside of the `with` statement are written to this dictionary.

chainer.**get_current_reporter**()
Returns the current reporter object.

chainer.**report**(*values*, *observer=None*)
Reports observed values with the current reporter object.

Any reporter object can be set current by the `with` statement. This function calls the `Report.report()` method of the current reporter. If no reporter object is current, this function does nothing.

Example

The most typical example is a use within links and chains. Suppose that a link is registered to the current reporter as an observer (for example, the target link of the optimizer is automatically registered to the reporter of the *Trainer*). We can report some values from the link as follows:

```
class MyRegressor(chainer.Chain):
    def __init__(self, predictor):
        super(MyRegressor, self).__init__(predictor=predictor)

    def __call__(self, x, y):
        # This chain just computes the mean absolute and squared
        # errors between the prediction and y.
        pred = self.predictor(x)
        abs_error = F.sum(F.abs(pred - y)) / len(x.data)
        loss = F.mean_squared_error(pred, y)

        # Report the mean absolute and squared errors.
        report({'abs_error': abs_error, 'squared_error': loss}, self)

        return loss
```

If the link is named `'main'` in the hierarchy (which is the default name of the target link in the *StandardUpdater*), these reported values are named `'main/abs_error'` and `'main/squared_error'`. If these values are reported inside the `Evaluator` extension, `'validation/'` is added at the head of the link name, thus the item names are changed to `'validation/main/abs_error'` and `'validation/main/squared_error'` (`'validation'` is the default name of the Evaluator extension).

Parameters

- **values** (*dict*) – Dictionary of observed values.
- **observer** – Observer object. Its object ID is used to retrieve the observer name, which is used as the prefix of the registration name of the observed value.

chainer.**report_scope**(**args*, ***kwds*)
Returns a report scope with the current reporter.

This is equivalent to `get_current_reporter().scope(observation)`, except that it does not make the reporter current redundantly.

Summary and DictSummary

class chainer.**Summary**
Online summarization of a sequence of scalars.

Summary computes the statistics of given scalars online.

add(*value*)
Adds a scalar value.

Parameters **value** – Scalar value to accumulate. It is either a NumPy scalar or a zero-dimensional array (on CPU or GPU).

compute_mean()
Computes the mean.

make_statistics()
Computes and returns the mean and standard deviation values.

Returns Mean and standard deviation values.

Return type tuple

class chainer.**DictSummary**
Online summarization of a sequence of dictionaries.

DictSummary computes the statistics of a given set of scalars online. It only computes the statistics for scalar values and variables of scalar values in the dictionaries.

add(*d*)
Adds a dictionary of scalars.

Parameters **d** (*dict*) – Dictionary of scalars to accumulate. Only elements of scalars, zero-dimensional arrays, and variables of zero-dimensional arrays are accumulated.

compute_mean()
Creates a dictionary of mean values.

It returns a single dictionary that holds a mean value for each entry added to the summary.

Returns Dictionary of mean values.

Return type dict

make_statistics()
Creates a dictionary of statistics.

It returns a single dictionary that holds mean and standard deviation values for every entry added to the summary. For an entry of name 'key', these values are added to the dictionary by names 'key' and 'key.std', respectively.

Returns Dictionary of statistics of all entries.

Return type dict

Experimental feature annotation

chainer.utils.**experimental**(*api_name*)
Declares that user is using an experimental feature.

The developer of an API can mark it as *experimental* by calling this function. When users call experimental APIs, FutureWarning is issued. The presentation of FutureWarning is disabled by setting chainer.disable_experimental_warning to True, which is False by default.

The basic usage is to call it in the function or method we want to mark as experimental along with the API name.

```
from chainer import utils

def f(x):
    utils.experimental('chainer.foo.bar.f')
    # concrete implementation of f follows

f(1)
```

```
... FutureWarning: chainer.foo.bar.f is experimental. The interface can change in␣
↪the future. ...
```

We can also make a whole class experimental. In that case, we should call this function in its __init__ method.

```
class C():
    def __init__(self):
      utils.experimental('chainer.foo.C')

C()
```

```
... FutureWarning: chainer.foo.C is experimental. The interface can change in the␣
↪future. ...
```

If we want to mark __init__ method only, rather than class itself, it is recommended that we explicitly feed its API name.

```
class D():
    def __init__(self):
        utils.experimental('D.__init__')

D()
```

```
...  FutureWarning: D.__init__ is experimental. The interface can change in the␣
↪future. ...
```

Currently, we do not have any sophisticated way to mark some usage of non-experimental function as experimental. But we can support such usage by explicitly branching it.

```
def g(x, experimental_arg=None):
    if experimental_arg is not None:
        utils.experimental('experimental_arg of chainer.foo.g')
```

Parameters **api_name** (*str*) – The name of an API marked as experimental.

Assertion and Testing

Chainer provides some facilities to make debugging easy.

Type checking utilities

Function uses a systematic type checking of the *chainer.utils.type_check* module. It enables users to easily find bugs of forward and backward implementations. You can find examples of type checking in some function implementations.

class chainer.utils.type_check.**Expr**(*priority*)

Abstract syntax tree of an expression.

It represents an abstract syntax tree, and isn't a value. You can get its actual value with *eval()* function, and get syntax representation with the __str__() method. Each comparison operator (e.g. ==) generates a new *Expr* object which represents the result of comparison between two expressions.

Example

Let x and y be instances of *Expr*, then

```
>>> x = Variable(1, 'x')
>>> y = Variable(1, 'y')
>>> c = (x == y)
```

is also an instance of *Expr*. To evaluate and get its value, call *eval()* method:

```
>>> c.eval()
True
```

Call `str` function to get a representation of the original equation:

```
>>> str(c)
'x == y'
```

You can actually compare an expression with a value:

```
>>> (x == 1).eval()
True
```

Note that you can't use boolean operators such as `and`, as they try to cast expressions to boolean values:

```
>>> z = Variable(1, 'z')
>>> x == y and y == z  # raises an error
Traceback (most recent call last):
RuntimeError: Don't convert Expr to bool. Please call Expr.eval method to␣
↪evaluate expression.
```

---

eval()
Evaluates the tree to get actual value.

Behavior of this function depends on an implementation class. For example, a binary operator + calls the __add__ function with the two results of *eval()* function.

chainer.utils.type_check.**expect**(**bool_exprs*)
Evaluates and tests all given expressions.

This function evaluates given boolean expressions in order. When at least one expression is evaluated as `False`, that means the given condition is not satisfied. You can check conditions with this function.

Parameters **bool_exprs** (*tuple of Bool expressions*) – Bool expressions you want to evaluate.

class chainer.utils.type_check.**TypeInfo**(*shape*, *dtype*)
Type information of an input/gradient array.

It contains type information of an array, such as the shape of array and the number of dimensions. This information is independent of CPU or GPU array.

class chainer.utils.type_check.**TypeInfoTuple**
Type information of input/gradient tuples.

It is a sub-class of tuple containing *TypeInfo*. The i-th element of this object contains type information of the i-th input/gradient data. As each element is *Expr*, you can easily check its validity.

size()
Returns an expression representing its length.

Returns An expression object representing length of the tuple.

Return type *Expr*

Gradient checking utilities

Most function implementations are numerically tested by *gradient checking*. This method computes numerical gradients of forward routines and compares their results with the corresponding backward routines. It enables us to make the source of issues clear when we hit an error of gradient computations. The *chainer.gradient_check* module makes it easy to implement the gradient checking.

chainer.gradient_check.**check_backward**(*func*, *x_data*, *y_grad*, *params=()*, *eps=0.001*, *atol=1e-05*, *rtol=0.0001*, *no_grads=None*, *dtype=None*)

Test backward procedure of a given function.

This function automatically check backward-process of given function. For example, when you have a *Function* class `MyFunc`, that gets two arguments and returns one value, you can make its test like this:

```
>> def test_my_func(self):
>>   func = MyFunc()
>>   x1_data = xp.array(...)
>>   x2_data = xp.array(...)
>>   gy_data = xp.array(...)
>>   check_backward(func, (x1_data, x2_data), gy_data)
```

This method creates *Variable* objects with `x_data` and calls `func` with the *Variable* s to get its result as *Variable*. Then, it sets `y_grad` array to `grad` attribute of the result and calls `backward` method to get gradients of the inputs. To check correctness of the gradients, the function calls *numerical_grad()* to calculate numerically the gradients and compares the types of gradients with *chainer.testing.assert_allclose()*. If input objects (`x1_data` or/and `x2_data` in this example) represent integer variables, their gradients are ignored.

You can simplify a test when `MyFunc` gets only one argument:

```
>>   check_backward(func, x1_data, gy_data)
```

If `MyFunc` is a loss function which returns a zero-dimensional array, pass `None` to `gy_data`. In this case, it sets `1` to `grad` attribute of the result:

```
>>   check_backward(my_loss_func, (x1_data, x2_data), None)
```

If `MyFunc` returns multiple outputs, pass all gradients for outputs as a tuple:

```
>>   gy1_data = xp.array(...)
>>   gy2_data = xp.array(...)
>>   check_backward(func, x1_data, (gy1_data, gy2_data))
```

You can also test a *Link*. To check gradients of parameters of the link, set a tuple of the parameters to `params` arguments:

```
>>   check_backward(my_link, (x1_data, x2_data), gy_data,
>>                  (my_link.W, my_link.b))
```

Note that `params` are not `ndarray` s, but `Variables` s.

Function objects are acceptable as `func` argument:

```
>>   check_backward(lambda x1, x2: f(x1, x2),
>>                  (x1_data, x2_data), gy_data)
```

Note: `func` is called many times to get numerical gradients for all inputs. This function doesn't work correctly when `func` behaves randomly as it gets different gradients.

Parameters

- **func** (*callable*) – A function which gets *Variable* s and returns *Variable* s. `func` must returns a tuple of *Variable* s or one *Variable*. You can use *Function* object, *Link* object or a function satisfying the condition.
- **x_data** (*ndarray or tuple of ndarrays*) – A set of `ndarray` s to be passed to `func`. If `x_data` is one `ndarray` object, it is treated as `(x_data,)`.
- **y_grad** (*ndarray or tuple of ndarrays or None*) – A set of `ndarray` s representing gradients of return-values of `func`. If `y_grad` is one `ndarray` object, it is treated as `(y_grad,)`. If `func` is a loss-function, `y_grad` should be set to `None`.
- **params** (*Variable or tuple of ~chainder.Variable*) – A set of *Variable* s whose gradients are checked. When `func` is a *Link* object, set its parameters as `params`. If `params` is one *Variable* object, it is treated as `(params,)`.
- **eps** (*float*) – Epsilon value to be passed to *numerical_grad()*.
- **atol** (*float*) – Absolute tolerance to be passed to *chainer.testing.assert_allclose()*.
- **rtol** (*float*) – Relative tolerance to be passed to *chainer.testing.assert_allclose()*.
- **no_grads** (*list of bool*) – Flag to skip variable for gradient assertion. It should be same length as `x_data`.
- **dtype** (*dtype*) – `x_data` and `y_grad` are casted to this dtype when calculating numerical gradients. Only float types and `None` are allowed.

See: *numerical_grad()*

`chainer.gradient_check.`**numerical_grad**(*f*, *inputs*, *grad_outputs*, *eps=0.001*)

Computes numerical gradient by finite differences.

This function is used to implement gradient check. For usage example, see unit tests of *chainer.functions*.

Parameters

- **f** (`function`) – Python function with no arguments that runs forward computation and returns the result.
- **inputs** (*tuple of arrays*) – Tuple of arrays that should be treated as inputs. Each element of them is slightly modified to realize numerical gradient by finite differences.
- **grad_outputs** (*tuple of arrays*) – Tuple of arrays that are treated as output gradients.
- **eps** (*float*) – Epsilon value of finite differences.

Returns Numerical gradient arrays corresponding to `inputs`.

Return type tuple

Standard Assertions

The assertions have same names as NumPy's ones. The difference from NumPy is that they can accept both `numpy.ndarray` and *`cupy.ndarray`*.

`chainer.testing.`**`assert_allclose`**(*x*, *y*, *atol=1e-05*, *rtol=0.0001*, *verbose=True*)

Asserts if some corresponding element of x and y differs too much.

This function can handle both CPU and GPU arrays simultaneously.

Parameters

- **`x`** – Left-hand-side array.
- **`y`** – Right-hand-side array.
- **`atol`** (*`float`*) – Absolute tolerance.
- **`rtol`** (*`float`*) – Relative tolerance.
- **`verbose`** (*`bool`*) – If `True`, it outputs verbose messages on error.

Function testing utilities

Chainer provides some utilities for testing its functions.

`chainer.testing.`**`unary_math_function_unittest`**(*func*, *func_expected=None*, *label_expected=None*, *make_data=None*)

Decorator for testing unary mathematical Chainer functions.

This decorator makes test classes test unary mathematical Chainer functions. Tested are forward and backward computations on CPU and GPU across parameterized `shape` and `dtype`.

Parameters

- **`func`** (`Function`) – Chainer function to be tested by the decorated test class.
- **`func_expected`** – Function used to provide expected values for testing forward computation. If not given, a corresponsing numpy function for `func` is implicitly picked up by its class name.
- **`label_expected`** (*`string`*) – String used to test labels of Chainer functions. If not given, the class name of `func` lowered is implicitly used.
- **`make_data`** – Function to customize input and gradient data used in the tests. It takes `shape` and `dtype` as its arguments, and returns a tuple of input and gradient data. By default, uniform destribution ranged `[-1, 1]` is used for both.

The decorated test class tests forward and backward computations on CPU and GPU across the following `parameterize()` ed parameters:

•shape: rank of zero, and rank of more than zero

•dtype: `numpy.float16`, `numpy.float32` and `numpy.float64`

Additionally, it tests the label of the Chainer function.

Chainer functions tested by the test class decorated with the decorator should have the following properties:

•Unary, taking one parameter and returning one value

•`dtype` of input and output are the same

•Elementwise operation for the supplied ndarray

Example

The following code defines a test class that tests *sin()* Chainer function, which takes a parameter with `dtype` of float and returns a value with the same `dtype`.

```
>>> import unittest
>>> from chainer import testing
>>> from chainer import functions as F
>>>
>>> @testing.unary_math_function_unittest(F.Sin())
... class TestSin(unittest.TestCase):
...     pass
```

Because the test methods are implicitly injected to `TestSin` class by the decorator, it is enough to place `pass` in the class definition.

Now the test is run with `nose` module.

```
>>> import nose
>>> nose.run(
...     defaultTest=__name__, argv=['', '-a', '!gpu'], exit=False)
True
```

To customize test data, `make_data` optional parameter can be used. The following is an example of testing `sqrt` Chainer function, which is tested in positive value domain here instead of the default input.

```
>>> import numpy
>>>
>>> def make_data(shape, dtype):
...     x = numpy.random.uniform(0.1, 1, shape).astype(dtype)
...     gy = numpy.random.uniform(-1, 1, shape).astype(dtype)
...     return x, gy
...
>>> @testing.unary_math_function_unittest(F.Sqrt(),
...                                       make_data=make_data)
... class TestSqrt(unittest.TestCase):
...     pass
...
>>> nose.run(
...     defaultTest=__name__, argv=['', '-a', '!gpu'], exit=False)
True
```

`make_data` function which returns input and gradient data generated in proper value domains with given `shape` and `dtype` parameters is defined, then passed to the decorator's `make_data` parameter.

Standard Function implementations

Chainer provides basic *Function* implementations in the *chainer.functions* package. Most of them are wrapped by plain Python functions, which users should use.

Note: As of v1.5, the concept of parameterized functions are gone, and they are replaced by corresponding *Link* implementations. They are still put in the *functions* namespace for backward compatibility, though it is strongly recommended to use them via the *chainer.links* package.

Activation functions

clipped_relu

chainer.functions.**clipped_relu**(*x*, *z=20.0*)

Clipped Rectifier Unit function.

For a clipping value $z(> 0)$, it computes

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_n)$-shaped float array.
- **z** (*float*) – Clipping value. (default = 20.0)

Returns Output variable. A $(s_1, s_2, ..., s_n)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.random.uniform(-100, 100, (10, 20)).astype('f')
>>> z = 10.0
>>> np.any(x < 0)
True
>>> np.any(x > z)
True
>>> y = F.clipped_relu(x, z=z)
>>> np.any(y.data < 0)
False
>>> np.any(y.data > z)
False
```

crelu

chainer.functions.**crelu**(*x*, *axis=1*)

Concatenated Rectified Linear Unit function.

This function is expressed as follows

$$f(x) = (\max(0, x), \max(0, -x)).$$

Here, two output values are concatenated along an axis.

See: https://arxiv.org/abs/1603.05201

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **axis** (*int*) – Axis that the output values are concatenated along. Default is 1.

Returns Output variable of concatenated array. If the axis is 1, A $(s_1, s_2 \times 2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.array([[-1, 0], [2, -3]], 'f')
>>> x
array([[-1.,  0.],
       [ 2., -3.]], dtype=float32)
>>> y = F.crelu(x, axis=1)
>>> y.data
array([[ 0.,  0.,  1.,  0.],
       [ 2.,  0.,  0.,  3.]], dtype=float32)
```

elu

chainer.functions.**elu**(*x*, *alpha=1.0*)

Exponential Linear Unit function.

For a parameter α, it is expressed as

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(\exp(x) - 1) & \text{if } x < 0, \end{cases}$$

See: https://arxiv.org/abs/1511.07289

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **alpha** (*float*) – Parameter α. Default is 1.0.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.array([[-1, 0], [2, -3]], 'f')
>>> x
array([[-1.,  0.],
       [ 2., -3.]], dtype=float32)
>>> y = F.elu(x, alpha=1.)
>>> y.data
array([[-0.63212055,  0.        ],
       [ 2.        , -0.95021296]], dtype=float32)
```

hard_sigmoid

chainer.functions.**hard_sigmoid**(*x*)

Element-wise hard-sigmoid function.

This function is defined as

$$f(x) = \begin{cases} 0 & \text{if } x < -2.5 \\ 0.2x + 0.5 & \text{if } -2.5 < x < 2.5 \\ 1 & \text{if } 2.5 < x. \end{cases}$$

Parameters **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

It maps the input values into the range of $[0, 1]$.

```
>>> x = np.array([-2.6, -1, 0, 1, 2.6])
>>> x
array([-2.6, -1. ,  0. ,  1. ,  2.6])
>>> F.hard_sigmoid(x).data
array([ 0. ,  0.3,  0.5,  0.7,  1. ])
```

leaky_relu

chainer.functions.**leaky_relu**(*x*, *slope=0.2*)

Leaky Rectified Linear Unit function.

This function is expressed as

$$f(x) = \max(x, ax),$$

where a is a configurable slope value.

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **slope** (*float*) – Slope value a.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.array([[-1, 0], [2, -3], [-2, 1]], 'f')
>>> x
array([[-1.,  0.],
       [ 2., -3.],
       [-2.,  1.]], dtype=float32)
>>> F.leaky_relu(x, slope=0.2).data
array([[-0.2       ,  0.        ],
       [ 2.        , -0.60000002],
       [-0.40000001,  1.        ]], dtype=float32)
```

log_softmax

chainer.functions.**log_softmax**(*x*, *use_cudnn=True*)

Channelwise log-softmax function.

This function computes its logarithm of softmax along the second axis. Let $x = (x_1, x_2, \ldots, x_d)^\top$ be the d dimensional index array and $f(x_1, \ldots, x_d)$ be the corresponding input array. For each index x in the input array, it computes the logarithm of the probability $\log p(x)$ defined as

$$p(x) = \frac{\exp(f(x_1, x_2, \ldots, x_d))}{\sum_{x'_2} \exp(f(x_1, x'_2, \ldots, x_d))}.$$

This method is theoretically equivalent to `log(softmax(x))` but is more stable.

Note: `log(softmax(x))` may cause underflow when `x` is too small, because `softmax(x)` may returns `0`. `log_softmax` method is more stable.

Parameters

- **x** (*Variable*) – Input variable.
- **use_cudnn** (*bool*) – If `True`, cuDNN is enabled and cuDNN ver. 3 or later is used, then this function uses cuDNN as the core implementation.

Returns Output variable.

Return type *Variable*

See also:

softmax()

lstm

chainer.functions.**lstm**(*c_prev*, *x*)

Long Short-Term Memory units as an activation function.

This function implements LSTM units with forget gates. Let the previous cell state c_{prev} and the incoming signal x.

First, the incoming signal x is split into four arrays a, i, f, o of the same shapes along the second axis. It means that x 's second axis must have 4 times the length of c_{prev}.

The split input signals are corresponding to:

- a : sources of cell input
- i : sources of input gate
- f : sources of forget gate
- o : sources of output gate

Second, it computes outputs as:

$$c = \tanh(a)\text{sigmoid}(i) + c_{\text{prev}}\text{sigmoid}(f),$$
$$h = \tanh(c)\text{sigmoid}(o).$$

These are returned as a tuple of two variables.

This function supports variable length inputs. The mini-batch size of the current input must be equal to or smaller than that of the previous one. When mini-batch size of `x` is smaller than that of `c`, this function only updates `c[0:len(x)]` and doesn't change the rest of `c`, `c[len(x):]`. So, please sort input sequences in descending order of lengths before applying the function.

Parameters

- **c_prev** (*Variable*) – Variable that holds the previous cell state. The cell state should be a zero array or the output of the previous call of LSTM.
- **x** (*Variable*) – Variable that holds the incoming signal. It must have the second dimension four times of that of the cell state,

Returns

Two *Variable* objects c and h. c is the updated cell state. `h` indicates the outgoing signal.

Return type tuple

See the original paper proposing LSTM with forget gates: Long Short-Term Memory in Recurrent Neural Networks.

Example

Assuming `y` is the current input signal, `c` is the previous cell state, and `h` is the previous output signal from an `lstm` function. Each of `y`, `c` and `h` has `n_units` channels. Most typical preparation of `x` is:

```
>>> n_units = 100
>>> y = chainer.Variable(np.zeros((1, n_units), 'f'))
>>> h = chainer.Variable(np.zeros((1, n_units), 'f'))
>>> c = chainer.Variable(np.zeros((1, n_units), 'f'))
>>> model = chainer.Chain(w=L.Linear(n_units, 4 * n_units),
...                       v=L.Linear(n_units, 4 * n_units),)
>>> x = model.w(y) + model.v(h)
>>> c, h = F.lstm(c, x)
```

It corresponds to calculate the input sources a, i, f, o from the current input `y` and the previous output `h`. Different parameters are used for different kind of input sources.

---

maxout

chainer.functions.**maxout**(*x*, *pool_size*, *axis=1*)

Maxout activation function.

It accepts an input tensor `x`, reshapes the `axis` dimension (say the size being `M * pool_size`) into two dimensions `(M, pool_size)`, and takes maximum along the `axis` dimension.

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A n-dimensional ($n \geq$ `axis`) float array. In general, its first dimension is assumed to be the *minibatch dimension*. The other dimensions are treated as one concatenated dimension.
- **pool_size** (*int*) – The size used for downsampling of pooling layer.
- **axis** (*int*) – The `axis` dimension to be reshaped. The size of `axis` dimension should be `M * pool_size`.

Returns Output variable. The shape of the output is same as `x` except that `axis` dimension is transformed from `M * pool_size` to `M`.

Return type *Variable*

See also:

Maxout

Example

Typically, x is the output of a linear layer or a convolution layer. The following is the example where we use *maxout()* in combination with a Linear link.

```
>>> in_size, out_size, pool_size = 10, 10, 10
>>> bias = np.arange(out_size * pool_size).astype('f')
>>> l = L.Linear(in_size, out_size * pool_size, initial_bias=bias)
>>> x = np.zeros((1, in_size), 'f')  # prepare data
>>> x = l(x)
>>> y = F.maxout(x, pool_size)
>>> x.shape
(1, 100)
>>> y.shape
(1, 10)
>>> x.reshape((out_size, pool_size)).data
array([[  0.,   1.,   2.,   3.,   4.,   5.,   6.,   7.,   8.,   9.],
       [ 10.,  11.,  12.,  13.,  14.,  15.,  16.,  17.,  18.,  19.],
       [ 20.,  21.,  22.,  23.,  24.,  25.,  26.,  27.,  28.,  29.],
       [ 30.,  31.,  32.,  33.,  34.,  35.,  36.,  37.,  38.,  39.],
       [ 40.,  41.,  42.,  43.,  44.,  45.,  46.,  47.,  48.,  49.],
       [ 50.,  51.,  52.,  53.,  54.,  55.,  56.,  57.,  58.,  59.],
       [ 60.,  61.,  62.,  63.,  64.,  65.,  66.,  67.,  68.,  69.],
       [ 70.,  71.,  72.,  73.,  74.,  75.,  76.,  77.,  78.,  79.],
       [ 80.,  81.,  82.,  83.,  84.,  85.,  86.,  87.,  88.,  89.],
       [ 90.,  91.,  92.,  93.,  94.,  95.,  96.,  97.,  98.,  99.]],␣
↪dtype=float32)
>>> y.data
array([[  9.,  19.,  29.,  39.,  49.,  59.,  69.,  79.,  89.,  99.]],␣
↪dtype=float32)
```

---

prelu

chainer.functions.**prelu**(*x*, *W*)

Parametric ReLU function.

It accepts two arguments: an input x and a weight array W and computes the output as $PReLU(x) = \max(x, W * x)$, where $*$ is an elementwise multiplication for each sample in the batch.

When the PReLU function is combined with two-dimensional convolution, the elements of parameter a are typically shared across the same filter of different pixels. In order to support such usage, this function supports the shape of parameter array that indicates leading dimensions of input arrays except the batch dimension.

For example W has the shape of $(2, 3, 4)$, x must have the shape of $(B, 2, 3, 4, S1, ..., SN)$ where B is batch size and the number of trailing S's is arbitrary non-negative integer.

Parameters

- **x** (Variable) – Input variable. Its first argument is assumed to be the minibatch dimension.
- **W** (Variable) – Weight variable.

Returns Output variable

Return type *Variable*

See also:

PReLU

relu

chainer.functions.**relu**(*x*, *use_cudnn=True*)

Rectified Linear Unit function.

$$f(x) = \max(0, x).$$

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **use_cudnn** (*bool*) – If True and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.array([[-1, 0], [2, -3], [-2, 1]], 'f')
>>> np.any(x < 0)
True
>>> y = F.relu(x)
>>> np.any(y.data < 0)
False
>>> y.shape
(3, 2)
```

sigmoid

chainer.functions.**sigmoid**(*x*, *use_cudnn=True*)

Element-wise sigmoid logistic function.

$$f(x) = (1 + \exp(-x))^{-1}.$$

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **use_cudnn** (*bool*) – If True and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

It maps the input values into the range of $[0, 1]$.

```
>>> x = np.arange(-2, 3, 2).astype('f')
>>> x
array([-2.,  0.,  2.], dtype=float32)
>>> F.sigmoid(x).data
array([ 0.11920291,  0.5       ,  0.88079709], dtype=float32)
```

slstm

chainer.functions.**slstm**(*c_prev1*, *c_prev2*, *x1*, *x2*)

S-LSTM units as an activation function.

This function implements S-LSTM unit. It is an extension of LSTM unit applied to tree structures. The function is applied to binary trees. Each node has two child nodes. It gets four arguments, previous cell states c_1 and c_2, and incoming signals x_1 and x_2.

First both input signals x_1 and x_2 are split into eight arrays a_1, i_1, f_1, o_1, and a_2, i_2, f_2, o_2. They have the same shape along the second axis. It means that x_1 and x_2 's second axis must have 4 times the length of $c_{1\text{prev}}$ and $c_{2\text{prev}}$.

The split input signals are corresponding to:

- a_i : sources of cell input
- i_i : sources of input gate
- f_i : sources of forget gate
- o_i : sources of output gate

It computes outputs as:

$$c = \tanh(a_1 + a_2)\sigma(i_1 + i_2) + c_{1\text{prev}}\sigma(f_1) + c_{2\text{prev}}\sigma(f_2),$$
$$h = \tanh(c)\sigma(o_1 + o_2),$$

where σ is the elementwise sigmoid function. The function returns c and h as a tuple.

Parameters

- **c_prev1** (Variable) – Variable that holds the previous cell state of the first child node. The cell state should be a zero array or the output of the previous call of LSTM.
- **c_prev2** (Variable) – Variable that holds the previous cell state of the second child node.
- **x1** (Variable) – Variable that holds the incoming signal from the first child node. It must have the second dimension four times of that of the cell state,
- **x2** (Variable) – Variable that holds the incoming signal from the second child node.

Returns

Two *Variable* objects c and h. c is the cell state. h indicates the outgoing signal.

Return type tuple

See detail in paper: Long Short-Term Memory Over Tree Structures.

softmax

chainer.functions.**softmax**(*x*, *use_cudnn=True*)

Channelwise softmax function.

This function computes its softmax along the second axis. Let $x = (x_1, x_2, \ldots, x_d)^\top$ be the d dimensional index array and $f(x)$ be the d dimensional input array. For each index x of the input array $f(x)$, it computes the probability $p(x)$ defined as $p(x) = \frac{\exp(f(x))}{\sum_{x_2} \exp(f(x))}$.

Parameters

- **x** (*Variable*) – Input variable.
- **use_cudnn** (*bool*) – If True and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable.

Return type *Variable*

softplus

chainer.functions.**softplus**(*x*, *beta=1.0*)

Element-wise softplus function.

The softplus function is the smooth approximation of ReLU.

$$f(x) = \frac{1}{\beta} \log(1 + \exp(\beta x)),$$

where β is a parameter. The function becomes curved and akin to ReLU as the β is increasing.

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **beta** (*float*) – Parameter β.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.arange(-2, 3, 2).astype('f')
>>> x
array([-2.,  0.,  2.], dtype=float32)
>>> F.softplus(x, beta=1.0).data
array([ 0.126928  ,  0.69314718,  2.12692809], dtype=float32)
```

tanh

chainer.functions.**tanh**(*x*, *use_cudnn=True*)

Elementwise hyperbolic tangent function.

$$f(x) = \tanh(x).$$

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **use_cudnn** (*bool*) – If True and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable. A $(s_1, s_2, ..., s_N)$-shaped float array.

Return type *Variable*

Example

```
>>> x = np.arange(-1, 4, 2).astype('f')
>>> x
array([-1.,  1.,  3.], dtype=float32)
>>> F.tanh(x).data
array([-0.76159418,  0.76159418,  0.99505478], dtype=float32)
```

Array manipulations

broadcast

chainer.functions.**broadcast**(**args*)

Broadcast given variables.

Parameters **args** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variables to be broadcasted. Each dimension of the shapes of the input variables must have the same size.

Returns *Variable* or tuple of *Variable* objects which are broadcasted from given arguments.

Return type *Variable*

Example

```
>>> x = np.random.uniform(0, 1, (3, 2)).astype('f')
>>> y = F.broadcast(x)
>>> np.all(x == y.data)
True
>>> z = np.random.uniform(0, 1, (3, 2)).astype('f')
>>> y, w = F.broadcast(x, z)
>>> np.all(x == y.data) & np.all(z == w.data)
True
```

broadcast_to

chainer.functions.**broadcast_to**(*x*, *shape*)

Broadcast a given variable to a given shape.

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable be broadcasted. A $(s_1, s_2, ..., s_N)$-shaped float array.

- **shape** (*tuple*) – Tuple of `int` of the shape of the output variable.

Returns Output variable broadcasted to the given shape.

Return type *Variable*

Example

```
>>> x = np.arange(0, 3)
>>> x
array([0, 1, 2])
>>> y = F.broadcast_to(x, (3, 3))
>>> y.data
array([[0, 1, 2],
       [0, 1, 2],
       [0, 1, 2]])
```

cast

`chainer.functions.`**`cast`**(*x*, *typ*)

Cast an input variable to a given type.

Parameters

- **x** (*Variable* or `numpy.ndarray` or *cupy.ndarray*) – Input variable to be casted. A $(s_1, s_2, ..., s_N)$-shaped float array.
- **typ** (`str` of dtype or `numpy.dtype`) – Typecode or data type to cast.

Returns Variable holding a casted array.

Return type *Variable*

Example

```
>>> x = np.arange(0, 3, dtype=np.float64)
>>> x.dtype
dtype('float64')
>>> y = F.cast(x, np.float32)
>>> y.dtype
dtype('float32')
>>> y = F.cast(x, 'float16')
>>> y.dtype
dtype('float16')
```

concat

`chainer.functions.`**`concat`**(*xs*, *axis=1*)

Concatenates given variables along an axis.

Parameters

- **xs** (tuple of *Variable* or `numpy.ndarray` or *cupy.ndarray*) – Input variables to be concatenated. The variables must have the same shape, except in the dimension corresponding to axis.

- **axis** (*int*) – The axis along which the arrays will be joined. Default is 1.

Returns The concatenated variable.

Return type *Variable*

Example

```
>>> x = np.arange(0, 12).reshape(3, 4)
>>> x
array([[ 0,  1,  2,  3],
       [ 4,  5,  6,  7],
       [ 8,  9, 10, 11]])
>>> y = np.arange(0, 3).reshape(3, 1)
>>> y
array([[0],
       [1],
       [2]])
>>> z = F.concat((x, y), axis=1)
>>> z.data
array([[ 0,  1,  2,  3,  0],
       [ 4,  5,  6,  7,  1],
       [ 8,  9, 10, 11,  2]])
```

copy

chainer.functions.**copy**(*x*, *dst*)

Copies the input variable onto the specified device.

This function copies the array of input variable onto the device specified by `dst`. When `dst == -1`, it copies the array onto the host memory. This function supports copies from host to device, from device to device and from device to host.

Parameters

- **x** (`Variable`) – Variable to be copied.
- **dst** – Target device specifier.

Returns Output variable.

Return type *Variable*

depth2space

chainer.functions.**depth2space**(*X*, *r*)

Computes the depth2space transformation for subpixel calculations.

Parameters

- **X** (`Variable`) – Variable holding a 4d array of
- **shape** (*batch, channel * r, dim1, dim2*) –
- **r** (*int*) – int specifying the upscaling factor.

Returns A variable holding the upscaled array from interspersed depth layers.

Return type *Variable*

Note: This can be used to compute super-resolution transformations. See https://arxiv.org/abs/1609.05158 for details.

dstack

chainer.functions.**dstack**(*xs*)
Concatenate variables along third axis (depth wise).

Parameters **xs** (*list of chainer.Variable*) – Variables to be concatenated.

Returns Output variable.

Return type *Variable*

expand_dims

chainer.functions.**expand_dims**(*x*, *axis*)
Expands dimensions of an input variable without copy.

Parameters

- **x** (Variable) – Input variable.
- **axis** (*int*) – Position where new axis is to be inserted.

Returns Variable that holds a expanded input.

Return type *Variable*

flatten

chainer.functions.**flatten**(*x*)
Flatten a given array.

Parameters **x** (*Varaiable*) – Input variable.

Returns Output variable.

Return type *Variable*

fliplr

chainer.functions.**fliplr**(*a*)
Flip array in the left/right direction.

Parameters **xs** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

flipud

chainer.functions.**flipud**(*a*)
Flip array in the up/down direction.

Parameters **xs** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

get_item

chainer.functions.**get_item**(*x*, *slices*)
Extract elements from array with specified shape, axes and offsets.

Parameters

- **x** (Variable) – A variable to be sliced.
- **slices** (*int*, *slice*, *Ellipsis*, *None*, *integer array-like*, *boolean array-like or tuple of them*) – It is an integer, a slice, an ellipsis, a numpy.newaxis, an integer array-like, a boolean array-like or tuple of them.

Returns

***Variable* object** which contains sliced array of x.

Return type *Variable*

Note: It only supports types that are supported by CUDA's atomicAdd when an integer array is included in slices. The supported types are numpy.float32, numpy.int32, numpy.uint32, numpy.uint64 and numpy.ulonglong.

Note: It does not support slices that contains multiple boolean arrays.

Note: See NumPy document for details of indexing.

---

hstack

chainer.functions.**hstack**(*xs*)
Concatenate variables horizontally (column wise).

Parameters **xs** (*list of chainer.Variable*) – Variables to be concatenated.

Returns Output variable.

Return type *Variable*

pad

chainer.functions.**pad**(*x*, *pad_width*, *mode*, ***keywords*)
Pad an input variable.

Parameters

- **x** (chainer.Variable or :class:`numpy.ndarray` or cupy.ndarray) – Input data.
- **pad_width** (*int or array-like*) – Number of values padded to the edges of each axis.
- **mode** (*str*) – Specifies how the function fills the periphery of the array. *constant*

  Pads with a constant values.
- **constant_values** (*int or array-like*) – The values are padded for each axis.

Returns Output variable.

Return type *Variable*

permutate

`chainer.functions.`**`permutate`**(*x*, *indices*, *axis=0*, *inv=False*)

Permutates a given variable along an axis.

This function permutate `x` with given `indices`. That means `y[i] = x[indices[i]]` for all `i`. Note that this result is same as `y = x.take(indices)`. `indices` must be a permutation of `[0, 1, ..., len(x) - 1]`.

When `inv` is `True`, `indices` is treated as its inverse. That means `y[indices[i]] = x[i]`.

Parameters

- **x** (`Variable`) – Variable to permutate.
- **indices** (`Variable`) – Indices to extract from the variable.
- **axis** (*int*) – Axis that the input array is permutate along.
- **inv** (*bool*) – If `True`, `indices` is treated as its inverse.

Returns Output variable.

Return type *Variable*

reshape

`chainer.functions.`**`reshape`**(*x*, *shape*)

Reshapes an input variable without copy.

Parameters

- **x** (`Variable`) – Input variable.
- **shape** (*tuple of ints*) – Target shape.

Returns

Variable that holds a reshaped version of the input variable.

Return type *Variable*

resize_images

chainer.functions.**resize_images**(*x*, *output_shape*)
Resize images to the given shape.

This function resizes 2D data to `output_shape`. Currently, only bilinear interpolation is supported as the sampling method.

Notatition: here is a notation for dimensionalities.

- n is the batch size.
- c_I is the number of the input channels.
- h and w are the height and width of the input image, respectively.
- h_O and w_O are the height and width of the output image.

Parameters

- **x** (Variable) – Input variable of shape (n, c_I, h, w).
- **output_shape** (*tuple*) – This is a tuple of length 2 whose values are `(h_O, w_O)`. Note that the order of height and width is opposite of the one in OpenCV.

Returns Resized image whose shape is (n, c_I, h_O, w_O).

Return type *Variable*

rollaxis

chainer.functions.**rollaxis**(*x*, *axis*, *start=0*)
Roll the axis backwards to the given position.

Parameters

- **x** (Variable) – Input variable.
- **axis** (*int*) – The axis to roll backwards.
- **start** (*int*) – The place to which the axis is moved.

Returns Variable whose axis is rolled.

Return type *Variable*

select_item

chainer.functions.**select_item**(*x*, *t*)
Select elements stored in given indices.

This function returns `t.choose(x.T)`, that means `y[i] == x[i, t[i]]` for all `i`.

Parameters

- **x** (Variable) – Variable storing arrays.
- **t** (Variable) – Variable storing index numbers.

Returns Variable that holds `t`-th element of `x`.

Return type *Variable*

separate

chainer.functions.**separate**(*x*, *axis=0*)

Separates an array along a given axis.

This function separates an array along a given axis. For example, shape of an array is `(2, 3, 4)`. When it separates the array with `axis=1`, it returns three `(2, 4)` arrays.

This function is an inverse of *chainer.functions.stack()*.

Parameters

- **x** (chainer.Variable) – Variable to be separated.
- **axis** (*int*) – Axis along which variables are separated.

Returns Output variables.

Return type tuple of chainer.Variable

See also:

chainer.functions.stack()

space2depth

chainer.functions.**space2depth**(*X*, *r*)

Computes the space2depth transformation for subpixel calculations.

Parameters

- **X** (Variable) – Variable holding a 4d array of
- **shape** (*batch, channel, dim1, dim2*) –
- **r** (*int*) – int specifying the upscaling factor.

Returns A variable holding the downscaled layer array from subpixel array sampling.

Return type *Variable*

Note: This can be used to compute inverse super-resolution transformations. See https://arxiv.org/abs/1609.05158 for details.

---

split_axis

chainer.functions.**split_axis**(*x*, *indices_or_sections*, *axis*, *force_tuple=False*)

Splits given variables along an axis.

Parameters

- **x** (*tuple of Variables*) – Variables to be split.
- **indices_or_sections** (*int or 1-D array*) – If this argument is an integer, N, the array will be divided into N equal arrays along axis. If it is a 1-D array of sorted integers, it indicates the positions where the array is split.
- **axis** (*int*) – Axis that the input array is split along.
- **force_tuple** (*bool*) – If `True`, this method returns a tuple even when the number of outputs is one.

Returns

Tuple of *Variable* objects if the number of outputs is more than 1 or *Variable* otherwise. When `force_tuple` is `True`, returned value is always a tuple regardless of the number of outputs.

Return type tuple or *Variable*

Note: This function raises `ValueError` if at least one of the outputs is split to zero-size (i.e. `axis`-th value of its shape is zero).

---

squeeze

`chainer.functions.`**`squeeze`**(*x*, *axis=None*)

Remove demensions of size one from the shape of a ndarray.

Parameters

- **x** (chainer.Variable or :class:`numpy.ndarray` or cupy.ndarray) – Input data.
- **axis** (*None or int or tuple of ints*) – A subset of the single-dimensional entries in the shape to remove. If `None` is supplied, all of them are removed. The dimension index starts at zero. If an axis with dimension greater than one is selected, an error is raised.

Returns Variable whose dimensions of size 1 are removed.

Return type *Variable*

stack

`chainer.functions.`**`stack`**(*xs*, *axis=0*)

Concatenate variables along a new axis.

Parameters

- **xs** (*list of chainer.Variable*) – Variables to be concatenated.
- **axis** (*int*) – Axis of result along which variables are stacked.

Returns Output variable.

Return type *Variable*

swapaxes

`chainer.functions.`**`swapaxes`**(*x*, *axis1*, *axis2*)

Swap two axes of a variable.

Parameters

- **x** (`Variable`) – Input variable.
- **axis1** (*int*) – The first axis to swap.
- **axis2** (*int*) – The second axis to swap.

Returns Variable whose axes are swapped.

Return type *Variable*

tile

chainer.functions.**tile**(*x*, *reps*)
Construct an array by tiling a given array.

Parameters

- **x** (chainer.Variable or numpy.ndarray or cupy.ndarray) – Input data.
- **reps** (*int or tuple of ints*) – The number of times for each axis with which x is replicated.

Returns Variable tiled the given array.

Return type *Variable*

transpose

chainer.functions.**transpose**(*x*, *axes=None*)
Permute the dimensions of an input variable without copy.

Parameters

- **x** (Variable) – Input variable.
- **axes** (*tuple of ints*) – By default, reverse the dimensions, otherwise permute the axes according to the values given.

Returns Variable whose axes are permuted.

Return type *Variable*

transpose_sequence

chainer.functions.**transpose_sequence**(*xs*)
Transpose a list of Variables.

This function transposes a list of *Variable* s and returns a list of Variable s. For example a user gives [(0, 1, 2, 3), (4, 5), (6)], the function returns [(0, 4, 6), (1, 5), (2), (3)]. Note that a given list needs to be sorted by each length of *Variable*.

Parameters **xs** (*list of ~chainer.Variable*) – Variables to transpose.

Returns Transposed list.

Return type tuple or *Variable*

vstack

chainer.functions.**vstack**(*xs*)
Concatenate variables vertically (row wise).

Parameters **xs** (*list of chainer.Variable*) – Variables to be concatenated.

Returns Output variable.

Return type *Variable*

where

chainer.functions.**where**(*condition*, *x*, *y*)
Choose elements depending on condition.

This function choose values depending on a given `condition`. All `condition`, `x`, and `y` must have the same shape.

Parameters

- **condition** (Variable) – Variable containing the condition. Only boolean array is permitted.
- **x** (Variable) – Variable chosen when `condition` is `True`.
- **y** (Variable) – Variable chosen when `condition` is `False`.

Returns Variable containing chosen values.

Return type *Variable*

Neural network connections

bilinear

chainer.functions.**bilinear**(*e1*, *e2*, *W*, *V1=None*, *V2=None*, *b=None*)
Applies a bilinear function based on given parameters.

This is a building block of Neural Tensor Network (see the reference paper below). It takes two input variables and one or four parameters, and outputs one variable.

To be precise, denote six input arrays mathematically by $e^1 \in \mathbb{R}^{I \cdot J}$, $e^2 \in \mathbb{R}^{I \cdot K}$, $W \in \mathbb{R}^{J \cdot K \cdot L}$, $V^1 \in \mathbb{R}^{J \cdot L}$, $V^2 \in \mathbb{R}^{K \cdot L}$, and $b \in \mathbb{R}^L$, where I is mini-batch size. In this document, we call V^1, V^2, and b linear parameters.

The output of forward propagation is calculated as

$$y_{il} = \sum_{jk} e^1_{ij} e^2_{ik} W_{jkl} + \sum_j e^1_{ij} V^1_{jl} + \sum_k e^2_{ik} V^2_{kl} + b_l.$$

Note that V1, V2, b are optional. If these are not given, then this function omits the last three terms in the above equation.

Note: This function accepts an input variable `e1` or `e2` of a non-matrix array. In this case, the leading dimension is treated as the batch dimension, and the other dimensions are reduced to one dimension.

Note: In the original paper, J and K must be equal and the author denotes $[V^1 V^2]$ (concatenation of matrices) by V.

Parameters

- **e1** (Variable) – Left input variable.
- **e2** (Variable) – Right input variable.
- **W** (Variable) – Quadratic weight variable.
- **V1** (Variable) – Left coefficient variable.

- **V2** (Variable) – Right coefficient variable.
- **b** (Variable) – Bias variable.

Returns Output variable.

Return type *Variable*

See: Reasoning With Neural Tensor Networks for Knowledge Base Completion [Socher+, NIPS2013].

convolution_2d

chainer.functions.**convolution_2d**(*x*, *W*, *b=None*, *stride=1*, *pad=0*, *use_cudnn=True*, *cover_all=False*, *deterministic=False*)

Two-dimensional convolution function.

This is an implementation of two-dimensional convolution in ConvNets. It takes three variables: the input image `x`, the filter weight `W`, and the bias vector `b`.

Notation: here is a notation for dimensionalities.

- n is the batch size.
- c_I and c_O are the number of the input and output channels, respectively.
- h and w are the height and width of the input image, respectively.
- k_H and k_W are the height and width of the filters, respectively.

Parameters

- **x** (Variable) – Input variable of shape (n, c_I, h, w).
- **W** (Variable) – Weight variable of shape (c_O, c_I, k_H, k_W).
- **b** (Variable) – Bias variable of length c_O (optional).
- **stride** (*int or pair of ints*) – Stride of filter applications. `stride=s` and `stride=(s, s)` are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p)` are equivalent.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available.
- **cover_all** (*bool*) – If `True`, all spatial locations are convoluted into some output pixels. It may make the output size larger.
- **deterministic** (*bool*) – The output of this function can be non-deterministic when it uses cuDNN. If this option is `True`, then it forces cuDNN to use a deterministic algorithm. This option is only available for cuDNN version >= v3.

Returns Output variable.

Return type *Variable*

The two-dimensional convolution function is defined as follows. Then the `Convolution2D` function computes correlations between filters and patches of size (k_H, k_W) in `x`. Note that correlation here is equivalent to the inner product between expanded vectors. Patches are extracted at positions shifted by multiples of `stride` from the first position `-pad` for each spatial axis. The right-most (or bottom-most) patches do not run over the padded spatial size.

Let (s_Y, s_X) be the stride of filter application, and (p_H, p_W) the spatial padding size. Then, the output size (h_O, w_O) is determined by the following equations:

$$h_O = (h + 2p_H - k_H)/s_Y + 1,$$
$$w_O = (w + 2p_W - k_W)/s_X + 1.$$

If the bias vector is given, then it is added to all spatial locations of the output of convolution.

See also:

Convolution2D

convolution_nd

chainer.functions.**convolution_nd**(*x*, *W*, *b=None*, *stride=1*, *pad=0*, *use_cudnn=True*, *cover_all=False*)

N-dimensional convolution function.

This is an implementation of N-dimensional convolution which is generalized two-dimensional convolution in ConvNets. It takes three variables: the input `x`, the filter weight `W` and the bias vector `b`.

Notation: here is a notation for dimensionalities.

- N is the number of spatial dimensions.
- n is the batch size.
- c_I and c_O are the number of the input and output channels, respectively.
- $d_1, d_2, ..., d_N$ are the size of each axis of the input's spatial dimensions, respectively.
- $k_1, k_2, ..., k_N$ are the size of each axis of the filters, respectively.

Parameters

- **x** (*Variable*) – Input variable of shape $(n, c_I, d_1, d_2, ..., d_N)$.
- **W** (*Variable*) – Weight variable of shape $(c_O, c_I, k_1, k_2, ..., k_N)$.
- **b** (*Variable*) – One-dimensional bias variable with length c_O (optional).
- **stride** (*int or tuple of ints*) – Stride of filter applications $(s_1, s_2, ..., s_N)$. `stride=s` is equivalent to `(s, s, ..., s)`.
- **pad** (*int or tuple of ints*) – Spatial padding width for input arrays $(p_1, p_2, ..., p_N)$. `pad=p` is equivalent to `(p, p, ..., p)`.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available. See below for the excact conditions.
- **cover_all** (*bool*) – If `True`, all spatial locations are convoluted into some output pixels. It may make the output size larger. *cover_all* needs to be `False` if you want to use cuDNN.

Returns Output variable.

Return type *Variable*

This function uses cuDNN implementation for its forward and backward computation if ALL of the following conditions are satisfied:

- `cuda.cudnn_enabled` is `True`
- `use_cudnn` is `True`

- The number of spatial dimensions is more than one.
- `cover_all` is `False`
- The input's `dtype` is equal to the filter weight's.
- The `dtype` is FP32, FP64 or FP16(cuDNN version is equal to or greater than v3)

See also:

ConvolutionND, *convolution_2d()*

deconvolution_2d

chainer.functions.**deconvolution_2d**(*x*, *W*, *b=None*, *stride=1*, *pad=0*, *outsize=None*, *use_cudnn=True*, *deterministic=False*)

Two dimensional deconvolution function.

This is an implementation of two-dimensional deconvolution. It takes three variables: input image `x`, the filter weight `W`, and the bias vector `b`.

Parameters

- **x** (*Variable*) – Input variable of shape (n, c_I, h, w).
- **W** (*Variable*) – Weight variable of shape (c_I, c_O, k_H, k_W).
- **b** (*Variable*) – Bias variable of length c_O (optional).
- **stride** (*int or pair of ints*) – Stride of filter applications. `stride=s` and `stride=(s, s)` are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p)` are equivalent.
- **outsize** (*tuple*) – Expected output size of deconvolutional operation. It should be pair of height and width (out_H, out_W). Default value is `None` and the outsize is estimated by input size, stride and pad.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available.
- **deterministic** (*bool*) – The output of this function can be non-deterministic when it uses cuDNN. If this option is `True`, then it forces cuDNN to use a deterministic algorithm. This option is only available for cuDNN version >= v3.

The filter weight has four dimensions (c_I, c_O, k_H, k_W) which indicate the number of input channels, output channels, height and width of the kernels, respectively.

The bias vector is of size c_O.

Let X be the input tensor of dimensions (n, c_I, h, w), (s_Y, s_X) the stride of filter application, and (p_H, p_W) the spatial padding size. Then, the output size (h_O, w_O) is determined by the following equations:

$$h_O = s_Y(h - 1) + k_H - 2p_H,$$
$$w_O = s_X(w - 1) + k_W - 2p_W.$$

deconvolution_nd

chainer.functions.**deconvolution_nd**(*x*, *W*, *b=None*, *stride=1*, *pad=0*, *outsize=None*, *use_cudnn=True*)

N-dimensional deconvolution function.

This is an implementation of N-dimensional deconvolution which generalizes two-dimensional one. It takes three variables: input `x`, the filter weight `W`, and the bias vector `b`.

Parameters

- **x** (chainer.Variable or `numpy.ndarray` or cupy.ndarray) – Input data of shape $(n, c_I, d_1, d_2, ..., d_N)$.
- **W** (chainer.Variable or `numpy.ndarray` or cupy.ndarray) – Weight data of shape $(c_I, c_O, k_1, k_2, ..., k_N)$.
- **b** (chainer.Variable or `numpy.ndarray` or cupy.ndarray) – Bias vector of length c_O (optional).
- **stride** (*int or tuple of ints*) – Stride of filter applications $(s_1, s_2, ..., s_N)$. `stride=s` is equivalent to `(s, s, ..., s)`.
- **pad** (*int or tuple of ints*) – Spatial padding size for input arrays $(p_1, p_2, ..., p_N)$. `pad=p` is equivalent to `(p, p, ..., p)`.
- **outsize** (*tuple of ints*) – Expected output size of deconvolutional operation. It should be a tuple of ints $(out_1, out_2, ..., out_N)$. Default value is `None` and the outsize is estimated by input size, stride and pad.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available. Note that cuDNN supports more than one-dimensional deconvolution operations only.

Returns Output variable.

Return type *Variable*

The filter weight has the following dimensions $(c_I, c_O, k_1, k_2, ..., k_N)$ which indicate the number of input channels, that of output channels and the filter's spatial sizes, respectively.

The one-dimensional bias vector is of size c_O.

Let X be the input tensor of dimensions $(n, c_I, d_1, d_2, ..., d_N)$, $(s_1, s_2, ..., s_N)$ the stride of filter applications, and $(p_1, p_2, ..., p_N)$ the spacial padding size. Then the output size $(out_1, out_2, ..., out_N)$ is determined by the following equations:

$$
\begin{aligned}
out_1 &= s_1(d_1 - 1) + k_1 - 2p_1, \\
out_2 &= s_2(d_2 - 1) + k_2 - 2p_2, \\
&..., \\
out_N &= s_N(d_N - 1) + k_N - 2p_N.
\end{aligned}
$$

See also:

`links.DeconvolutionND`, *deconvolution_2d()*

dilated_convolution_2d

`chainer.functions.`**dilated_convolution_2d**(*x*, *W*, *b=None*, *stride=1*, *pad=0*, *dilate=1*, *use_cudnn=True*, *cover_all=False*)

Two-dimensional dilated convolution function.

This is an implementation of two-dimensional dilated convolution in ConvNets. It takes three variables: the input image `x`, the filter weight `W`, and the bias vector `b`.

Notation: here is a notation for dimensionalities.

- n is the batch size.

- c_I and c_O are the number of the input and output, respectively.
- h and w are the height and width of the input image, respectively.
- k_H and k_W are the height and width of the filters, respectively.

Parameters

- **x** (Variable) – Input variable of shape (n, c_I, h, w).
- **W** (Variable) – Weight variable of shape (c_O, c_I, k_H, k_W).
- **b** (Variable) – Bias variable of length c_O (optional).
- **stride** (*int or pair of ints*) – Stride of filter applications. `stride=s` and `stride=(s, s)` are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p)` are equivalent.
- **dilate** (*int or pair of ints*) – Dilation factor of filter applications. `dilate=d` and `dilate=(d, d)` are equivalent.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available.
- **cover_all** (*bool*) – If `True`, all spatial locations are convoluted into some output pixels. It may make the output size larger.

Returns Output variable.

Return type *Variable*

The two-dimensional dilated convolution function is defined as follows. Then the `DilatedConvolution2D` function computes correlations between filters and patches of size (k_H, k_W) in `x`. Patches here are extracted at intervals of the dilation factor. Note that correlation here is equivalent to the inner product between expanded vectors. Patches are extracted at intervals of the dilation factor and at positions shifted by multiples of `stride` from the first position `-pad` for each spatial axis. The right-most (or bottom-most) patches do not run over the padded spatial size.

Let (s_Y, s_X) be the stride of filter application, (p_H, p_W) the spatial padding size, and (d_Y, d_X) the dilation factor of filter application. Then, the output size (h_O, w_O) is determined by the following equations:

$$h_O = (h + 2p_H - k_H - (k_H - 1) * (d_Y - 1))/s_Y + 1,$$
$$w_O = (w + 2p_W - k_W - (k_W - 1) * (d_X - 1))/s_X + 1.$$

If the bias vector is given, then it is added to all spatial locations of the output of convolution.

See also:

`DilatedConvolution2D`

embed_id

`chainer.functions.`**`embed_id`**(*x*, *W*, *ignore_label=None*)

Efficient linear function for one-hot input.

This function implements so called *word embedding*. It takes two arguments: a set of IDs (words) `x` in B dimensional integer vector, and a set of all ID (word) embeddings `W` in $V \times d$ float32 matrix. It outputs $B \times d$ matrix whose `i`-th column is the `x[i]`-th column of `W`.

This function is only differentiable on the input `W`.

Parameters

- **x** (Variable) – Batch vectors of IDs.
- **W** (Variable) – Representation of each ID (a.k.a. word embeddings).
- **ignore_label** (*int or None*) – If ignore_label is an int value, i-th column of return value is filled with 0.

Returns Output variable.

Return type *Variable*

See also:

EmbedID

linear

chainer.functions.**linear**(*x*, *W*, *b=None*)

Linear function, or affine transformation.

It accepts two or three arguments: an input minibatch x, a weight matrix W, and optionally a bias vector b. It computes

$$Y = xW^\top + b.$$

Parameters

- **x** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Input variable, which is a $(s_B, s_1, s_2, ..., s_n)$-shaped float array. Its first dimension (s_B) is assumed to be the *minibatch dimension*. The other dimensions are treated as concatenated one dimension whose size must be $(s_1 * ... * s_n = N)$.
- **W** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Weight variable of shape (M, N), where $(N = s_1 * ... * s_n)$.
- **b** (*Variable* or numpy.ndarray or *cupy.ndarray*) – Bias variable (optional) of shape $(M,)$.

Returns Output variable. A float array with shape of (s_B, M).

Return type *Variable*

See also:

Linear

Example

```
>>> x = np.random.uniform(0, 1, (3, 4)).astype('f')
>>> W = np.random.uniform(0, 1, (5, 4)).astype('f')
>>> b = np.random.uniform(0, 1, (5,)).astype('f')
>>> y = F.linear(x, W, b)
>>> y.shape
(3, 5)
```

n_step_lstm

chainer.functions.**n_step_lstm**(*n_layers*, *dropout_ratio*, *hx*, *cx*, *ws*, *bs*, *xs*, *train=True*, *use_cudnn=True*)

Stacked Long Short-Term Memory function for sequence inputs.

This function calculates stacked LSTM with sequences. This function gets an initial hidden state h_0, an initial cell state c_0, an input sequence x, weight matrices W, and bias vectors b. This function calculates hidden states h_t and c_t for each time t from input x_t.

$$
\begin{aligned}
i_t &= \sigma(W_0 x_t + W_4 h_{t-1} + b_0 + b_4) \\
f_t &= \sigma(W_1 x_t + W_5 h_{t-1} + b_1 + b_5) \\
o_t &= \sigma(W_2 x_t + W_6 h_{t-1} + b_2 + b_6) \\
a_t &= \tanh(W_3 x_t + W_7 h_{t-1} + b_3 + b_7) \\
c_t &= f_t \dot{c}_{t-1} + i_t \dot{a}_t \\
h_t &= o_t \dot{\tanh}(c_t)
\end{aligned}
$$

As the function accepts a sequence, it calculates h_t for all t with one call. Eight weight matrices and eight bias vectors are required for each layers. So, when S layers exists, you need to prepare $8S$ weigth matrices and $8S$ bias vectors.

If the number of layers `n_layers` is greather than 1, input of `k`-th layer is hidden state `h_t` of `k-1`-th layer. Note that all input variables except first layer may have different shape from the first layer.

Parameters

- **n_layers** (*int*) – Number of layers.
- **dropout_ratio** (*float*) – Dropout ratio.
- **hx** (`chainer.Variable`) – Variable holding stacked hidden states. Its shape is `(S, B, N)` where `S` is number of layers and is equal to `n_layers`, `B` is mini-batch size, and `N` is dimention of hidden units.
- **cx** (`chainer.Variable`) – Variable holding stacked cell states. It has the same shape as `hx`.
- **ws** (*list of list of chainer.Variable*) – Weight matrices. `ws[i]` represents weights for i-th layer. Each `ws[i]` is a list containing eight matrices. `ws[i][j]` is corresponding with `W_j` in the equation. Only `ws[0][j]` where `0 <= j < 4` is `(I, N)` shape as they are multiplied with input variables. All other matrices has `(N, N)` shape.
- **bs** (*list of list of chainer.Variable*) – Bias vectors. `bs[i]` represnents biases for i-th layer. Each `bs[i]` is a list containing eight vectors. `bs[i][j]` is corresponding with `b_j` in the equation. Shape of each matrix is `(N,)` where `N` is dimention of hidden units.
- **xs** (*list of chainer.Variable*) – A list of *Variable* holding input values. Each element `xs[t]` holds input value for time `t`. Its shape is `(B_t, I)`, where `B_t` is mini-batch size for time `t`, and `I` is size of input units. Note that this functions supports variable length sequences. When sequneces has different lengths, sort sequences in descending order by length, and transpose the sorted sequence. *transpose_sequence()* transpose a list of *Variable()* holding sequence. So `xs` needs to satisfy `xs[t].shape[0] >= xs[t + 1].shape[0]`.
- **train** (*bool*) – If `True`, this function executes dropout.
- **use_cudnn** (*bool*) – If `True`, this function uses cuDNN if available.

Returns

This functions returns a tuple concaining three elements, `hy`, `cy` and `ys`.

- `hy` is an updated hidden states whose shape is same as `hx`.
- `cy` is an updated cell states whose shape is same as `cx`.
- `ys` is a list of *Variable* . Each element `ys[t]` holds hidden states of the last layer corresponding to an input `xs[t]`. Its shape is `(B_t, N)` where `B_t` is mini-batch size for time `t`, and `N` is size of hidden units. Note that `B_t` is the same value as `xs[t]`.

Return type tuple

See also:

chainer.functions.lstm()

Evaluation functions

accuracy

chainer.functions.**accuracy**(*y*, *t*, *ignore_label=None*)

Computes multiclass classification accuracy of the minibatch.

Parameters

- **y** (Variable) – Variable holding a matrix whose (i, j)-th element indicates the score of the class j at the i-th example.
- **t** (Variable) – Variable holding an int32 vector of ground truth labels.
- **ignore_label** (*int or None*) – Skip calculating accuracy if the true label is `ignore_label`.

Returns A variable holding a scalar array of the accuracy.

Return type *Variable*

Note: This function is non-differentiable.

binary_accuracy

chainer.functions.**binary_accuracy**(*y*, *t*)

Computes binary classification accuracy of the minibatch.

Parameters

- **y** (Variable) – Variable holding a matrix whose i-th element indicates the score of positive at the i-th example.
- **t** (Variable) – Variable holding an int32 vector of ground truth labels. If `t[i] == -1`, corresponding `x[i]` is ignored. Accuracy is zero if all ground truth labels are `-1`.

Returns A variable holding a scalar array of the accuracy.

Return type *Variable*

Note: This function is non-differentiable.

classification_summary

chainer.functions.**classification_summary**(*y*, *t*, *label_num=None*, *beta=1.0*, *ignore_label=-1*)

Calculates Precision, Recall, F beta Score, and support.

This function calculates the following quantities for each class.

•Precision: $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}$

•Recall: $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{tn}}$

•F beta Score: The weighted harmonic average of Precision and Recall.

•Support: The number of instances of each ground truth label.

Here, `tp`, `fp`, and `tn` stand for the number of true positives, false positives, and true negative, respectively.

`label_num` specifies the number of classes, that is, each value in `t` must be an integer in the range of `[0, label_num)`. If `label_num` is `None`, this function regards `label_num` as a maximum of in `t` plus one.

`ignore_label` determines which instances should be ignored. Specifically, instances with the given label are not taken into account for calculating the above quantities. By default, it is set to -1 so that all instances are taken into consideration, as labels are supposed to be non-negative integers. Setting `ignore_label` to a non-negative integer less than `label_num` is illegal and yields undefined behavior. In the current implementation, it arises `RuntimeWarning` and `ignore_label`-th entries in output arrays do not contain correct quantities.

Parameters

- **y** (Variable) – Variable holding a vector of scores.
- **t** (Variable) – Variable holding a vector of ground truth labels.
- **label_num** (*int*) – The number of classes.
- **beta** (*float*) – The parameter which determines the weight of precision in the F-beta score.
- **ignore_label** (*int*) – Instances with this label are ignored.

Returns 4-tuple of ~chainer.Variable of size `(label_num,)`. Each element represents precision, recall, F beta score, and support of this minibatch.

r2_score

chainer.functions.**r2_score**(*pred*, *true*, *sample_weight=None*, *multioutput='uniform_average'*)

Computes R^2(coefficient of determination) regression score function.

Parameters

- **pred** (Variable) – Variable holding a vector, matrix or tensor of estimated target values.
- **true** (Variable) – Variable holding a vector, matrix or tensor of correct target values.
- **sample_weight** – This argument is for compatibility with scikit-learn's implementation of r2_score. Current implementation admits None only.
- **multioutput** (*string*) – ['uniform_average', 'raw_values']. if 'uniform_average', this function returns an average of R^2 score of multiple output. If 'raw_average', this function return a set of R^2 score of multiple output.

Returns A Variable holding a scalar array of the R^2 score if 'multioutput' is 'uniform_average' or a vector of R^2 scores if 'multioutput' is 'raw_values'.

Return type *Variable*

Note: This function is non-differentiable.

Loss functions

bernoulli_nll

chainer.functions.**bernoulli_nll**(*x*, *y*)

Computes the negative log-likelihood of a Bernoulli distribution.

This function calculates the negative log-likelihood of a Bernoulli distribution.

$$-\log B(x;p) = -\sum_i \{x_i \log(p_i) + (1-x_i)\log(1-p_i)\},$$

where $p = \sigma(y)$, $\sigma(\cdot)$ is a sigmoid function, and $B(x;p)$ is a Bernoulli distribution.

Note: As this function uses a sigmoid function, you can pass a result of fully-connected layer (that means `Linear`) to this function directly.

Parameters

- **x** (`Variable`) – Input variable.
- **y** (`Variable`) – A variable representing the parameter of Bernoulli distribution.

Returns A variable representing negative log-likelihood.

Return type *Variable*

black_out

chainer.functions.**black_out**(*x*, *t*, *W*, *samples*)

BlackOut loss function.

BlackOut loss function is defined as

$$-\log(p(t)) - \sum_{s \in S} \log(1-p(s)),$$

where t is the correct label, S is a set of negative examples and $p(\cdot)$ is likelihood of a given label. And, p is defined as

$$p(y) = \frac{\exp(W_y^\top x)}{\sum_{s \in samples} \exp(W_s^\top x)}.$$

Parameters

- **x** (`Variable`) – Batch of input vectors.
- **t** (`Variable`) – Vector of ground truth labels.
- **W** (`Variable`) – Weight matrix.
- **samples** (`Variable`) – Negative samples.

Returns Loss value.

Return type *Variable*

See: BlackOut: Speeding up Recurrent Neural Network Language Models With Very Large Vocabularies

See also:

BlackOut.

connectionist_temporal_classification

chainer.functions.**connectionist_temporal_classification**(*x*, *t*, *blank_symbol*, *input_length=None*, *label_length=None*)

Connectionist Temporal Classification loss function.

Connectionist Temporal Classification(CTC) *[Graves2006]* is a loss function of sequence labeling where the alignment between the inputs and target is unknown. See also *[Graves2012]*

Parameters

- **x** (*sequence of Variable*) – RNN output at each time. `x` must be a list of *Variable* s. Each element of `x`, `x[i]` is a *Variable* representing output of RNN at time `i`.
- **t** (*Variable*) – Expected label sequence.
- **blank_symbol** (*int*) – Index of blank_symbol. This value must be non-negative.
- **input_length** (*Variable*) – Length of valid sequence for each of mini batch `x` (optional). If input_length is skipped, It regards that all of `x` is valid input.
- **label_length** (*Variable*) – Length of valid sequence for each of mini batch `t` (optional). If label_length is skipped, It regards that all of `t` is valid input.

Returns A variable holding a scalar value of the CTC loss.

Return type *Variable*

Note: You need to input `x` without applying to activation functions(e.g. softmax function), because this function applies softmax functions to `x` before calculating CTC loss to avoid numerical limitations. You also need to apply softmax function to forwarded values before you decode it.

Note: This function is differentiable only by `x`.

Note: This function supports (batch, sequence, 1-dimensional input)-data.

---

contrastive

chainer.functions.**contrastive**(*x0*, *x1*, *y*, *margin=1*)

Computes contrastive loss.

It takes a pair of variables and a label as inputs. The label is 1 when those two input variables are similar, or 0 when they are dissimilar. Let N and K denote mini-batch size and the dimension of input variables, respectively. The shape of both input variables should be `(N, K)`.

$$L = \frac{1}{2N}\left(\sum_{n=1}^{N} y_n d_n^2 + (1-y_n)\max(\text{margin} - d_n, 0)^2\right)$$

where $d_n = \|\mathbf{x_0}_n - \mathbf{x_1}_n\|_2$. N denotes the mini-batch size. Input variables, x0 and x1, have N vectors, and each vector is K-dimensional. Therefore, $\mathbf{x_0}_n$ and $\mathbf{x_1}_n$ are n-th K-dimensional vectors of x0 and x1.

Parameters

- **x0** (*Variable*) – The first input variable. The shape should be (N, K), where N denotes the mini-batch size, and K denotes the dimension of `x0`.
- **x1** (*Variable*) – The second input variable. The shape should be the same as `x0`.
- **y** (*Variable*) – Labels. All values should be 0 or 1. The shape should be `(N,)`, where N denotes the mini-batch size.
- **margin** (*float*) – A parameter for contrastive loss. It should be positive value.

Returns

A variable holding a scalar that is the loss value calculated by the above equation.

Return type *Variable*

Note: This cost can be used to train siamese networks. See Learning a Similarity Metric Discriminatively, with Application to Face Verification for details.

---

crf1d

chainer.functions.**crf1d**(*cost*, *xs*, *ys*)

Calculates negative log-likelihood of linear-chain CRF.

It takes a transition cost matrix, a sequence of costs, and a sequence of labels. Let c_{st} be a transition cost from a label s to a label t, x_{it} be a cost of a label t at position i, and y_i be an expected label at position i. The negative log-likelihood of linear-chain CRF is defined as

$$L = -\left(\sum_{i=1}^{l} x_{iy_i} + \sum_{i=1}^{l-1} c_{y_i y_{i+1}} - \log(Z)\right),$$

where l is the length of the input sequence and Z is the normalizing constant called partition function.

Note: When you want to calculate the negative log-likelihood of sequences which have different lengths, sort the sequences in descending order of lengths and transpose the sequences. For example, you have three input seuqnces:

```
>>> a1 = a2 = a3 = a4 = np.random.uniform(-1, 1, 3).astype('f')
>>> b1 = b2 = b3 = np.random.uniform(-1, 1, 3).astype('f')
>>> c1 = c2 = np.random.uniform(-1, 1, 3).astype('f')
```

```
>>> a = [a1, a2, a3, a4]
>>> b = [b1, b2, b3]
>>> c = [c1, c2]
```

where `a1` and all other variables are arrays with `(K,)` shape. Make a transpose of the sequences:

```
>>> x1 = np.stack([a1, b1, c1])
>>> x2 = np.stack([a2, b2, c2])
>>> x3 = np.stack([a3, b3])
>>> x4 = np.stack([a4])
```

and make a list of the arrays:

```
>>> xs = [x1, x2, x3, x4]
```

You need to make label sequences in the same fashion. And then, call the function:

```
>>> cost = chainer.Variable(
...     np.random.uniform(-1, 1, (3, 3)).astype('f'))
>>> ys = [np.zeros(x.shape[0:1], dtype='i') for x in xs]
>>> loss = F.crf1d(cost, xs, ys)
```

It calculates sum of the negative log-likelihood of the three sequences.

Parameters

- **cost** (Variable) – A $K \times K$ matrix which holds transition cost between two labels, where K is the number of labels.
- **xs** (*list of Variable*) – Input vector for each label. `len(xs)` denotes the length of the sequence, and each *Variable* holds a $B \times K$ matrix, where B is mini-batch size, K is the number of labels. Note that B s in all the variables are not necessary the same, i.e., it accepts the input sequences with different lengths.
- **ys** (*list of Variable*) – Expected output labels. It needs to have the same length as `xs`. Each *Variable* holds a B integer vector. When `x` in `xs` has the different B, corresponding `y` has the same B. In other words, `ys` must satisfy `ys[i].shape == xs[i].shape[0:1]` for all `i`.

Returns

A variable holding the average negative log-likelihood of the input sequences.

Return type *Variable*

Note: See detail in the original paper: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

`chainer.functions.`**`argmax_crf1d`**(*cost*, *xs*)

Computes a state that maximizes a joint probability of the given CRF.

Parameters

- **cost** (Variable) – A $K \times K$ matrix which holds transition cost between two labels, where K is the number of labels.
- **xs** (*list of Variable*) – Input vector for each label. `len(xs)` denotes the length of the sequence, and each *Variable* holds a $B \times K$ matrix, where B is mini-batch size, K is the number of labels. Note that B s in all the variables are not necessary the same, i.e., it accepts the input sequences with different lengths.

Returns

A tuple of *Variable* object `s` and a `list` `ps`. The shape of `s` is `(B,)`, where `B` is the mini-batch size. i-th element of `s`, `s[i]`, represents log-likelihood of i-th data. `ps` is a list of `numpy.ndarray` or *`cupy.ndarray`*, and denotes the state that maximizes the point probability. `len(ps)` is equal to `len(xs)`, and shape of each `ps[i]` is the mini-batch size of the corresponding `xs[i]`. That means, `ps[i].shape == xs[i].shape[0:1]`.

Return type tuple

cross_covariance

`chainer.functions.`**`cross_covariance`**(*y*, *z*)

Computes the sum-squared cross-covariance penalty between `y` and `z`

Parameters

- **`y`** (`Variable`) – Variable holding a matrix where the first dimension corresponds to the batches.
- **`z`** (`Variable`) – Variable holding a matrix where the first dimension corresponds to the batches.

Returns A variable holding a scalar of the cross covariance loss.

Return type *Variable*

Note: This cost can be used to disentangle variables. See https://arxiv.org/abs/1412.6583v3 for details.

---

gaussian_kl_divergence

`chainer.functions.`**`gaussian_kl_divergence`**(*mean*, *ln_var*)

Computes the KL-divergence of Gaussian variables from the standard one.

Given two variable `mean` representing μ and `ln_var` representing $\log(\sigma^2)$, this function returns a variable representing the KL-divergence between the given multi-dimensional Gaussian $N(\mu, S)$ and the standard Gaussian $N(0, I)$

$$D_{\mathbf{KL}}(N(\mu, S) \| N(0, I)),$$

where S is a diagonal matrix such that $S_{ii} = \sigma_i^2$ and I is an identity matrix.

Parameters

- **`mean`** (`Variable`) – A variable representing mean of given gaussian distribution, μ.
- **`ln_var`** (`Variable`) – A variable representing logarithm of variance of given gaussian distribution, $\log(\sigma^2)$.

Returns

A variable representing KL-divergence between given gaussian distribution and the standard gaussian.

Return type *Variable*

gaussian_nll

chainer.functions.**gaussian_nll**(*x*, *mean*, *ln_var*)

Computes the negative log-likelihood of a Gaussian distribution.

Given two variable `mean` representing μ and `ln_var` representing $\log(\sigma^2)$, this function returns the negative log-likelihood of x on a Gaussian distribution $N(\mu, S)$,

$$-\log N(x; \mu, \sigma^2) = \log\left(\sqrt{(2\pi)^D |S|}\right) + \frac{1}{2}(x-\mu)^\top S^{-1}(x-\mu),$$

where D is a dimension of x and S is a diagonal matrix where $S_{ii} = \sigma_i^2$.

Parameters

- **x** (Variable) – Input variable.
- **mean** (Variable) – A variable representing mean of a Gaussian distribution, μ.
- **ln_var** (Variable) – A variable representing logarithm of variance of a Gaussian distribution, $\log(\sigma^2)$.

Returns A variable representing the negative log-likelihood.

Return type *Variable*

hinge

chainer.functions.**hinge**(*x*, *t*, *norm='L1'*)

Computes the hinge loss for a one-of-many classification task.

$$L = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}[\max(0, 1 - \delta\{l_n = k\}t_{nk})]^p$$

where N denotes the batch size, K is the number of classes of interest,

$$\delta\{\text{condition}\} = \begin{cases} 1 & \text{if condition} \\ -1 & \text{otherwise,} \end{cases}$$

and

$$p = \begin{cases} 1 & \text{if norm} = \text{'L1'} \\ 2 & \text{if norm} = \text{'L2'.} \end{cases}$$

Parameters

- **x** (Variable) – Input variable. The shape of `x` should be (N, K).
- **t** (Variable) – The N-dimensional label vector $\mathbf{l}$ with values $l_n \in \{0, 1, 2, \ldots, K-1\}$. The shape of `t` should be (N,).
- **norm** (*string*) – Specifies norm type. Only either 'L1' or 'L2' is acceptable.

Returns

A variable object holding a scalar array of the hinge loss L.

Return type *Variable*

huber_loss

chainer.functions.**huber_loss**(*x*, *t*, *delta*)

Loss function which is less sensitive to outliers in data than MSE.

$$a = x - t$$

and

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{if } |a| \leq \delta \\ \delta(|a| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases}$$

Parameters

- **x** (Variable) – Input variable. The shape of x should be (N, K).
- **t** (Variable) – Target variable for regression. The shape of t should be (N, K).
- **delta** (*float*) – Constant variable for huber loss function as used in definition.

Returns

A variable object holding a scalar array of the huber loss L_δ.

Return type *Variable*

See: Huber loss - Wikipedia.

mean_absolute_error

chainer.functions.**mean_absolute_error**(*x0*, *x1*)

Mean absolute error function.

This function computes mean absolute error between two variables. The mean is taken over the minibatch.

mean_squared_error

chainer.functions.**mean_squared_error**(*x0*, *x1*)

Mean squared error function.

This function computes mean squared error between two variables. The mean is taken over the minibatch. Note that the error is not scaled by 1/2.

negative_sampling

chainer.functions.**negative_sampling**(*x*, *t*, *W*, *sampler*, *sample_size*)

Negative sampling loss function.

In natural language processing, especially language modeling, the number of words in a vocabulary can be very large. Therefore, you need to spend a lot of time calculating the gradient of the embedding matrix.

By using the negative sampling trick you only need to calculate the gradient for a few sampled negative examples.

The objective function is below:

$$f(x, p) = \log \sigma(x^\top w_p) + \ kE_{i \sim P(i)}[\log \sigma(-x^\top w_i)],$$

where $\sigma(\cdot)$ is a sigmoid function, w_i is the weight vector for the word i, and p is a positive example. It is approximated with k examples N sampled from probability $P(i)$, like this:

$$f(x, p) \approx \log \sigma(x^\top w_p) + \sum_{n \in N} \log \sigma(-x^\top w_n).$$

Each sample of N is drawn from the word distribution $P(w)$. This is calculated as $P(w) = \frac{1}{Z} c(w)^\alpha$, where $c(w)$ is the unigram count of the word w, α is a hyper-parameter, and Z is the normalization constant.

Parameters

- **x** (Variable) – Batch of input vectors.
- **t** (Variable) – Vector of ground truth labels.
- **W** (Variable) – Weight matrix.
- **sampler** (*FunctionType*) – Sampling function. It takes a shape and returns an integer array of the shape. Each element of this array is a sample from the word distribution. A *WalkerAlias* object built with the power distribution of word frequency is recommended.
- **sample_size** (*int*) – Number of samples.

See: Distributed Representations of Words and Phrases and their Compositionality

See also:

NegativeSampling.

sigmoid_cross_entropy

chainer.functions.**sigmoid_cross_entropy**(*x*, *t*, *use_cudnn=True*, *normalize=True*)

Computes cross entropy loss for pre-sigmoid activations.

Parameters

- **x** (Variable) – A variable object holding a matrix whose (i, j)-th element indicates the unnormalized log probability of the j-th unit at the i-th example.
- **t** (Variable) – Variable holding an int32 vector of ground truth labels. If `t[i] == -1`, corresponding `x[i]` is ignored. Loss is zero if all ground truth labels are `-1`.
- **normalize** (*bool*) – Variable holding a boolean value which determines the normalization constant. If true, this function normalizes the cross entropy loss across all instances. If else, it only normalizes along a batch size.

Returns

A variable object holding a scalar array of the cross entropy loss.

Return type *Variable*

Note: This function is differentiable only by `x`.

---

softmax_cross_entropy

chainer.functions.**softmax_cross_entropy**(*x*, *t*, *use_cudnn=True*, *normalize=True*, *cache_score=True*, *class_weight=None*, *ignore_label=-1*)

Computes cross entropy loss for pre-softmax activations.

Parameters

- **x** (Variable) – Variable holding a multidimensional array whose element indicates unnormalized log probability: the first axis of the variable represents the number of samples, and the second axis represents the number of classes. While this function computes a usual softmax cross entropy if the number of dimensions is equal to 2, it computes a cross entropy of the replicated softmax if the number of dimensions is greater than 2.
- **t** (Variable) – Variable holding an int32 vector of ground truth labels. If `t[i] == ignore_label`, corresponding `x[i]` is ignored.
- **normalize** (*bool*) – If `True`, this function normalizes the cross entropy loss across all instances. If `False`, it only normalizes along a batch size.
- **cache_score** (*bool*) – When it is `True`, the function stores result of forward computation to use it on backward computation. It reduces computational cost though consumes more memory.
- **class_weight** (*ndarray or ndarray*) – An array that contains constant weights that will be multiplied with the loss values along with the second dimension. The shape of this array should be `(x.shape[1],)`. If this is not `None`, each class weight `class_weight[i]` is actually multiplied to `y[:, i]` that is the corresponding log-softmax output of `x` and has the same shape as `x` before calculating the actual loss value.
- **ignore_label** (*int*) – Label value you want to ignore. Its default value is `-1`. See description of the argument *t*.

Returns A variable holding a scalar array of the cross entropy loss.

Return type *Variable*

Note: This function is differentiable only by `x`.

---

triplet

`chainer.functions.`**triplet**(*anchor*, *positive*, *negative*, *margin=0.2*)

Computes triplet loss.

It takes a triplet of variables as inputs, a, p and n: anchor, positive example and negative example respectively. The triplet defines a relative similarity between samples. Let N and K denote mini-batch size and the dimension of input variables, respectively. The shape of all input variables should be (N, K).

$$L(a, p, n) = \frac{1}{N}\left(\sum_{i=1}^{N} \max\{d(a_i, p_i) - d(a_i, n_i) + \mathrm{margin}, 0\}\right)$$

where $d(x_i, y_i) = \|\mathbf{x}_i - \mathbf{y}_i\|_2^2$.

Parameters

- **anchor** (Variable) – The anchor example variable. The shape should be (N, K), where N denotes the minibatch size, and K denotes the dimension of the anchor.
- **positive** (Variable) – The positive example variable. The shape should be the same as anchor.
- **negative** (Variable) – The negative example variable. The shape should be the same as anchor.
- **margin** (*float*) – A parameter for triplet loss. It should be a positive value.

Returns

A variable holding a scalar that is the loss value calculated by the above equation.

Return type *Variable*

Note: This cost can be used to train triplet networks. See Learning Fine-grained Image Similarity with Deep Ranking for details.

Mathematical functions

arccos

chainer.functions.**arccos**(*x*)

Elementwise arccosine function.

$$y_i = \arccos x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

arcsin

chainer.functions.**arcsin**(*x*)

Elementwise arcsine function.

$$y_i = \arcsin x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

arctan

chainer.functions.**arctan**(*x*)

Elementwise arctangent function.

$$y_i = \arctan x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

argmax

chainer.functions.**argmax**(*x*, *axis=None*)

Returns index which holds maximum of array elements over a given axis.

Parameters

- **x** (Variable) – Array to find maximum elements.
- **axis** (*None or int*) – Axis over which a max is performed. The default (axis = None) is perform a max over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

argmin

chainer.functions.**argmin**(*x*, *axis=None*)

Returns index which holds minimum of array elements over a given axis.

Parameters

- **x** (Variable) – Array to find minimum elements.
- **axis** (*None or int*) – Axis over which a min is performed. The default (axis = None) is perform a min over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

average

chainer.functions.**average**(*x*, *axis=None*, *weights=None*)

Calculate weighted average of array elements over a given axis.

Parameters

- **x** (Variable) – Elements to sum.
- **axis** (*None or int*) – Axis which the method is performed. With the default (axis = None) it performs a mean over all the dimensions of the input array.
- **weights** (*None or chainer.Variable*) – An array holding weights to calculate weighted average. If it is None, all weights are assumed to be one. When axis is None, weights must have the same shape of x. And when axis is int, it must be 1-D array satisfing weights.shape == (x.shape[axis],).

Returns Output variable.

Return type *Variable*

batch_inv

chainer.functions.**batch_inv**(*a*)

Computes the inverse of a batch of square matrices.

Parameters **a** (Variable) – Input array to compute the inverse for. Shape of the array should be `(m, n, n)` where `m` is the number of matrices in the batch, and `n` is the dimensionality of a square matrix.

Returns Inverse of every matrix in the batch of matrices.

Return type *Variable*

batch_l2_norm_squared

chainer.functions.**batch_l2_norm_squared**(*x*)

L2 norm (a.k.a. Euclidean norm) squared.

This function implements the square of L2 norm on a vector. No reduction along batch axis is done.

Parameters **x** (Variable) – Input variable. The first dimension is assumed to be the *minibatch dimension*. If `x` has more than two dimensions all but the first dimension are flattened to one dimension.

Returns Two dimensional output variable.

Return type *Variable*

batch_matmul

chainer.functions.**batch_matmul**(*a*, *b*, *transa=False*, *transb=False*)

Computes the batch matrix multiplications of two sets of arrays.

Parameters

- **a** (Variable) – The left operand of the batch matrix multiplications. A 2-D array of shape `(B, N)` is considered as B $N \times 1$ matrices. A 3-D array of shape `(B, M, N)` is considered as B $M \times N$ matrices.
- **b** (Variable) – The right operand of the batch matrix multiplications. Its array is treated as matrices in the same way as `a`'s array.
- **transa** (*bool*) – If `True`, transpose each matrix in `a`.
- **transb** (*bool*) – If `True`, transpose each matrix in `b`.

Returns

The result of the batch matrix multiplications as a 3-D array.

Return type *Variable*

bias

chainer.functions.**bias**(*x*, *y*, *axis=1*)

Elementwise summation with broadcasting.

Computes a elementwise summation of two input variables, with the shape of the latter variable broadcasted to match the shape of the former. `axis` is the first axis of the first variable along which the second variable is applied.

The term "broadcasting" here comes from Caffe's bias layer so the "broadcasting" with the following arguments:

```
   x : 100 x 3 x 40 x 60
   y : 3 x 40
axis : 1
```

is equivalent to the following numpy broadcasting:

```
x : 100 x 3 x 40 x 60
y :   1 x 3 x 40 x 1
```

Note that how the `axis` indicates to which axis of `x` we apply `y`.

Parameters

- **x** (`Variable`) – Input variable to be summed.
- **y** (`Variable`) – Input variable to sum, broadcasted.
- **axis** (*int*) – The first axis of `x` along which `y` is applied.

Returns Output variable.

Return type *Variable*

ceil

`chainer.functions.`**`ceil`**(*x*)

Elementwise ceil function.

$$y_i = \lceil x_i \rceil$$

Parameters **x** (`Variable`) – Input variable.

Returns Output variable.

Return type *Variable*

clip

`chainer.functions.`**`clip`**(*x*, *x_min*, *x_max*)

Clips (limits) elements of input variable.

Given an interval `[x_min, xmax]`, elements outside the interval are clipped to the interval edges.

Parameters

- **x** (`Variable`) – Input variable to be clipped.
- **x_min** (*float*) – Minimum value.
- **x_max** (*float*) – Maximum value.

Returns Output variable.

Return type *Variable*

cos

`chainer.functions.`**`cos`**(*x*)

Elementwise cos function.

cosh

chainer.functions.**cosh**(*x*)
Elementwise hyperbolic cosine function.

$$y_i = \cosh x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

exp

chainer.functions.**exp**(*x*)
Elementwise exponential function.

fmod

chainer.functions.**fmod**(*x*, *divisor*)
Elementwise mod function.

$$y_i = x_i \bmod \text{divisor}.$$

Parameters

- **x** (Variable) – Input variable.
- **divisor** (Variable) – Input divisor.

Returns Output variable.

Return type *Variable*

floor

chainer.functions.**floor**(*x*)
Elementwise floor function.

$$y_i = \lfloor x_i \rfloor$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

identity

chainer.functions.**identity**(**inputs*)
Just returns input variables.

inv

chainer.functions.**inv**(*a*)
Computes the inverse of square matrix.

Parameters **a** (Variable) – Input array to compute the inverse for. Shape of the array should be `(n, n)` where `n` is the dimensionality of a square matrix.

Returns Matrix inverse of `a`.

Return type *Variable*

linear_interpolate

chainer.functions.**linear_interpolate**(*p*, *x*, *y*)
Elementwise linear-interpolation function.

This function is defined as

$$f(p, x, y) = px + (1 - p)y.$$

Parameters

- **p** (Variable) – Input variable.
- **x** (Variable) – Input variable.
- **y** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

log

chainer.functions.**log**(*x*)
Elementwise natural logarithm function.

log10

chainer.functions.**log10**(*x*)
Elementwise logarithm function to the base 10.

$$y_i = \log_{10} x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

log1p

chainer.functions.**log1p**(*x*)
Elementwise natural logarithm plus one function.

log2

chainer.functions.**log2**(*x*)
Elementwise logarithm function to the base 2.

$$y_i = \log_2 x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

logsumexp

chainer.functions.**logsumexp**(*x*, *axis=None*)
Log-sum-exp of array elements over a given axis.

This function calculates logarithm of sum of exponential of array elements.

$$y_i = \log\left(\sum_j \exp(x_{ij})\right)$$

Parameters

- **x** (Variable) – Elements to log-sum-exp.
- **axis** (*None, int, or tuple of int*) – Axis which a sum is performed. The default (axis = None) is perform a sum over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

matmul

chainer.functions.**matmul**(*a*, *b*, *transa=False*, *transb=False*)
Computes the matrix multiplication of two arrays.

Parameters

- **a** (Variable) – The left operand of the matrix multiplication. A 1-D array of shape (N,) is considered as an $N \times 1$ matrix. A 2-D array of shape (M, N) is considered as an $M \times N$ matrix.
- **b** (Variable) – The right operand of the matrix multiplication. Its array is treated as a matrix in the same way as a's array.
- **transa** (*bool*) – If True, transpose a.
- **transb** (*bool*) – If True, transpose b.

Returns

The result of the matrix multiplication as a 2-D array.

Return type *Variable*

max

chainer.functions.**max**(*x*, *axis=None*, *keepdims=False*)
Maximum of array elements over a given axis.

Parameters

- **x** (Variable) – Array to be maximized.
- **axis** (*None, int, or tuple of int*) – Axis over which a max is performed. The default (axis = None) is perform a max over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

maximum

chainer.functions.**maximum**(*x1*, *x2*)
Element-wise maximum of input variables.

Parameters

- **x1** (Variable) – Input variables to be compared.
- **x2** (Variable) – Input variables to be compared.

Returns Output variable.

Return type *Variable*

min

chainer.functions.**min**(*x*, *axis=None*, *keepdims=False*)
Minimum of array elements over a given axis.

Parameters

- **x** (Variable) – Array to be minimized.
- **axis** (*None, int, or tuple of int*) – Axis over which a min is performed. The default (axis = None) is perform a min over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

minimum

chainer.functions.**minimum**(*x1*, *x2*)
Element-wise minimum of input variables.

Parameters

- **x1** (Variable) – Input variables to be compared.
- **x2** (Variable) – Input variables to be compared.

Returns Output variable.

Return type *Variable*

rsqrt

chainer.functions.**rsqrt**(*x*)

Computes elementwise reciprocal of square root of input x_i.

$$y_i = \frac{1}{\sqrt{x_i}}.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

See also:

sqrt()

scale

chainer.functions.**scale**(*x*, *y*, *axis=1*)

Elementwise product with broadcasting.

Computes a elementwise product of two input variables, with the shape of the latter variable broadcasted to match the shape of the former. `axis` is the first axis of the first variable along which the second variable is applied.

The term "broadcasting" here comes from Caffe's scale layer so the "broadcasting" with the following arguments:

```
   x : 100 x 3 x 40 x 60
   y : 3 x 40
axis : 1
```

is equivalent to the following numpy broadcasting:

```
x : 100 x 3 x 40 x 60
y :   1 x 3 x 40 x 1
```

Note that how the `axis` indicates to which axis of `x` we apply `y`.

Parameters

- **x** (Variable) – Input variable to be scaled.
- **y** (Variable) – Input variable to scale, broadcasted.
- **axis** (*int*) – The first axis of `x` along which `y` is applied.

Returns Output variable.

Return type *Variable*

sin

chainer.functions.**sin**(*x*)

Elementwise sin function.

sinh

chainer.functions.**sinh**(*x*)
Elementwise hyperbolic sine function.

$$y_i = \sinh x_i.$$

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

sqrt

chainer.functions.**sqrt**(*x*)
Elementwise square root function.

$$y_i = \sqrt{x_i}.$$

If the value of x_i is negative, it returns Nan for y_i respect to underlying numpy and cupy specification.

Parameters **x** (Variable) – Input variable.

Returns Output variable.

Return type *Variable*

square

chainer.functions.**square**(*x*)
Elementwise square function.

$$y_i = x_i^2.$$

Parameters **x** (chainer.Variable or numpy.ndarray or cupy.ndarray) – Input variable.

Returns Output variable.

Return type *Variable*

squared_difference

chainer.functions.**squared_difference**(*x1*, *x2*)
Squared difference of input variables.

Parameters

- **x1** (Variable) – Input variables to be compared.
- **x2** (Variable) – Input variables to be compared.

Returns `(x1 - x2) ** 2` element-wise.

Return type *Variable*

sum

chainer.functions.**sum**(*x*, *axis=None*)
Sum of array elements over a given axis.

Parameters

- **x** (Variable) – Elements to sum.
- **axis** (*None, int, or tuple of int*) – Axis which a sum is performed. The default (axis = None) is perform a sum over all the dimensions of the input array.

Returns Output variable.

Return type *Variable*

tanh

Hyperbolic tangent function is described in "Activation functions" section.

See also:

tanh()

tan

chainer.functions.**tan**(*x*)
Elementwise tan function.

Noise injections

dropout

chainer.functions.**dropout**(*x*, *ratio=0.5*, *train=True*)
Drops elements of input variable randomly.

This function drops input elements randomly with probability `ratio` and scales the remaining elements by factor `1 / (1 - ratio)`. In testing mode, it does nothing and just returns `x`.

Parameters

- **x** (Variable) – Input variable.
- **ratio** (*float*) – Dropout ratio.
- **train** (*bool*) – If `True`, executes dropout. Otherwise, does nothing.

Returns Output variable.

Return type *Variable*

See the paper by G. Hinton: Improving neural networks by preventing co-adaptation of feature detectors.

gaussian

chainer.functions.**gaussian**(*mean*, *ln_var*)
Gaussian sampling function.

It takes mean μ and logarithm of variance $\log(\sigma^2)$ as input and output a sample drawn from gaussian $N(\mu, \sigma)$.

Parameters

- **mean** (Variable) – Input variable representing mean μ.
- **ln_var** (Variable) – Input variable representing logarithm of variance $\log(\sigma^2)$.

Returns Output variable.

Return type *Variable*

Normalization functions

batch_normalization

chainer.functions.**batch_normalization**(*x*, *gamma*, *beta*, *eps=2e-05*, *running_mean=None*, *running_var=None*, *decay=0.9*, *use_cudnn=True*)

Batch normalization function.

It takes the input variable `x` and two parameter variables `gamma` and `beta`. The parameter variables must both have the same dimensionality, which is referred to as the channel shape. This channel shape corresponds to the dimensions in the input which are not averaged over. Since the first dimension of the input corresponds to the batch size, the second dimension of *x* will correspond to the first dimension of the channel shape, the third dimension of *x* will correspond to the second channel dimension (if it exists) and so on. Therefore, the dimensionality of the input must be at least one plus the number of channel dimensions. The total effective "batch size" will then be considered to be the product of all dimensions in *x* except for the channel dimensions.

As an example, if the input is four dimensional and the parameter variables are one dimensional, then it is assumed that the first dimension of the input is the batch size, the second dimension is the channel size, and the remaining two dimensions are considered to be spatial dimensions that will be averaged over along with the batch size in the batch normalization computations. That is, the total batch size will be considered to be the product of all input dimensions except the second dimension.

Note: If this function is called, it will not be possible to access the updated running mean and variance statistics, because they are members of the function object, which cannot be accessed by the caller. If it is desired to access the updated running statistics, it is necessary to get a new instance of the function object, call the object, and then access the running_mean and/or running_var attributes. See the corresponding Link class for an example of how to do this.

Parameters

- **x** (Variable) – Input variable.
- **gamma** (Variable) – Scaling parameter of normalized data.
- **beta** (Variable) – Shifting parameter of scaled normalized data.
- **eps** (*float*) – Epsilon value for numerical stability.
- **running_mean** (array) – Running average of the mean. This is a running average of the mean over several mini-batches using the decay parameter. If `None`, the running average is not computed. If this is `None`, then `runnng_var` must also be `None`.
- **running_var** (array) – Running average of the variance. This is a running average of the variance over several mini-batches using the decay parameter. If `None`, the running average is not computed. If this is `None`, then `running_mean` must also be `None`.
- **decay** (*float*) – Decay rate of moving average. It is used during training.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation.

See: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

See also:

`links.BatchNormalization`

fixed_batch_normalization

`chainer.functions.`**`fixed_batch_normalization`**(*x*, *gamma*, *beta*, *mean*, *var*, *eps=2e-05*, *use_cudnn=True*)

Batch normalization function with fixed statistics.

This is a variant of batch normalization, where the mean and variance statistics are given by the caller as fixed variables. This is used on testing mode of the batch normalization layer, where batch statistics cannot be used for prediction consistency.

Parameters

- **x** (`Variable`) – Input variable.
- **gamma** (`Variable`) – Scaling parameter of normalized data.
- **beta** (`Variable`) – Shifting parameter of scaled normalized data.
- **mean** (`Variable`) – Shifting parameter of input.
- **var** (`Variable`) – Square of scaling parameter of input.
- **eps** (*float*) – Epsilon value for numerical stability.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation.

See also:

`functions.batch_normalization()`, `links.BatchNormalization`

local_response_normalization

`chainer.functions.`**`local_response_normalization`**(*x*, *n=5*, *k=2*, *alpha=0.0001*, *beta=0.75*)

Local response normalization across neighboring channels.

This function implements normalization across channels. Let x an input image with N channels. Then, this function computes an output image y by following formula:

$$y_i = \frac{x_i}{\left(k + \ \alpha \sum_{j=\max 1, i-n/2}^{\min N, i+n/2} x_j^2\right)^\beta}.$$

Parameters

- **x** (`Variable`) – Input variable.
- **n** (*int*) – Normalization window width.
- **k** (*float*) – Smoothing parameter.
- **alpha** (*float*) – Normalizer scaling parameter.
- **beta** (*float*) – Normalizer power parameter.

Returns Output variable.

Return type *Variable*

See: Section 3.3 of ImageNet Classification with Deep Convolutional Neural Networks

normalize

chainer.functions.**normalize**(*x*, *eps=1e-05*, *axis=1*)

L2 norm squared (a.k.a. Euclidean norm).

This function implements L2 normalization on a vector along the given axis. No reduction is done along the normalization axis.

In the case when `axis=1` and x is a vector of dimension (N, K), where N and K denote mini-batch size and the dimension of the input variable, this function computes an output vector y by the following equation:

$$y_i = \frac{x_i}{\|x_i\|_2 + \epsilon}$$

`eps` is used to avoid division by zero when norm of x along the given axis is zero.

The default value of `axis` is determined for backward compatibility.

Parameters

- **x** (Variable) – Two dimensional output variable. The first dimension is assumed to be the mini-batch dimension.
- **eps** (*float*) – Epsilon value for numerical stability.
- **axis** (*int*) – Axis along which to normalize.

Returns The output variable which has the same shape as x.

Return type *Variable*

Spatial pooling

average_pooling_2d

chainer.functions.**average_pooling_2d**(*x*, *ksize*, *stride=None*, *pad=0*, *use_cudnn=True*)

Spatial average pooling function.

This function acts similarly to `Convolution2D`, but it computes the average of input spatial patch for each channel without any parameter instead of computing the inner products.

Parameters

- **x** (Variable) – Input variable.
- **ksize** (*int or pair of ints*) – Size of pooling window. `ksize=k` and `ksize=(k, k)` are equivalent.
- **stride** (*int or pair of ints or None*) – Stride of pooling applications. `stride=s` and `stride=(s, s)` are equivalent. If `None` is specified, then it uses same stride as the pooling window size.
- **pad** (*int or pair of ints*) – Spatial padding width for the input array. `pad=p` and `pad=(p, p)` are equivalent.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable.

Return type *Variable*

Note: This function currently does not support `cover_all` mode as *`max_pooling_2d()`*. Average pooling runs in non-cover-all mode.

---

average_pooling_nd

`chainer.functions.`**`average_pooling_nd`**(*x*, *ksize*, *stride=None*, *pad=0*, *use_cudnn=True*)

N-dimensionally spatial average pooling function.

This function provides a N-dimensionally generalized version of `average_pooling_2d()`. This acts similarly to `ConvolutionND`, but it computes the average of input spatial patch for each channel without any parameter instead of computing the inner products.

Parameters

- **`x`** (`Variable`) – Input variable.
- **`ksize`** (*`int or tuple of ints`*) – Size of pooling window. `ksize=k` and `ksize=(k, k, ..., k)` are equivalent.
- **`stride`** (*`int or tuple of ints or None`*) – Stride of pooling applications. `stride=s` and `stride=(s, s, ..., s)` are equivalent. If `None` is specified, then it uses same stride as the pooling window size.
- **`pad`** (*`int or tuple of ints`*) – Spatial padding width for the input array. `pad=p` and `pad=(p, p, ..., p)` are equivalent.
- **`use_cudnn`** (*`bool`*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation. cuDNN supports more than one-dimensional pooling.

Returns Output variable.

Return type *Variable*

Note: This function currently does not support `cover_all` mode as *`max_pooling_nd()`*. Average pooling runs in non-cover-all mode.

---

max_pooling_2d

`chainer.functions.`**`max_pooling_2d`**(*x*, *ksize*, *stride=None*, *pad=0*, *cover_all=True*, *use_cudnn=True*)

Spatial max pooling function.

This function acts similarly to `Convolution2D`, but it computes the maximum of input spatial patch for each channel without any parameter instead of computing the inner products.

Parameters

- **`x`** (`Variable`) – Input variable.
- **`ksize`** (*`int or pair of ints`*) – Size of pooling window. `ksize=k` and `ksize=(k, k)` are equivalent.
- **`stride`** (*`int or pair of ints or None`*) – Stride of pooling applications. `stride=s` and `stride=(s, s)` are equivalent. If `None` is specified, then it uses same stride as the pooling window size.

- **pad** (*int or pair of ints*) – Spatial padding width for the input array. `pad=p` and `pad=(p, p)` are equivalent.
- **cover_all** (*bool*) – If `True`, all spatial locations are pooled into some output pixels. It may make the output size larger.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns Output variable.

Return type *Variable*

max_pooling_nd

`chainer.functions.`**max_pooling_nd**(*x*, *ksize*, *stride=None*, *pad=0*, *cover_all=True*, *use_cudnn=True*)

N-dimensionally spatial max pooling function.

This function provides a N-dimensionally generalized version of `max_pooling_2d()`. This acts similarly to `ConvolutionND`, but it computes the maximum of input spatial patch for each channel without any parameter instead of computing the inner products.

Parameters

- **x** (`Variable`) – Input variable.
- **ksize** (*int or tuple of ints*) – Size of pooling window. `ksize=k` and `ksize=(k, k, ..., k)` are equivalent.
- **stride** (*int or tuple of ints or None*) – Stride of pooling applications. `stride=s` and `stride=(s,s, ..., s)` are equivalent. If `None` is specified, then it uses same stride as the pooling window size.
- **pad** (*int or tuple of ints*) – Spatial padding width for the input array. `pad=p` and `pad=(p, p, ..., p)` are equivalent.
- **cover_all** (*bool*) – If `True`, all spatial locations are pooled into some output pixels. It may make the output size larger.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation. cuDNN supports more than one-dimensional pooling.

Returns Output variable.

Return type *Variable*

roi_pooling_2d

`chainer.functions.`**roi_pooling_2d**(*x*, *rois*, *outh*, *outw*, *spatial_scale*)

Spatial Region of Interest (ROI) pooling function.

This function acts similarly to `MaxPooling2D`, but it computes the maximum of input spatial patch for each channel with the region of interest.

Parameters

- **x** (`Variable`) – Input variable. The shape is expected to be 4 dimentional: (n: batch, c: channel, h, height, w: width).
- **rois** (`Variable`) – Input roi variable. The shape is expected to be (n: data size, 5), and each datum is set as below: (batch_index, x_min, y_min, x_max, y_max).

- **outh** (*int*) – Height of output image after pooled.
- **outw** (*int*) – Width of output image after pooled.
- **spatial_scale** (*float*) – Scale of the roi is resized.

Returns Output variable.

Return type *Variable*

See the original paper proposing ROIPooling: Fast R-CNN.

spatial_pyramid_pooling_2d

chainer.functions.**spatial_pyramid_pooling_2d**(*x*, *pyramid_height*, *pooling_class*, *use_cudnn=True*)

Spatial pyramid pooling function.

It outputs a fixed-length vector regardless of input feature map size.

It performs pooling operation to the input 4D-array `x` with different kernel sizes and padding sizes, and then flattens all dimensions except first dimension of all pooling results, and finally concatenates them along second dimension.

At i-th pyramid level, the kernel size $(k_h^{(i)}, k_w^{(i)})$ and padding size $(p_h^{(i)}, p_w^{(i)})$ of pooling operation are calculated as below:

$$
\begin{aligned}
k_h^{(i)} &= \lceil b_h / 2^i \rceil, \\
k_w^{(i)} &= \lceil b_w / 2^i \rceil, \\
p_h^{(i)} &= (2^i k_h^{(i)} - b_h)/2, \\
p_w^{(i)} &= (2^i k_w^{(i)} - b_w)/2,
\end{aligned}
$$

where $\lceil \cdot \rceil$ denotes the ceiling function, and b_h, b_w are height and width of input variable `x`, respectively. Note that index of pyramid level i is zero-based.

See detail in paper: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.

Parameters

- **x** (`Variable`) – Input variable. The shape of `x` should be `(batchsize, # of channels, height, width)`.
- **pyramid_height** (*int*) – Number of pyramid levels
- **pooling_class** (*MaxPooling2D or AveragePooling2D*) – Only MaxPooling2D class can be available for now.
- **use_cudnn** (*bool*) – If `True` and cuDNN is enabled, then this function uses cuDNN as the core implementation.

Returns

Output variable. The shape of the output variable will be $(batchsize, c\sum_{h=0}^{H-1} 2^{2h}, 1, 1)$, where c is the number of channels of input variable `x` and H is the number of pyramid levels.

Return type *Variable*

Note: This function uses some pooling classes as components to perform spatial pyramid pooling. Now it supports only `MaxPooling2D` as elemental pooling operator so far.

---

unpooling_2d

chainer.functions.**unpooling_2d**(*x*, *ksize*, *stride=None*, *pad=0*, *outsize=None*, *cover_all=True*)

Inverse operation of pooling for 2d array.

This function acts similarly to `Deconvolution2D`, but it spreads input 2d array's value without any parameter instead of computing the inner products.

Parameters

- **x** (*Variable*) – Input variable.
- **ksize** (*int or pair of ints*) – Size of pooling window. `ksize=k` and `ksize=(k, k)` are equivalent.
- **stride** (*int, pair of ints or None*) – Stride of pooling applications. `stride=s` and `stride=(s, s)` are equivalent. If `None` is specified, then it uses same stride as the pooling window size.
- **pad** (*int or pair of ints*) – Spatial padding width for the input array. `pad=p` and `pad=(p, p)` are equivalent.
- **outsize** (*None or pair of ints*) – Expected output size (height, width) of array after the operation. If `None`, the size (height or width) is estimated from the size of input array in first batch with `get_deconv_outsize()`. If outsize is not `None`, the result of outsize applied to `get_conv_outsize()` must be equal to the shape of the 2d array in the input batch `x`.
- **cover_all** (*bool*) – If `True`, the output size may be smaller than the size if `cover_all` is `False`. This flag serves to align behavior to the pooling functions which can cover all input locations, see *max_pooling_2d()* and *convolution_2d()*.

Returns Output variable.

Return type *Variable*

upsampling_2d

chainer.functions.**upsampling_2d**(*x*, *indexes*, *ksize*, *stride=None*, *pad=0*, *outsize=None*, *cover_all=True*)

Upsampling using pooling indices.

This function produces an upsampled image using pooling indices.

Example

It should be noted that you need to specify `use_cudnn=False` when you create MaxPooling2D object because if cuDNN used for operating max pooling, `indexes` is never created and stored in the MaxPooling2D object.

```
>>> p = F.MaxPooling2D(2, 2, use_cudnn=False)
>>> x = np.arange(1, 37).reshape(1, 1, 6, 6).astype('f')
>>> x = chainer.Variable(x)
>>> x.data
array([[[[  1.,   2.,   3.,   4.,   5.,   6.],
         [  7.,   8.,   9.,  10.,  11.,  12.],
         [ 13.,  14.,  15.,  16.,  17.,  18.],
         [ 19.,  20.,  21.,  22.,  23.,  24.],
         [ 25.,  26.,  27.,  28.,  29.,  30.],
         [ 31.,  32.,  33.,  34.,  35.,  36.]]]], dtype=float32)
```

This is the original x before max pooling.

```
>>> pooled_x = p(x)
>>> pooled_x.data
array([[[[  8.,  10.,  12.],
         [ 20.,  22.,  24.],
         [ 32.,  34.,  36.]]]], dtype=float32)
```

This is the output of the max pooling operation. upsampling_2d needs indexes array stored in the max pooling object p.

```
>>> upsampled_x = F.upsampling_2d(
...     pooled_x, p.indexes, p.kh, p.sy, p.ph, x.shape[2:])
>>> upsampled_x.shape
(1, 1, 6, 6)
>>> upsampled_x.data
array([[[[  0.,   0.,   0.,   0.,   0.,   0.],
         [  0.,   8.,   0.,  10.,   0.,  12.],
         [  0.,   0.,   0.,   0.,   0.,   0.],
         [  0.,  20.,   0.,  22.,   0.,  24.],
         [  0.,   0.,   0.,   0.,   0.,   0.],
         [  0.,  32.,   0.,  34.,   0.,  36.]]]], dtype=float32)
```

Parameters

- **x** (*Variable*) – Input variable.
- **indexes** (*ndarray or ndarray*) – Index array that was used to calculate x with MaxPooling2D.
- **ksize** (*int or (int, int)*) – ksize attribute of MaxPooling2D object that is used to calculate x
- **stride** (*int or (int, int)*) – stride attribute of MaxPooling2D object that is used to calculate x
- **pad** (*int or (int, int)*) – pad attribute of MaxPooling2D object that is used to calculate x
- **outsize** (*(int, int)*) – Expected output size (height, width).
- **cover_all** (*bool*) – Whether cover_all is used in the MaxPooling2D object or not.

Returns Output variable.

Return type *Variable*

Utility functions

forget

chainer.functions.**forget**(*func*, **xs*)

Call a function without storing internal results.

On a forward propagation Chainer stores all internal results of Function on a computational graph as they are required on backward-propagation. These results consume too much memory when the internal results are too large. This method **forgets** such internal results on forward propagation, and still supports back-propagation with recalculation.

In a forward propagation, this method calls a given function with given variables without creating a computational graph. That means, no internal results are stored. In a backward propagation this method calls the given function again to create a computational graph to execute back-propagation.

This method reduces internal memory usage. Instead it requires more calculation time as it calls the function twice.

Example

Let `f` be a function defined as:

```
>>> def f(a, b):
...   return a + b + a
```

and, `x` and `y` be *Variable*:

```
>>> x = chainer.Variable(np.random.uniform(-1, 1, 5).astype('f'))
>>> y = chainer.Variable(np.random.uniform(-1, 1, 5).astype('f'))
```

When `z` is calculated as `z = f(x, y)`, its internal result `x + y` is stored in memory. Instead if you call `f` with *forget()*:

```
>>> z = F.forget(f, x, y)
```

internal `x + y` is forgotten.

Note: The method does not support functions behaving randomly, such as *dropout()* and *negative_sampling()*. It is because first results of these function differ from the second one.

Parameters

- **func** (*callable*) – A function to call. It needs to be called with *Variable* object(s) and to return a *Variable* object or a tuple of *Variable* objects.
- **xs** (Variable) – Argument variables of the function.

Returns A variable `func` returns. If it returns a tuple, the method returns a tuple too.

Return type *Variable*

Standard Link implementations

Chainer provides many *Link* implementations in the *chainer.links* package.

Note: Some of the links are originally defined in the *chainer.functions* namespace. They are still left in the namespace for backward compatibility, though it is strongly recommended to use them via the *chainer.links* package.

Learnable connections

Bias

class chainer.links.**Bias**(*axis=1*, *shape=None*)
Broadcasted elementwise summation with learnable parameters.

Computes a elementwise summation as *bias()* function does except that its second input is a learnable bias parameter b the link has.

Parameters

- **axis** (*int*) – The first axis of the first input of *bias()* function along which its second input is applied.
- **shape** (*tuple of ints*) – Shape of the learnable bias parameter. If None, this link does not have learnable parameters so an explicit bias needs to be given to its __call__ method's second input.

See also:

See *bias()* for details.

Variables **b** (Variable) – Bias parameter if shape is given. Otherwise, no attributes.

Bilinear

class chainer.links.**Bilinear**(*left_size*, *right_size*, *out_size*, *nobias=False*, *initialW=None*, *initial_bias=None*)
Bilinear layer that performs tensor multiplication.

Bilinear is a primitive link that wraps the *bilinear()* functions. It holds parameters W, V1, V2, and b corresponding to the arguments of *bilinear()*.

Parameters

- **left_size** (*int*) – Dimension of input vector e^1 (J)
- **right_size** (*int*) – Dimension of input vector e^2 (K)
- **out_size** (*int*) – Dimension of output vector y (L)
- **nobias** (*bool*) – If True, parameters V1, V2, and b are omitted.
- **initialW** (*3-D numpy array*) – Initial value of W. Shape of this argument must be (left_size, right_size, out_size). If None, W is initialized by centered Gaussian distribution properly scaled according to the dimension of inputs and outputs. May also be a callable that takes numpy.ndarray or cupy.ndarray and edits its value.
- **initial_bias** (*tuple*) – Initial values of V^1, V^2 and b. The length this argument must be 3. Each element of this tuple must have the shapes of (left_size, output_size), (right_size, output_size), and (output_size,), respectively. If None, V^1 and V^2 is initialized by scaled centered Gaussian distributions and b is set to 0. May also be a tuple of callables that take numpy.ndarray or cupy.ndarray and edit its value.

See also:

See *chainer.functions.bilinear()* for details.

Variables

- **W** (Variable) – Bilinear weight parameter.

- **V1** (Variable) – Linear weight parameter for the first argument.
- **V2** (Variable) – Linear weight parameter for the second argument.
- **b** (Variable) – Bias parameter.

Convolution2D

class chainer.links.**Convolution2D**(*in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *wscale=1*, *bias=0*, *nobias=False*, *use_cudnn=True*, *initialW=None*, *initial_bias=None*, *deterministic=False*)

Two-dimensional convolutional layer.

This link wraps the *convolution_2d()* function and holds the filter weight and bias vector as parameters.

Parameters

- **in_channels** (*int*) – Number of channels of input arrays. If None, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_channels** (*int*) – Number of channels of output arrays.
- **ksize** (*int or pair of ints*) – Size of filters (a.k.a. kernels). ksize=k and ksize=(k, k) are equivalent.
- **stride** (*int or pair of ints*) – Stride of filter applications. stride=s and stride=(s, s) are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. pad=p and pad=(p, p) are equivalent.
- **wscale** (*float*) – Scaling factor of the initial weight.
- **bias** (*float*) – Initial bias value.
- **nobias** (*bool*) – If True, then this link does not use the bias term.
- **use_cudnn** (*bool*) – If True, then this link uses cuDNN if available.
- **initialW** (*4-D array*) – Initial weight value. If None, then this function uses Gaussian distribution scaled by w_scale to initialize weight. May also be a callable that takes numpy.ndarray or cupy.ndarray and edits its value.
- **initial_bias** (*1-D array*) – Initial bias value. If None, then this function uses bias to initialize bias. May also be a callable that takes numpy.ndarray or cupy.ndarray and edits its value.
- **deterministic** (*bool*) – The output of this link can be non-deterministic when it uses cuDNN. If this option is True, then it forces cuDNN to use a deterministic algorithm. This option is only available for cuDNN version >= v4.

See also:

See *chainer.functions.convolution_2d()* for the definition of two-dimensional convolution.

Variables

- **W** (Variable) – Weight parameter.
- **b** (Variable) – Bias parameter.

ConvolutionND

class chainer.links.**ConvolutionND**(*ndim*, *in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *initialW=None*, *initial_bias=None*, *use_cudnn=True*, *cover_all=False*)

N-dimensional convolution layer.

This link wraps the *convolution_nd()* function and holds the filter weight and bias vector as parameters.

Parameters

- **ndim** (*int*) – Number of spatial dimensions.
- **in_channels** (*int*) – Number of channels of input arrays.
- **out_channels** (*int*) – Number of channels of output arrays.
- **ksize** (*int or tuple of ints*) – Size of filters (a.k.a. kernels). `ksize=k` and `ksize=(k, k, ..., k)` are equivalent.
- **stride** (*int or tuple of ints*) – Stride of filter application. `stride=s` and `stride=(s, s, ..., s)` are equivalent.
- **pad** (*int or tuple of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p, ..., p)` are equivalent.
- **initialW** – Value used to initialize the filter weight. May be an initializer instance or another value that *init_weight()* helper function can take.
- **initial_bias** – Value used to initialize the bias vector. May be an initializer instance or another value except `None` that *init_weight()* helper function can take. If `None` is given, this link does not use the bias vector.
- **use_cudnn** (*bool*) – If `True`, then this link uses cuDNN if available. See *convolution_nd()* for exact conditions of cuDNN availability.
- **cover_all** (*bool*) – If `True`, all spatial locations are convoluted into some output pixels. It may make the output size larger. `cover_all` needs to be `False` if you want to use cuDNN.

See also:

See *convolution_nd()* for the definition of N-dimensional convolution. See *convolution_2d()* for the definition of two-dimensional convolution.

Variables

- **W** (`Variable`) – Weight parameter.
- **b** (`Variable`) – Bias parameter. If `initial_bias` is `None`, set to `None`.

Deconvolution2D

class chainer.links.**Deconvolution2D**(*in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *wscale=1*, *bias=0*, *nobias=False*, *outsize=None*, *use_cudnn=True*, *initialW=None*, *initial_bias=None*, *deterministic=False*)

Two dimensional deconvolution function.

This link wraps the *deconvolution_2d()* function and holds the filter weight and bias vector as parameters.

Parameters

- **in_channels** (*int*) – Number of channels of input arrays. If `None`, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_channels** (*int*) – Number of channels of output arrays.
- **ksize** (*int or pair of ints*) – Size of filters (a.k.a. kernels). `ksize=k` and `ksize=(k, k)` are equivalent.
- **stride** (*int or pair of ints*) – Stride of filter applications. `stride=s` and `stride=(s, s)` are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p)` are equivalent.
- **wscale** (*float*) – Scaling factor of the initial weight.
- **bias** (*float*) – Initial bias value.
- **nobias** (*bool*) – If `True`, then this function does not use the bias term.
- **outsize** (*tuple*) – Expected output size of deconvolutional operation. It should be pair of height and width (out_H, out_W). Default value is `None` and the outsize is estimated by input size, stride and pad.
- **use_cudnn** (*bool*) – If `True`, then this function uses cuDNN if available.
- **initialW** (*4-D array*) – Initial weight value. If `None`, then this function uses Gaussian distribution scaled by `w_scale` to initialize weight. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **initial_bias** (*1-D array*) – Initial bias value. If `None`, then this function uses `bias` to initialize bias. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **deterministic** (*bool*) – The output of this link can be non-deterministic when it uses cuDNN. If this option is `True`, then it forces cuDNN to use a deterministic algorithm. This option is only available for cuDNN version >= v4.

The filter weight has four dimensions (c_I, c_O, k_H, k_W) which indicate the number of input channels, output channels, height and width of the kernels, respectively. The filter weight is initialized with i.i.d. Gaussian random samples, each of which has zero mean and deviation $\sqrt{1/(c_I k_H k_W)}$ by default. The deviation is scaled by `wscale` if specified.

The bias vector is of size c_O. Its elements are initialized by `bias` argument. If `nobias` argument is set to True, then this function does not hold the bias parameter.

See also:

See *chainer.functions.deconvolution_2d()* for the definition of two-dimensional convolution.

DeconvolutionND

class `chainer.links.`**DeconvolutionND**(*ndim*, *in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *outsize=None*, *initialW=None*, *initial_bias=0*, *use_cudnn=True*)

N-dimensional deconvolution function.

This link wraps *deconvolution_nd()* function and holds the filter weight and bias vector as its parameters.

Parameters

- **ndim** (*int*) – Number of spatial dimensions.

- **in_channels** (*int*) – Number of channels of input arrays.
- **out_channels** (*int*) – Number of channels of output arrays.
- **ksize** (*int or tuple of ints*) – Size of filters (a.k.a. kernels). `ksize=k` and `ksize=(k, k, ..., k)` are equivalent.
- **stride** (*int or tuple of ints*) – Stride of filter application. `stride=s` and `stride=(s, s, ..., s)` are equivalent.
- **pad** (*int or tuple of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p, ..., p)` are equivalent.
- **outsize** (*tuple of ints*) – Expected output size of deconvolutional operation. It should be a tuple of ints that represents the output size of each dimension. Default value is `None` and the outsize is estimated with input size, stride and pad.
- **initialW** – Value used to initialize the filter weight. May be an initializer instance of another value the same with that *init_weight()* function can take.
- **initial_bias** – Value used to initialize the bias vector. May be an initializer instance or another value except `None` the same with that *init_weight()* function can take. If `None` is supplied, this link does not use the bias vector.
- **use_cudnn** (*bool*) – If `True`, then this link uses cuDNN if available.

See also:

deconvolution_nd()

Variables

- **W** (`Variable`) – Weight parameter.
- **b** (`Variable`) – Bias parameter. If `initial_bias` is `None`, set to `None`.

DilatedConvolution2D

class `chainer.links.`**DilatedConvolution2D**(*in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *dilate=1*, *wscale=1*, *bias=0*, *nobias=False*, *use_cudnn=True*, *initialW=None*, *initial_bias=None*)

Two-dimensional dilated convolutional layer.

This link wraps the *dilated_convolution_2d()* function and holds the filter weight and bias vector as parameters.

Parameters

- **in_channels** (*int*) – Number of channels of input arrays. If `None`, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_channels** (*int*) – Number of channels of output arrays.
- **ksize** (*int or pair of ints*) – Size of filters (a.k.a. kernels). `ksize=k` and `ksize=(k, k)` are equivalent.
- **stride** (*int or pair of ints*) – Stride of filter applications. `stride=s` and `stride=(s, s)` are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays. `pad=p` and `pad=(p, p)` are equivalent.

- **dilate** (*int or pair of ints*) – Dilation factor of filter applications. `dilate=d` and `dilate=(d, d)` are equivalent.
- **wscale** (*float*) – Scaling factor of the initial weight.
- **bias** (*float*) – Initial bias value.
- **nobias** (*bool*) – If `True`, then this link does not use the bias term.
- **use_cudnn** (*bool*) – If `True`, then this link uses cuDNN if available.
- **initialW** (*4-D array*) – Initial weight value. If `None`, then this function uses scaled Gaussian distribution to initialize weight. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **initial_bias** (*1-D array*) – Initial bias value. If `None`, then this function uses `bias` to initialize bias. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.

See also:

See *chainer.functions.dilated_convolution_2d()* for the definition of two-dimensional dilated convolution.

Variables

- **W** (`Variable`) – Weight parameter.
- **b** (`Variable`) – Bias parameter.

EmbedID

class `chainer.links.`**EmbedID**(*in_size*, *out_size*, *initialW=None*, *ignore_label=None*)
Efficient linear layer for one-hot input.

This is a link that wraps the *embed_id()* function. This link holds the ID (word) embedding matrix `W` as a parameter.

Parameters

- **in_size** (*int*) – Number of different identifiers (a.k.a. vocabulary size).
- **out_size** (*int*) – Size of embedding vector.
- **initialW** (*2-D array*) – Initial weight value. If `None`, then the matrix is initialized from the standard normal distribution. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **ignore_label** (*int or None*) – If `ignore_label` is an int value, `i`-th column of return value is filled with `0`.

See also:

chainer.functions.embed_id()

Variables **W** (`Variable`) – Embedding parameter matrix.

GRU

class `chainer.links.`**GRU**(*n_units*, *n_inputs=None*, *init=None*, *inner_init=None*, *bias_init=0*)
Stateless Gated Recurrent Unit function (GRU).

GRU function has six parameters W_r, W_z, W, U_r, U_z, and U. All these parameters are $n \times n$ matrices, where n is the dimension of hidden vectors.

Given two inputs a previous hidden vector h and an input vector x, GRU returns the next hidden vector h' defined as

$$\begin{aligned} r &= \sigma(W_r x + U_r h), \\ z &= \sigma(W_z x + U_z h), \\ \bar{h} &= \tanh(W x + U(r \odot h)), \\ h' &= (1 - z) \odot h + z \odot \bar{h}, \end{aligned}$$

where σ is the sigmoid function, and $\odot$ is the element-wise product.

GRU does not hold the value of hidden vector h. So this is *stateless*. Use *StatefulGRU* as a *stateful* GRU.

Parameters

- **n_units** (*int*) – Dimension of hidden vector h.
- **n_inputs** (*int*) – Dimension of input vector x. If None, it is set to the same value as n_units.

See:

- On the Properties of Neural Machine Translation: Encoder-Decoder Approaches [Cho+, SSST2014].
- Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [Chung+NIPS2014 DLWorkshop].

See also:

StatefulGRU

Highway

class chainer.links.**Highway**(*in_out_size*, *nobias=False*, *activate=<function relu>*, *init_Wh=None*, *init_Wt=None*, *init_bh=None*, *init_bt=-1*)

Highway module.

In highway network, two gates are added to the ordinal non-linear transformation ($H(x) = activate(W_h x + b_h)$). One gate is the transform gate $T(x) = \sigma(W_t x + b_t)$, and the other is the carry gate $C(x)$. For simplicity, the author defined $C = 1 - T$. Highway module returns y defined as

$$y = activate(W_h x + b_h) \odot \sigma(W_t x + b_t) + x \odot (1 - \sigma(W_t x + b_t))$$

The output array has the same spatial size as the input. In order to satisfy this, W_h and W_t must be square matrices.

Parameters

- **in_out_size** (*int*) – Dimension of input and output vectors.
- **nobias** (*bool*) – If True, then this function does not use the bias.
- **activate** – Activation function of plain array. $tanh$ is also available.

- **init_Wh** (*2-D array*) – Initial weight value of plain array. If `None`, then this function uses Gaussian distribution scaled by `w_scale` to initialize W_h. May also be a callable that takes `numpy.ndarray` or‘‘cupy.ndarray‘‘ and edits its value.
- **init_bh** (*1-D array*) – Initial bias value of plain array. If `None`, then this function uses zero vector to initialize b_h. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **init_Wt** (*2-D array*) – Initial weight value of transform array. If `None`, then this function uses Gaussian distribution scaled by `w_scale` to initialize W_t. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **init_bt** (*1-D array*) – Initial bias value of transform array. Default value is -1 vector. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. Negative value is recommended by the author of the paper. (e.g. -1, -3, ...).

See: Highway Networks.

Inception

class `chainer.links.`**Inception**(*in_channels*, *out1*, *proj3*, *out3*, *proj5*, *out5*, *proj_pool*, *conv_init=None*, *bias_init=None*)

Inception module of GoogLeNet.

It applies four different functions to the input array and concatenates their outputs along the channel dimension. Three of them are 2D convolutions of sizes 1x1, 3x3 and 5x5. Convolution paths of 3x3 and 5x5 sizes have 1x1 convolutions (called projections) ahead of them. The other path consists of 1x1 convolution (projection) and 3x3 max pooling.

The output array has the same spatial size as the input. In order to satisfy this, Inception module uses appropriate padding for each convolution and pooling.

See: Going Deeper with Convolutions.

Parameters

- **in_channels** (*int*) – Number of channels of input arrays.
- **out1** (*int*) – Output size of 1x1 convolution path.
- **proj3** (*int*) – Projection size of 3x3 convolution path.
- **out3** (*int*) – Output size of 3x3 convolution path.
- **proj5** (*int*) – Projection size of 5x5 convolution path.
- **out5** (*int*) – Output size of 5x5 convolution path.
- **proj_pool** (*int*) – Projection size of max pooling path.
- **conv_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the convolution matrix weights. Maybe be `None` to use default initialization.
- **bias_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the convolution bias weights. Maybe be `None` to use default initialization.

InceptionBN

class chainer.links.**InceptionBN**(*in_channels*, *out1*, *proj3*, *out3*, *proj33*, *out33*, *pooltype*, *proj_pool=None*, *stride=1*, *conv_init=None*, *dtype=<type 'numpy.float32'>*)

Inception module of the new GoogLeNet with BatchNormalization.

This chain acts like *Inception*, while InceptionBN uses the *BatchNormalization* on top of each convolution, the 5x5 convolution path is replaced by two consecutive 3x3 convolution applications, and the pooling method is configurable.

See: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

Parameters

- **in_channels** (*int*) – Number of channels of input arrays.
- **out1** (*int*) – Output size of the 1x1 convolution path.
- **proj3** (*int*) – Projection size of the single 3x3 convolution path.
- **out3** (*int*) – Output size of the single 3x3 convolution path.
- **proj33** (*int*) – Projection size of the double 3x3 convolutions path.
- **out33** (*int*) – Output size of the double 3x3 convolutions path.
- **pooltype** (*str*) – Pooling type. It must be either `'max'` or `'avg'`.
- **proj_pool** (*bool*) – If `True`, do projection in the pooling path.
- **stride** (*int*) – Stride parameter of the last convolution of each path.
- **conv_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the convolution matrix weights. Maybe be `None` to use default initialization.
- **dtype** (*numpy.dtype*) – Type to use in `~batch_normalization.BatchNormalization`.

See also:

Inception

Variables **train** (*bool*) – If `True`, then batch normalization layers are used in training mode. If `False`, they are used in testing mode.

Linear

class chainer.links.**Linear**(*in_size*, *out_size*, *wscale=1*, *bias=0*, *nobias=False*, *initialW=None*, *initial_bias=None*)

Linear layer (a.k.a. fully-connected layer).

This is a link that wraps the *linear()* function, and holds a weight matrix `W` and optionally a bias vector `b` as parameters.

The weight matrix `W` is initialized with i.i.d. Gaussian samples, each of which has zero mean and deviation $\sqrt{1/}$

Parameters

- **in_size** (*int*) – Dimension of input vectors. If `None`, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_size** (*int*) – Dimension of output vectors.
- **wscale** (*float*) – Scaling factor of the weight matrix.

- **bias** (*float*) – Initial bias value.
- **nobias** (*bool*) – If `True`, then this function does not use the bias.
- **initialW** (*2-D array*) – Initial weight value. If `None`, then this function uses Gaussian distribution scaled by `w_scale` to initialize weight. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.
- **initial_bias** (*1-D array*) – Initial bias value. If `None`, then this function uses `bias` to initialize bias. May also be a callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value.

See also:

linear()

Variables

- **W** (`Variable`) – Weight parameter.
- **b** (`Variable`) – Bias parameter.

LSTM

class `chainer.links.`**LSTM**(*in_size*, *out_size*, ***kwargs*)

Fully-connected LSTM layer.

This is a fully-connected LSTM layer as a chain. Unlike the *lstm()* function, which is defined as a stateless activation function, this chain holds upward and lateral connections as child links.

It also maintains *states*, including the cell state and the output at the previous time step. Therefore, it can be used as a *stateful LSTM*.

This link supports variable length inputs. The mini-batch size of the current input must be equal to or smaller than that of the previous one. The mini-batch size of `c` and `h` is determined as that of the first input `x`. When mini-batch size of `i`-th input is smaller than that of the previous input, this link only updates `c[0:len(x)]` and `h[0:len(x)]` and doesn't change the rest of `c` and `h`. So, please sort input sequences in descending order of lengths before applying the function.

Parameters

- **in_size** (*int*) – Dimension of input vectors. If `None`, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_size** (*int*) – Dimensionality of output vectors.
- **lateral_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the lateral connections. Maybe be `None` to use default initialization.
- **upward_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the upward connections. Maybe be `None` to use default initialization.
- **bias_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value It is used for initialization of the biases of cell input, input gate and output gate.and gates of the upward connection. Maybe a scalar, in that case, the bias is initialized by this value. Maybe be `None` to use default initialization.
- **forget_bias_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value It is used for initialization of the biases of the forget gate of the upward connection. Maybe a scalar, in that case, the bias is initialized by this value. Maybe be `None` to use default initialization.

Variables

- **upward** (Linear) – Linear layer of upward connections.
- **lateral** (Linear) – Linear layer of lateral connections.
- **c** (Variable) – Cell states of LSTM units.
- **h** (Variable) – Output at the previous time step.

reset_state()
Resets the internal state.

It sets None to the c and h attributes.

set_state(*c*, *h*)
Sets the internal state.

It sets the c and h attributes.

Parameters

- **c** (Variable) – A new cell states of LSTM units.
- **h** (Variable) – A new output at the previous time step.

MLPConvolution2D

class chainer.links.**MLPConvolution2D**(*in_channels*, *out_channels*, *ksize*, *stride=1*, *pad=0*, *wscale=1*, *activation=<function relu>*, *use_cudnn=True*, *conv_init=None*, *bias_init=None*)
Two-dimensional MLP convolution layer of Network in Network.

This is an "mlpconv" layer from the Network in Network paper. This layer is a two-dimensional convolution layer followed by 1x1 convolution layers and interleaved activation functions.

Note that it does not apply the activation function to the output of the last 1x1 convolution layer.

Parameters

- **in_channels** (*int or None*) – Number of channels of input arrays. If None, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_channels** (*tuple of ints*) – Tuple of number of channels. The i-th integer indicates the number of filters of the i-th convolution.
- **ksize** (*int or pair of ints*) – Size of filters (a.k.a. kernels) of the first convolution layer. ksize=k and ksize=(k, k) are equivalent.
- **stride** (*int or pair of ints*) – Stride of filter applications at the first convolution layer. stride=s and stride=(s, s) are equivalent.
- **pad** (*int or pair of ints*) – Spatial padding width for input arrays at the first convolution layer. pad=p and pad=(p, p) are equivalent.
- **activation** (function) – Activation function for internal hidden units. Note that this function is not applied to the output of this link.
- **use_cudnn** (*bool*) – If True, then this link uses cuDNN if available.
- **conv_init** – An initializer of weight matrices passed to the convolution layers.
- **bias_init** – An initializer of bias vectors passed to the convolution layers.

See: Network in Network.

Variables **activation** (function) – Activation function.

NStepLSTM

class chainer.links.**NStepLSTM**(*n_layers*, *in_size*, *out_size*, *dropout*, *use_cudnn=True*)
Stacked LSTM for sequnces.

This link is stacked version of LSTM for sequences. It calculates hidden and cell states of all layer at end-of-string, and all hidden states of the last layer for each time.

Unlike *chainer.functions.n_step_lstm()*, this function automatically sort inputs in descending order by length, and transpose the seuqnece. Users just need to call the link with a list of *chainer.Variable* holding sequences.

Parameters

- **n_layers** (*int*) – Number of layers.
- **in_size** (*int*) – Dimensionality of input vectors.
- **out_size** (*int*) – Dimensionality of hidden states and output vectors.
- **dropout** (*float*) – Dropout ratio.
- **use_cudnn** (*bool*) – Use cuDNN.

See also:

chainer.functions.n_step_lstm()

Scale

class chainer.links.**Scale**(*axis=1*, *W_shape=None*, *bias_term=False*, *bias_shape=None*)
Broadcasted elementwise product with learnable parameters.

Computes a elementwise product as *scale()* function does except that its second input is a learnable weight parameter W the link has.

Parameters

- **axis** (*int*) – The first axis of the first input of *scale()* function along which its second input is applied.
- **W_shape** (*tuple of ints*) – Shape of learnable weight parameter. If None, this link does not have learnable weight parameter so an explicit weight needs to be given to its __call__ method's second input.
- **bias_term** (*bool*) – Whether to also learn a bias (equivalent to Scale link + Bias link).
- **bias_shape** (*tuple of ints*) – Shape of learnable bias. If W_shape is None, this should be given to determine the shape. Otherwise, the bias has the same shape W_shape with the weight parameter and bias_shape is ignored.

See also:

See *scale()* for details.

Variables

- **W** (Variable) – Weight parameter if W_shape is given. Otherwise, no W attribute.
- ***bias*** (Bias) – Bias term if bias_term is True. Otherwise, no bias attribute.

StatefulGRU

class chainer.links.**StatefulGRU**(*in_size*, *out_size*, *init=None*, *inner_init=None*, *bias_init=0*)
Stateful Gated Recurrent Unit function (GRU).

Stateful GRU function has six parameters W_r, W_z, W, U_r, U_z, and U. All these parameters are $n \times n$ matrices, where n is the dimension of hidden vectors.

Given input vector x, Stateful GRU returns the next hidden vector h' defined as

$$\begin{aligned} r &= \sigma(W_r x + U_r h), \\ z &= \sigma(W_z x + U_z h), \\ \bar{h} &= \tanh(W x + U(r \odot h)), \\ h' &= (1 - z) \odot h + z \odot \bar{h}, \end{aligned}$$

where h is current hidden vector.

As the name indicates, *StatefulGRU* is *stateful*, meaning that it also holds the next hidden vector *h'* as a state. Use *GRU* as a stateless version of GRU.

Parameters

- **in_size** (*int*) – Dimension of input vector x.
- **out_size** (*int*) – Dimension of hidden vector h.
- **init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the GRU's input units (W). Maybe be *None* to use default initialization.
- **inner_init** – A callable that takes `numpy.ndarray` or `cupy.ndarray` and edits its value. It is used for initialization of the GRU's inner recurrent units (U). Maybe be `None` to use default initialization.
- **bias_init** – A callable or scalar used to initialize the bias values for both the GRU's inner and input units. Maybe be `None` to use default initialization.

Variables **h** (*Variable*) – Hidden vector that indicates the state of *StatefulGRU*.

See also:

GRU

StatefulPeepholeLSTM

class chainer.links.**StatefulPeepholeLSTM**(*in_size*, *out_size*)
Fully-connected LSTM layer with peephole connections.

This is a fully-connected LSTM layer with peephole connections as a chain. Unlike the *LSTM* link, this chain holds `peep_i`, `peep_f` and `peep_o` as child links besides `upward` and `lateral`.

Given a input vector x, Peephole returns the next hidden vector h' defined as

$$
\begin{aligned}
a &= \\
&\tanh(upwardx + lateralh), \\
i &= \\
&\sigma(upwardx + lateralh + peep_i c), \\
f &= \\
&\sigma(upwardx + lateralh + peep_f c), \\
c' &= \\
&a \odot i + f \odot c, \\
o &= \\
&\sigma(upwardx + lateralh + peep_o c'), \\
h' &= \\
&o \tanh(c'),
\end{aligned}
$$

where σ is the sigmoid function, $\odot$ is the element-wise product, c is the current cell state, c' is the next cell state and h is the current hidden vector.

Parameters

- **in_size** (*int*) – Dimension of the input vector x.
- **out_size** (*int*) – Dimension of the hidden vector h.

Variables

- **upward** (Linear) – Linear layer of upward connections.
- **lateral** (Linear) – Linear layer of lateral connections.
- **peep_i** (Linear) – Linear layer of peephole connections to the input gate.
- **peep_f** (Linear) – Linear layer of peephole connections to the forget gate.
- **peep_o** (Linear) – Linear layer of peephole connections to the output gate.
- **c** (Variable) – Cell states of LSTM units.
- **h** (Variable) – Output at the current time step.

reset_state()

Resets the internal states.

It sets None to the c and h attributes.

StatelessLSTM

class chainer.links.**StatelessLSTM**(*in_size*, *out_size*, *lateral_init=None*, *upward_init=None*, *bias_init=0*, *forget_bias_init=0*)

Stateless LSTM layer.

This is a fully-connected LSTM layer as a chain. Unlike the *lstm()* function, this chain holds upward and lateral connections as child links. This link doesn't keep cell and hidden states.

Parameters

- **in_size** (*int*) – Dimension of input vectors. If None, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **out_size** (*int*) – Dimensionality of output vectors.

Variables

- **upward** (chainer.links.Linear) – Linear layer of upward connections.
- **lateral** (chainer.links.Linear) – Linear layer of lateral connections.

Activation/loss/normalization functions with parameters

BatchNormalization

class chainer.links.**BatchNormalization**(*size*, *decay=0.9*, *eps=2e-05*, *dtype=<type 'numpy.float32'>*, *use_gamma=True*, *use_beta=True*, *initial_gamma=None*, *initial_beta=None*, *use_cudnn=True*)

Batch normalization layer on outputs of linear or convolution functions.

This link wraps the *batch_normalization()* and *fixed_batch_normalization()* functions.

It runs in three modes: training mode, fine-tuning mode, and testing mode.

In training mode, it normalizes the input by *batch statistics*. It also maintains approximated population statistics by moving averages, which can be used for instant evaluation in testing mode.

In fine-tuning mode, it accumulates the input to compute *population statistics*. In order to correctly compute the population statistics, a user must use this mode to feed mini-batches running through whole training dataset.

In testing mode, it uses pre-computed population statistics to normalize the input variable. The population statistics is approximated if it is computed by training mode, or accurate if it is correctly computed by fine-tuning mode.

Parameters

- **size** (*int or tuple of ints*) – Size (or shape) of channel dimensions.
- **decay** (*float*) – Decay rate of moving average. It is used on training.
- **eps** (*float*) – Epsilon value for numerical stability.
- **dtype** (*numpy.dtype*) – Type to use in computing.
- **use_gamma** (*bool*) – If True, use scaling parameter. Otherwise, use unit(1) which makes no effect.
- **use_beta** (*bool*) – If True, use shifting parameter. Otherwise, use unit(0) which makes no effect.
- **use_cudnn** (*bool*) – If True, then this link uses cuDNN if available.

See: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

See also:

batch_normalization(), *fixed_batch_normalization()*

Variables

- **gamma** (Variable) – Scaling parameter.
- **beta** (Variable) – Shifting parameter.
- **avg_mean** (Variable) – Population mean.
- **avg_var** (Variable) – Population variance.
- **N** (*int*) – Count of batches given for fine-tuning.

- **decay** (*float*) – Decay rate of moving average. It is used on training.
- **eps** (*float*) – Epsilon value for numerical stability. This value is added to the batch variances.
- **use_cudnn** (*bool*) – If `True`, then this link uses cuDNN if available.

start_finetuning()

Resets the population count for collecting population statistics.

This method can be skipped if it is the first time to use the fine-tuning mode. Otherwise, this method should be called before starting the fine-tuning mode again.

LayerNormalization

class chainer.links.**LayerNormalization**(*size=None*, *eps=1e-06*, *initial_gamma=None*, *initial_beta=None*)

Layer normalization layer on outputs of linear functions.

This link implements a "layer normalization" layer which normalizes the input units by statistics that are computed along the second axis, scales and shifts them. Parameter initialization will be deferred until the first forward data pass at which time the size will be determined.

Parameters

- **size** (*int*) – Size of input units. If `None`, parameter initialization will be deferred until the first forward data pass at which time the size will be determined.
- **eps** (*float*) – Epsilon value for numerical stability of normalization.
- **initial_gamma** (*Initializer*) – Initializer for scaling vector. If `None`, then the vector is filled by 1. If a scalar, the vector is filled by it. If `numpy.ndarray`, the vector is set by it.
- **initial_beta** (*Initializer*) – Initializer for shifting vector. If `None`, then the vector is filled by 0. If a scalar, the vector is filled by it. If `numpy.ndarray`, the vector is set by it.

Variables

- **gamma** (`Variable`) – Scaling parameter.
- **beta** (`Variable`) – Shifting parameter.
- **eps** (*float*) – Epsilon value for numerical stability.

See: Layer Normalization

BinaryHierarchicalSoftmax

class chainer.links.**BinaryHierarchicalSoftmax**(*in_size*, *tree*)

Hierarchical softmax layer over binary tree.

In natural language applications, vocabulary size is too large to use softmax loss. Instead, the hierarchical softmax uses product of sigmoid functions. It costs only $O(\log(n))$ time where n is the vocabulary size in average.

At first a user need to prepare a binary tree whose each leaf is corresponding to a word in a vocabulary. When a word x is given, exactly one path from the root of the tree to the leaf of the word exists. Let $\mathrm{path}(x) =$

$((e_1, b_1), \ldots, (e_m, b_m))$ be the path of x, where e_i is an index of i-th internal node, and $b_i \in \{-1, 1\}$ indicates direction to move at i-th internal node (-1 is left, and 1 is right). Then, the probability of x is given as below:

$$\begin{aligned} P(x) &= \prod_{(e_i, b_i) \in \text{path}(x)} P(b_i | e_i) \\ &= \prod_{(e_i, b_i) \in \text{path}(x)} \sigma(b_i x^\top w_{e_i}), \end{aligned}$$

where $\sigma(\cdot)$ is a sigmoid function, and w is a weight matrix.

This function costs $O(\log(n))$ time as an average length of paths is $O(\log(n))$, and $O(n)$ memory as the number of internal nodes equals $n - 1$.

Parameters

- **in_size** (*int*) – Dimension of input vectors.
- **tree** – A binary tree made with tuples like *((1, 2), 3)*.

Variables **W** (Variable) – Weight parameter matrix.

See: Hierarchical Probabilistic Neural Network Language Model [Morin+, AISTAT2005].

static create_huffman_tree(*word_counts*)
Makes a Huffman tree from a dictionary containing word counts.

This method creates a binary Huffman tree, that is required for *BinaryHierarchicalSoftmax*. For example, `{0: 8, 1: 5, 2: 6, 3: 4}` is converted to `((3, 1), (2, 0))`.

Parameters **word_counts** (*dict of int key and int or float values*) – Dictionary representing counts of words.

Returns Binary Huffman tree with tuples and keys of `word_coutns`.

BlackOut

class chainer.links.**BlackOut**(*in_size*, *counts*, *sample_size*)
BlackOut loss layer.

See also:

black_out() for more detail.

Parameters

- **in_size** (*int*) – Dimension of input vectors.
- **counts** (*int list*) – Number of each identifiers.
- **sample_size** (*int*) – Number of negative samples.

Variables **W** (Variable) – Weight parameter matrix.

CRF1d

class chainer.links.**CRF1d**(*n_label*)
Linear-chain conditional random field loss layer.

This link wraps the *crf1d()* function. It holds a transition cost matrix as a parameter.

Parameters **n_label** (*int*) – Number of labels.

See also:

crf1d() for more detail.

Variables **cost** (Variable) – Transition cost parameter.

argmax(*xs*)

Computes a state that maximizes a joint probability.

Parameters **xs** (*list of Variable*) – Input vector for each label.

Returns

A tuple of *Variable* representing each log-likelihood and a list representing the argmax path.

Return type tuple

See also:

See `crf1d_argmax()` for more detail.

PReLU

class `chainer.links.`**PReLU**(*shape=()*, *init=0.25*)

Parametric ReLU function as a link.

Parameters

- **shape** (*tuple of ints*) – Shape of the parameter array.
- **init** (*float*) – Initial parameter value.

See the paper for details: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

See also:

chainer.functions.prelu()

Variables **W** (Variable) – Coefficient of parametric ReLU.

Maxout

class `chainer.links.`**Maxout**(*in_size*, *out_size*, *pool_size*, *wscale=1*, *initialW=None*, *initial_bias=0*)

Fully-connected maxout layer.

Let `M`, `P` and `N` be an input dimension, a pool size, and an output dimension, respectively. For an input vector x of size `M`, it computes

$$Y_i = \max_j (W_{ij.} x + b_{ij}).$$

Here W is a weight tensor of shape `(M, P, N)`, b an optional bias vector of shape `(M, P)` and $W_{ij.}$ is a sub-vector extracted from W by fixing first and second dimensions to i and j, respectively. Minibatch dimension is omitted in the above equation.

As for the actual implementation, this chain has a Linear link with a `(M * P, N)` weight matrix and an optional `M * P` dimensional bias vector.

Parameters

- **in_size** (*int*) – Dimension of input vectors.
- **out_size** (*int*) – Dimension of output vectors.
- **pool_size** (*int*) – Number of channels.
- **wscale** (*float*) – Scaling factor of the weight matrix.
- **initialW** (*3-D array or None*) – Initial weight value. If None, then this function uses Gaussian distribution scaled by w_scale to initialize weight.
- **initial_bias** (*2-D array, float or None*) – Initial bias value. If it is float, initial bias is filled with this value. If None, bias is omitted.

Variables ***linear*** (Link) – The Linear link that performs affine transformation.

See also:

maxout()

See also:

Goodfellow, I., Warde-farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout Networks. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 1319-1327). URL

NegativeSampling

class chainer.links.**NegativeSampling**(*in_size*, *counts*, *sample_size*, *power=0.75*)

Negative sampling loss layer.

This link wraps the *negative_sampling()* function. It holds the weight matrix as a parameter. It also builds a sampler internally given a list of word counts.

Parameters

- **in_size** (*int*) – Dimension of input vectors.
- **counts** (*int list*) – Number of each identifiers.
- **sample_size** (*int*) – Number of negative samples.
- **power** (*float*) – Power factor α.

See also:

negative_sampling() for more detail.

Variables **W** (Variable) – Weight parameter matrix.

Machine learning models

Classifier

class chainer.links.**Classifier**(*predictor*, *lossfun=<function softmax_cross_entropy>*, *accfun=<function accuracy>*)

A simple classifier model.

This is an example of chain that wraps another chain. It computes the loss and accuracy based on a given input/label pair.

Parameters

- **predictor** (Link) – Predictor network.

- **lossfun** (function) – Loss function.
- **accfun** (function) – Function that computes accuracy.

Variables

- **predictor** (Link) – Predictor network.
- **lossfun** (function) – Loss function.
- **accfun** (function) – Function that computes accuracy.
- **y** (Variable) – Prediction for the last minibatch.
- **loss** (Variable) – Loss value for the last minibatch.
- ***accuracy*** (Variable) – Accuracy for the last minibatch.
- **compute_accuracy** (*bool*) – If True, compute accuracy on the forward computation. The default value is True.

Pre-trained models

Pre-trained models are mainly used to achieve a good performance with a small dataset, or extract a semantic feature vector. Although CaffeFunction automatically loads a pre-trained model released as a caffemodel, the following link models provide an interface for automatically converting caffemodels, and easily extracting semantic feature vectors.

For example, to extract the feature vectors with VGG16Layers, which is a common pre-trained model in the field of image recognition, users need to write the following few lines:

```
from chainer.links import VGG16Layers
from PIL import Image

model = VGG16Layers()
img = Image.open("path/to/image.jpg")
feature = model.extract([img], layers=["fc7"])["fc7"]
```

where fc7 denotes a layer before the last fully-connected layer. Unlike the usual links, these classes automatically load all the parameters from the pre-trained models during initialization.

VGG16Layers

class chainer.links.**VGG16Layers**(*pretrained_model='auto'*)

A pre-trained CNN model with 16 layers provided by VGG team [1].

During initialization, this chain model automatically downloads the pre-trained caffemodel, convert to another chainer model, stores it on your local directory, and initializes all the parameters with it. This model would be useful when you want to extract a semantic feature vector from a given image, or fine-tune the model on a different dataset. Note that this pre-trained model is released under Creative Commons Attribution License.

If you want to manually convert the pre-trained caffemodel to a chainer model that can be specified in the constructor, please use convert_caffemodel_to_npz classmethod instead.

Parameters **pretrained_model** (*str*) – the destination of the pre-trained chainer model serialized as a .npz file. If this argument is specified as auto, it automatically downloads the caffemodel from the internet. Note that in this case the converted chainer model is stored on $CHAINER_DATASET_ROOT/pfnet/chainer/models directory, where $CHAINER_DATASET_ROOT is set as $HOME/.chainer/dataset unless you specify another value as a environment variable. The converted chainer model is automatically used from

the second time. If the argument is specified as `None`, all the parameters are not initialized by the pre-trained model, but the default initializer used in the original paper, i.e., `chainer.initializers.Normal(scale=0.01)`.

Variables **available_layers** (*list of str*) – The list of available layer names used by `__call__` and `extract` methods.

classmethod convert_caffemodel_to_npz (*path_caffemodel*, *path_npz*)
Converts a pre-trained caffemodel to a chainer model.

Parameters

- **path_caffemodel** (*str*) – Path of the pre-trained caffemodel.
- **path_npz** (*str*) – Path of the converted chainer model.

extract (*images, layers=['fc7'], size=(224, 224), test=True, volatile=OFF*)
Extracts all the feature maps of given images.

The difference of directly executing `__call__` is that it directly accepts images as an input and automatically transforms them to a proper variable. That is, it is also interpreted as a shortcut method that implicitly calls `prepare` and `__call__` functions.

Parameters

- **images** (*iterable of PIL.Image or numpy.ndarray*) – Input images.
- **layers** (*list of str*) – The list of layer names you want to extract.
- **size** (*pair of ints*) – The resolution of resized images used as an input of CNN. All the given images are not resized if this argument is `None`, but the resolutions of all the images should be the same.
- **test** (*bool*) – If `True`, dropout runs in test mode.
- **volatile** (`Flag`) – Volatility flag used for input variables.

Returns A directory in which the key contains the layer name and the value contains the corresponding feature map variable.

Return type Dictionary of ~chainer.Variable

predict (*images*, *oversample=True*)
Computes all the probabilities of given images.

Parameters

- **images** (*iterable of PIL.Image or numpy.ndarray*) – Input images.
- **oversample** (*bool*) – If `True`, it averages results across center, corners, and mirrors. Otherwise, it uses only the center.

Returns Output that contains the class probabilities of given images.

Return type *Variable*

`chainer.links.model.vision.vgg.`**prepare** (*image*, *size=(224, 224)*)
Converts the given image to the numpy array for VGG models.

Note that you have to call this method before `__call__` because the pre-trained vgg model requires to resize the given image, covert the RGB to the BGR, subtract the mean, and permute the dimensions before calling.

Parameters

- **image** (*PIL.Image or numpy.ndarray*) – Input image. If an input is `numpy.ndarray`, its shape must be `(height, width)`, `(height, width, channels)`, or `(channels, height, width)`, and the order of the channels must be RGB.
- **size** (*pair of ints*) – Size of converted images. If `None`, the given image is not resized.

Returns The converted output array.

Return type numpy.ndarray

Residual Networks

class `chainer.links.model.vision.resnet.`**ResNetLayers**(*pretrained_model*, *n_layers*)

A pre-trained CNN model provided by MSRA [1].

When you specify the path of the pre-trained chainer model serialized as a `.npz` file in the constructor, this chain model automatically initializes all the parameters with it. This model would be useful when you want to extract a semantic feature vector per image, or fine-tune the model on a different dataset. Note that unlike `VGG16Layers`, it does not automatically download a pre-trained caffemodel. This caffemodel can be downloaded at GitHub.

If you want to manually convert the pre-trained caffemodel to a chainer model that can be specified in the constructor, please use `convert_caffemodel_to_npz` classmethod instead.

Parameters

- **pretrained_model** (*str*) – the destination of the pre-trained chainer model serialized as a `.npz` file. If this argument is specified as `auto`, it automatically loads and converts the caffemodel from `$CHAINER_DATASET_ROOT/pfnet/chainer/models/ResNet-{n-layers}-model.caffemodel`, where `$CHAINER_DATASET_ROOT` is set as `$HOME/.chainer/dataset` unless you specify another value by modifying the environment variable and {n_layers} is replaced with the specified number of layers given as the first argment to this costructor. Note that in this case the converted chainer model is stored on the same directory and automatically used from the next time. If this argument is specified as `None`, all the parameters are not initialized by the pre-trained model, but the default initializer used in the original paper, i.e., `chainer.initializers.HeNormal(scale=1.0)`.
- **n_layers** (*int*) – The number of layers of this model. It should be either 50, 101, or 152.

Variables **available_layers** (*list of str*) – The list of available layer names used by `__call__` and `extract` methods.

classmethod **convert_caffemodel_to_npz**(*path_caffemodel*, *path_npz*, *n_layers=50*)

Converts a pre-trained caffemodel to a chainer model.

Parameters

- **path_caffemodel** (*str*) – Path of the pre-trained caffemodel.
- **path_npz** (*str*) – Path of the converted chainer model.

extract(*images*, *layers=['pool5']*, *size=(224, 224)*, *test=True*, *volatile=OFF*)

Extracts all the feature maps of given images.

The difference of directly executing `__call__` is that it directly accepts images as an input and automatically transforms them to a proper variable. That is, it is also interpreted as a shortcut method that implicitly calls `prepare` and `__call__` functions.

Parameters

- **images** (*iterable of PIL.Image or numpy.ndarray*) – Input images.
- **layers** (*list of str*) – The list of layer names you want to extract.
- **size** (*pair of ints*) – The resolution of resized images used as an input of CNN. All the given images are not resized if this argument is `None`, but the resolutions of all the images should be the same.
- **test** (*bool*) – If `True`, BatchNormalization runs in test mode.
- **volatile** (`Flag`) – Volatility flag used for input variables.

Returns A directory in which the key contains the layer name and the value contains the corresponding feature map variable.

Return type Dictionary of ~chainer.Variable

predict(*images*, *oversample=True*)

Computes all the probabilities of given images.

Parameters

- **images** (*iterable of PIL.Image or numpy.ndarray*) – Input images.
- **oversample** (*bool*) – If `True`, it averages results across center, corners, and mirrors. Otherwise, it uses only the center.

Returns Output that contains the class probabilities of given images.

Return type *Variable*

class chainer.links.**ResNet50Layers**(*pretrained_model='auto'*)

A pre-trained CNN model with 50 layers provided by MSRA [1].

When you specify the path of the pre-trained chainer model serialized as a `.npz` file in the constructor, this chain model automatically initializes all the parameters with it. This model would be useful when you want to extract a semantic feature vector per image, or fine-tune the model on a different dataset. Note that unlike `VGG16Layers`, it does not automatically download a pre-trained caffemodel. This caffemodel can be downloaded at GitHub.

If you want to manually convert the pre-trained caffemodel to a chainer model that can be specified in the constructor, please use `convert_caffemodel_to_npz` classmethod instead.

ResNet50 has 25,557,096 trainable parameters, and it's 58% and 43% fewer than ResNet101 and ResNet152, respectively. On the other hand, the top-5 classification accuracy on ImageNet dataset drops only 0.7% and 1.1% from ResNet101 and ResNet152, respectively. Therefore, ResNet50 may have the best balance between the accuracy and the model size. It would be basically just enough for many cases, but some advanced models for object detection or semantic segmentation use deeper ones as their building blocks, so these deeper ResNets are here for making reproduction work easier.

Parameters **pretrained_model** (*str*) – the destination of the pre-trained chainer model serialized as a `.npz` file. If this argument is specified as `auto`, it automatically loads and converts the caffemodel from `$CHAINER_DATASET_ROOT/pfnet/chainer/models/ResNet-50-model.caffemodel`, where `$CHAINER_DATASET_ROOT` is set as `$HOME/.chainer/dataset` unless you specify another value by modifying the environment variable. Note that in this case the converted chainer model is stored on the same directory and automatically used from the next time. If this argument is specified as `None`, all the parameters are not initialized by the pre-trained model, but the default initializer used in the original paper, i.e., `chainer.initializers.HeNormal(scale=1.0)`.

Variables **available_layers** (*list of str*) – The list of available layer names used by `__call__` and `extract` methods.

class chainer.links.**ResNet101Layers**(*pretrained_model='auto'*)
A pre-trained CNN model with 101 layers provided by MSRA [1].

When you specify the path of the pre-trained chainer model serialized as a `.npz` file in the constructor, this chain model automatically initializes all the parameters with it. This model would be useful when you want to extract a semantic feature vector per image, or fine-tune the model on a different dataset. Note that unlike `VGG16Layers`, it does not automatically download a pre-trained caffemodel. This caffemodel can be downloaded at GitHub.

If you want to manually convert the pre-trained caffemodel to a chainer model that can be specified in the constructor, please use `convert_caffemodel_to_npz` classmethod instead.

ResNet101 has 44,549,224 trainable parameters, and it's 43% fewer than ResNet152 model, while the top-5 classification accuracy on ImageNet dataset drops 1.1% from ResNet152. For many cases, ResNet50 may have the best balance between the accuracy and the model size.

Parameters **pretrained_model** (*str*) – the destination of the pre-trained chainer model serialized as a `.npz` file. If this argument is specified as `auto`, it automatically loads and converts the caffemodel from `$CHAINER_DATASET_ROOT/pfnet/chainer/models/ResNet-101-model.caffemodel`, where `$CHAINER_DATASET_ROOT` is set as `$HOME/.chainer/dataset` unless you specify another value by modifying the environment variable. Note that in this case the converted chainer model is stored on the same directory and automatically used from the next time. If this argument is specified as `None`, all the parameters are not initialized by the pre-trained model, but the default initializer used in the original paper, i.e., `chainer.initializers.HeNormal(scale=1.0)`.

Variables **available_layers** (*list of str*) – The list of available layer names used by `__call__` and `extract` methods.

class chainer.links.**ResNet152Layers**(*pretrained_model='auto'*)
A pre-trained CNN model with 152 layers provided by MSRA [1].

When you specify the path of the pre-trained chainer model serialized as a `.npz` file in the constructor, this chain model automatically initializes all the parameters with it. This model would be useful when you want to extract a semantic feature vector per image, or fine-tune the model on a different dataset. Note that unlike `VGG16Layers`, it does not automatically download a pre-trained caffemodel. This caffemodel can be downloaded at GitHub.

If you want to manually convert the pre-trained caffemodel to a chainer model that can be specified in the constructor, please use `convert_caffemodel_to_npz` classmethod instead.

ResNet152 has 60,192,872 trainable parameters, and it's the deepest ResNet model and it achieves the best result on ImageNet classification task in ILSVRC 2015.

Parameters **pretrained_model** (*str*) – the destination of the pre-trained chainer model serialized as a `.npz` file. If this argument is specified as `auto`, it automatically loads and converts the caffemodel from `$CHAINER_DATASET_ROOT/pfnet/chainer/models/ResNet-152-model.caffemodel`, where `$CHAINER_DATASET_ROOT` is set as `$HOME/.chainer/dataset` unless you specify another value by modifying the environment variable. Note that in this case the converted chainer model is stored on the same directory and automatically used from the next time. If this argument is specified as `None`, all the parameters are not initialized by the pre-trained model, but the default initializer used in the original paper, i.e., `chainer.initializers.HeNormal(scale=1.0)`.

Variables **available_layers** (*list of str*) – The list of available layer names used by `__call__` and `extract` methods.

chainer.links.model.vision.resnet.**prepare**(*image*, *size=(224, 224)*)
Converts the given image to the numpy array for ResNets.

Note that you have to call this method before __call__ because the pre-trained resnet model requires to resize the given image, covert the RGB to the BGR, subtract the mean, and permute the dimensions before calling.

Parameters

- **image** (*PIL.Image or numpy.ndarray*) – Input image. If an input is numpy.ndarray, its shape must be (height, width), (height, width, channels), or (channels, height, width), and the order of the channels must be RGB.
- **size** (*pair of ints*) – Size of converted images. If None, the given image is not resized.

Returns The converted output array.

Return type numpy.ndarray

Deprecated links

Parameter

class chainer.links.**Parameter**(*array*)
Link that just holds a parameter and returns it.

Deprecated since version v1.5: The parameters are stored as variables as of v1.5. Use them directly instead.

Parameters **array** – Initial parameter array.

Variables **W** (Variable) – Parameter variable.

Optimizers

class chainer.optimizers.**AdaDelta**(*rho=0.95*, *eps=1e-06*)
Zeiler's ADADELTA.

See: http://www.matthewzeiler.com/pubs/googleTR2012/googleTR2012.pdf

class chainer.optimizers.**AdaGrad**(*lr=0.001*, *eps=1e-08*)
AdaGrad implementation.

See: http://jmlr.org/papers/v12/duchi11a.html

class chainer.optimizers.**Adam**(*alpha=0.001*, *beta1=0.9*, *beta2=0.999*, *eps=1e-08*)
Adam optimization algorithm.

See: https://arxiv.org/abs/1412.6980v8

class chainer.optimizers.**MomentumSGD**(*lr=0.01*, *momentum=0.9*)
Classical momentum SGD.

class chainer.optimizers.**NesterovAG**(*lr=0.01*, *momentum=0.9*)
Nesterov's Accelerated Gradient.

Formulated as the linear combination coefficients of the velocity and gradient contributions at each iteration.

See: https://arxiv.org/abs/1212.0901

class chainer.optimizers.**RMSprop**(*lr=0.01*, *alpha=0.99*, *eps=1e-08*)
Hinton's RMSprop.

class chainer.optimizers.**RMSpropGraves**(*lr=0.0001*, *alpha=0.95*, *momentum=0.9*, *eps=0.0001*)
Alex Graves's RMSprop.

See https://arxiv.org/abs/1308.0850

class chainer.optimizers.**SGD**(*lr=0.01*)
Vanilla Stochastic Gradient Descent.

class chainer.optimizers.**SMORMS3**(*lr=0.001*, *eps=1e-16*)
Simon Funk's SMORMS3.

See http://sifter.org/~simon/journal/20150420.html.

Serializers

Serialization in NumPy NPZ format

NumPy serializers can be used in arbitrary environments that Chainer runs with. It consists of asymmetric serializer/deserializer due to the fact that numpy.savez() does not support online serialization. Therefore, serialization requires two-step manipulation: first packing the objects into a flat dictionary, and then serializing it into npz format.

class chainer.serializers.**DictionarySerializer**(*target=None*, *path=''*)
Serializer for dictionary.

This is the standard serializer in Chainer. The hierarchy of objects are simply mapped to a flat dictionary with keys representing the paths to objects in the hierarchy.

Note: Despite of its name, this serializer DOES NOT serialize the object into external files. It just build a flat dictionary of arrays that can be fed into numpy.savez() and numpy.savez_compressed(). If you want to use this serializer directly, you have to manually send a resulting dictionary to one of these functions.

Parameters

- **target** (*dict*) – The dictionary that this serializer saves the objects to. If target is None, then a new dictionary is created.
- **path** (*str*) – The base path in the hierarchy that this serializer indicates.

Variables **target** (*dict*) – The target dictionary. Once the serialization completes, this dictionary can be fed into numpy.savez() or numpy.savez_compressed() to serialize it in the NPZ format.

class chainer.serializers.**NpzDeserializer**(*npz*, *path=''*, *strict=True*)
Deserializer for NPZ format.

This is the standard deserializer in Chainer. This deserializer can be used to read an object serialized by *save_npz()*.

Parameters

- **npz** – *npz* file object.
- **path** – The base path that the deserialization starts from.
- **strict** (*bool*) – If True, the deserializer raises an error when an expected value is not found in the given NPZ file. Otherwise, it ignores the value and skip deserialization.

chainer.serializers.**save_npz**(*filename*, *obj*, *compression=True*)
Saves an object to the file in NPZ format.

This is a short-cut function to save only one object into an NPZ file.

Parameters

- **filename** (*str*) – Target file name.
- **obj** – Object to be serialized. It must support serialization protocol.
- **compression** (*bool*) – If True, compression in the resulting zip file is enabled.

chainer.serializers.**load_npz**(*filename*, *obj*)
Loads an object from the file in NPZ format.

This is a short-cut function to load from an *.npz* file that contains only one object.

Parameters

- **filename** (*str*) – Name of the file to be loaded.
- **obj** – Object to be deserialized. It must support serialization protocol.

Serialization in HDF5 format

class chainer.serializers.**HDF5Serializer**(*group*, *compression=4*)
Serializer for HDF5 format.

This is the standard serializer in Chainer. The chain hierarchy is simply mapped to HDF5 hierarchical groups.

Parameters

- **group** (*h5py.Group*) – The group that this serializer represents.
- **compression** (*int*) – Gzip compression level.

class chainer.serializers.**HDF5Deserializer**(*group*, *strict=True*)
Deserializer for HDF5 format.

This is the standard deserializer in Chainer. This deserializer can be used to read an object serialized by *HDF5Serializer*.

Parameters

- **group** (*h5py.Group*) – The group that the deserialization starts from.
- **strict** (*bool*) – If True, the deserializer raises an error when an expected value is not found in the given HDF5 file. Otherwise, it ignores the value and skip deserialization.

chainer.serializers.**save_hdf5**(*filename*, *obj*, *compression=4*)
Saves an object to the file in HDF5 format.

This is a short-cut function to save only one object into an HDF5 file. If you want to save multiple objects to one HDF5 file, use *HDF5Serializer* directly by passing appropriate h5py.Group objects.

Parameters

- **filename** (*str*) – Target file name.
- **obj** – Object to be serialized. It must support serialization protocol.
- **compression** (*int*) – Gzip compression level.

chainer.serializers.**load_hdf5**(*filename*, *obj*)
Loads an object from the file in HDF5 format.

This is a short-cut function to load from an HDF5 file that contains only one object. If you want to load multiple objects from one HDF5 file, use *HDF5Deserializer* directly by passing appropriate h5py.Group objects.

Parameters

- **filename** (*str*) – Name of the file to be loaded.
- **obj** – Object to be deserialized. It must support serialization protocol.

Function hooks

Chainer provides a function-hook mechanism that enriches the behavior of forward and backward propagation of *Function*.

Base class

class chainer.function.**FunctionHook**
Base class of hooks for Functions.

FunctionHook is an callback object that is registered to *Function*. Registered function hooks are invoked before and after forward and backward operations of each function.

Function hooks that derive *FunctionHook* are required to implement four methods: *forward_preprocess()*, *forward_postprocess()*, *backward_preprocess()*, and *backward_postprocess()*. By default, these methods do nothing.

Specifically, when __call__() method of some function is invoked, *forward_preprocess()* (resp. *forward_postprocess()*) of all function hooks registered to this function are called before (resp. after) forward propagation.

Likewise, when *backward()* of some *Variable* is invoked, *backward_preprocess()* (resp. *backward_postprocess()*) of all function hooks registered to the function which holds this variable as a gradient are called before (resp. after) backward propagation.

There are two ways to register *FunctionHook* objects to *Function* objects.

First one is to use with statement. Function hooks hooked in this way are registered to all functions within with statement and are unregistered at the end of with statement.

Example

The following code is a simple example in which we measure the elapsed time of a part of forward propagation procedure with *TimerHook*, which is a subclass of *FunctionHook*.

```
>>> from chainer import function_hooks
>>> class Model(chainer.Chain):
...     def __call__(self, x1):
...         return F.exp(self.l(x1))
>>> model1 = Model(l=L.Linear(10, 10))
>>> model2 = Model(l=L.Linear(10, 10))
>>> x = chainer.Variable(np.zeros((1, 10), 'f'))
>>> with chainer.function_hooks.TimerHook() as m:
...     _ = model1(x)
...     y = model2(x)
```

```
...     print("Total time : " + str(m.total_time()))
...     model3 = Model(l=L.Linear(10, 10))
...     z = model3(y)
Total time : ...
```

In this example, we measure the elapsed times for each forward propagation of all functions in `model1` and `model2` (specifically, `LinearFunction` and `Exp` of `model1` and `model2`). Note that `model3` is not a target of measurement as *`TimerHook`* is unregistered before forward propagation of `model3`.

Note: Chainer stores the dictionary of registered function hooks as a thread local object. So, function hooks registered are different depending on threads.

The other one is to register directly to *`Function`* object with *`add_hook()`* method. Function hooks registered in this way can be removed by *`delete_hook()`* method. Contrary to former registration method, function hooks are registered only to the function which *`add_hook()`* is called.

Parameters **name** (*`str`*) – Name of this function hook.

backward_postprocess (*function*, *in_data*, *out_grad*)
Callback function invoked after backward propagation.

Parameters

- **function** (`Function`) – Function object to which the function hook is registered.
- **in_data** (*`tuple of numpy.ndarray or tuple of cupy.ndarray`*) – Input of forward propagation.
- **out_grad** (*`tuple of numpy.ndarray or tuple of cupy.ndarray`*) – Gradient data of backward propagation.

backward_preprocess (*function*, *in_data*, *out_grad*)
Callback function invoked before backward propagation.

Parameters

- **function** (`Function`) – Function object to which the function hook is registered.
- **in_data** (*`tuple of numpy.ndarray or tuple of cupy.ndarray`*) – Input data of forward propagation.
- **out_grad** (*`tuple of numpy.ndarray or tuple of cupy.ndarray`*) – Gradient data of backward propagation.

forward_postprocess (*function*, *in_data*)
Callback function invoked after forward propagation.

Parameters

- **function** (`Function`) – Function object to which the function hook is registered.
- **in_data** (*`tuple of numpy.ndarray or tuple of cupy.ndarray`*) – Input data of forward propagation.

forward_preprocess (*function*, *in_data*)
Callback function invoked before forward propagation.

Parameters

- **function** (`Function`) – Function object to which the function hook is registered.

- **in_data** (*tuple of numpy.ndarray or tuple of cupy.ndarray*) – Input data of forward propagation.

Concrete function hooks

class chainer.function_hooks.**PrintHook**(*sep=''*, *end='n'*, *file=<open file '<stdout>'*, *mode 'w'>*, *flush=True*)

Function hook that prints debug information.

This function hook outputs the debug information of input arguments of `forward` and `backward` methods involved in the hooked functions at preprocessing time (that is, just before each method is called).

Unlike simple "debug print" technique, where users insert print functions at every function to be inspected, we can show the information of all functions involved with single `with` statement.

Further, this hook enables us to show the information of `backward` methods without inserting print functions into Chainer's library code.

Variables

- **sep** – Separator of print function.
- **end** – Character to be added at the end of print function.
- **file** – Output file_like object that that redirect to.
- **flush** – If `True`, this hook forcibly flushes the text stream at the end of preprocessing.

Example

The basic usage is to use it with `with` statement.

```
>>> from chainer import function_hooks
>>> l = L.Linear(10, 10)
>>> x = chainer.Variable(np.zeros((1, 10), 'f'))
>>> with chainer.function_hooks.PrintHook():
...     y = l(x)
...     z = F.sum(y)
...     z.backward()
```

In this example, `PrintHook` shows the debug information of forward propagation of `LinearFunction` (which is implicitly called by `l`) and `Sum` (called by `F.sum`) and backward propagation of `z` and `y`.

---

class chainer.function_hooks.**TimerHook**

Function hook for measuring elapsed time of functions.

Variables **call_history** – List of measurement results. It consists of pairs of the function that calls this hook and the elapsed time the function consumes.

total_time()

Returns total elapsed time in seconds.

Weight Initializers

Weight initializer is an instance of `Initializer` that destructively edits the contents of `numpy.ndarray` or *`cupy.ndarray`*. Typically, weight initializers are passed to `__init__` of *`Link`* and initializes its the weights and biases.

Base class

class chainer.initializer.**Initializer**(*dtype=None*)
Initializes array.

It initializes the given array.

Variables **dtype** – Data type specifier. It is for type check in __call__ function.

Concrete initializers

class chainer.initializers.**Identity**(*scale=1.0*, *dtype=None*)
Initializes array with the identity matrix.

It initializes the given array with the constant multiple of the identity matrix. Note that arrays to be passed must be 2D squared matrices.

Variables ***scale*** (*scalar*) – A constant to be multiplied to identity matrices.

class chainer.initializers.**Constant**(*fill_value*, *dtype=None*)
Initializes array with constant value.

Variables

- **fill_value** (*scalar or numpy.ndarray or cupy.ndarray*) – A constant to be assigned to the initialized array. Broadcast is allowed on this assignment.
- **dtype** – Data type specifier.

chainer.initializers.**Zero**(*dtype=None*)
Returns initializer that initializes array with the all-zero array.

Parameters **dtype** – Data type specifier.

Returns An initialized array.

Return type numpy.ndarray or *cupy.ndarray*

chainer.initializers.**One**(*dtype=None*)
Returns initializer that initializes array with the all-one array.

Parameters **dtype** – Data type specifier.

Returns An initialized array.

Return type numpy.ndarray or *cupy.ndarray*

class chainer.initializers.**Normal**(*scale=0.05*, *dtype=None*)
Initializes array with a normal distribution.

Each element of the array is initialized by the value drawn independently from Gaussian distribution whose mean is 0, and standard deviation is scale.

Parameters

- **scale** (*float*) – Standard deviation of Gaussian distribution.
- **dtype** – Data type specifier.

class chainer.initializers.**GlorotNormal**(*scale=1.0*, *dtype=None*)
Initializes array with scaled Gaussian distribution.

Each element of the array is initialized by the value drawn independently from Gaussian distribution whose mean is 0, and standard deviation is $scale \times \sqrt{\frac{2}{fan_{in}+fan_{out}}}$, where fan_{in} and fan_{out} are the number of input and output units, respectively.

Reference: Glorot & Bengio, AISTATS 2010

Parameters

- **scale** (*float*) – A constant that determines the scale of the standard deviation.
- **dtype** – Data type specifier.

class chainer.initializers.**HeNormal**(*scale=1.0*, *dtype=None*)
Initializes array with scaled Gaussian distribution.

Each element of the array is initialized by the value drawn independently from Gaussian distribution whose mean is 0, and standard deviation is $scale \times \sqrt{\frac{2}{fan_{in}}}$, where fan_{in} is the number of input units.

Reference: He et al., https://arxiv.org/abs/1502.01852

Parameters

- **scale** (*float*) – A constant that determines the scale of the standard deviation.
- **dtype** – Data type specifier.

class chainer.initializers.**Orthogonal**(*scale=1.1*, *dtype=None*)
Initializes array with an orthogonal system.

This initializer first makes a matrix of the same shape as the array to be initialized whose elements are drawn independently from standard Gaussian distribution. Next, it applies Singular Value Decomposition (SVD) to the matrix. Then, it initializes the array with either side of resultant orthogonal matrices, depending on the shape of the input array. Finally, the array is multiplied by the constant `scale`.

If the `ndim` of the input array is more than 2, we consider the array to be a matrix by concatenating all axes except the first one.

The number of vectors consisting of the orthogonal system (i.e. first element of the shape of the array) must be equal to or smaller than the dimension of each vector (i.e. second element of the shape of the array).

Variables

- ***scale*** (*float*) – A constant to be multiplied by.
- **dtype** – Data type specifier.

Reference: Saxe et al., https://arxiv.org/abs/1312.6120

class chainer.initializers.**Uniform**(*scale=0.05*, *dtype=None*)
Initializes array with a scaled uniform distribution.

Each element of the array is initialized by the value drawn independently from uniform distribution $[-scale, scale]$.

Variables

- ***scale*** (*float*) – A constant that determines the scale of the uniform distribution.
- **dtype** – Data type specifier.

class chainer.initializers.**LeCunUniform**(*scale=1.0*, *dtype=None*)
Initializes array with a scaled uniform distribution.

Each element of the array is initialized by the value drawn independently from uniform distribution $[-s, s]$ where $s = scale \times \sqrt{\frac{3}{fan_{in}}}$. Here fan_{in} is the number of input units.

Reference: LeCun 98, Efficient Backprop http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf

Variables

- ***scale*** (*float*) – A constant that determines the scale of the uniform distribution.
- **dtype** – Data type specifier.

class chainer.initializers.**GlorotUniform**(*scale=1.0*, *dtype=None*)
Initializes array with a scaled uniform distribution.

Each element of the array is initialized by the value drawn independently from uniform distribution $[-s, s]$ where $s = scale \times \sqrt{\frac{6}{fan_{in}+fan_{out}}}$. Here, fan_{in} and *fan_{out}* are the number of input and output units, respectively.

Variables

- ***scale*** (*float*) – A constant that determines the scale of the uniform distribution.
- **dtype** – Data type specifier.

class chainer.initializers.**HeUniform**(*scale=1.0*, *dtype=None*)
Initializes array with scaled uniform distribution.

Each element of the array is initialized by the value drawn independently from uniform distribution $[-s, s]$ where $s = scale \times \sqrt{\frac{6}{fan_{in}}}$. Here, fan_{in} is the number of input units.

Variables

- ***scale*** (*float*) – A constant that determines the scale of the uniform distribution.
- **dtype** – Data type specifier.

Helper function

chainer.**init_weight**(*weights*, *initializer*, *scale=1.0*)
Helper function for initialization of the weight tensor.

This function accepts several types of initializer, prepares the appropriate ~chainer.Initializer if necessary, and does the initialization.

Parameters

- **weights** (*numpy.ndarray or cupy.ndarray*) – Weight tensor to be initialized.
- **initializer** – The value used to initialize the data. May be None (in which case *HeNormal* is used as an initializer), a scalar to set all values to, an numpy.ndarray to be assigned, or a callable that takes numpy.ndarray or *cupy.ndarray* and edits its value.
- **scale** (*scalar*) – A constant to multiply initializer by.

Dataset examples

The most basic *dataset* implementation is an array. Both NumPy and CuPy arrays can be used directly as datasets.

In many cases, though, the simple arrays are not enough to write the training procedure. In order to cover most of such cases, Chainer provides many built-in implementations of datasets.

These built-in datasets are divided into two groups. One is a group of general datasets. Most of them are wrapper of other datasets to introduce some structures (e.g., tuple or dict) to each data point. The other one is a group of concrete, popular datasets. These concrete examples use the downloading utilities in the *chainer.dataset* module to cache downloaded and converted datasets.

General datasets

General datasets are further divided into three types.

The first one is *DictDataset* and *TupleDataset*, both of which combine other datasets and introduce some structures on them.

The second one is *SubDataset*, which represents a subset of an existing dataset. It can be used to separate a dataset for hold-out validation or cross validation. Convenient functions to make random splits are also provided.

The last one is a group of domain-specific datasets. Currently, *ImageDataset* and *LabeledImageDataset* are provided for datasets of images.

DictDataset

class chainer.datasets.**DictDataset**(***datasets*)
Dataset of a dictionary of datasets.

It combines multiple datasets into one dataset. Each example is represented by a dictionary mapping a key to an example of the corresponding dataset.

Parameters **datasets** – Underlying datasets. The keys are used as the keys of each example. All datasets must have the same length.

TupleDataset

class chainer.datasets.**TupleDataset**(**datasets*)
Dataset of a tuple of datasets.

It combines multiple datasets into one dataset. Each example is represented by a tuple whose `i`-th item corresponds to the i-th dataset.

Parameters **datasets** – Underlying datasets. The `i`-th one is used for the `i`-th item of each example. All datasets must have the same length.

SubDataset

class chainer.datasets.**SubDataset**(*dataset*, *start*, *finish*, *order=None*)
Subset of a base dataset.

SubDataset defines a subset of a given base dataset. The subset is defined as an interval of indexes, optionally with a given permutation.

If `order` is given, then the `i`-th example of this dataset is the `order[start + i]`-th example of the base dataset, where `i` is a non-negative integer. If `order` is not given, then the `i`-th example of this dataset is the `start + i`-th example of the base dataset. Negative indexing is also allowed: in this case, the term `start + i` is replaced by `finish + i`.

SubDataset is often used to split a dataset into training and validation subsets. The training set is used for training, while the validation set is used to track the generalization performance, i.e. how the learned model works well on unseen data. We can tune hyperparameters (e.g. number of hidden units, weight initializers,

learning rate, etc.) by comparing the validation performance. Note that we often use another set called test set to measure the quality of the tuned hyperparameter, which can be made by nesting multiple SubDatasets.

There are two ways to make training-validation splits. One is a single split, where the dataset is split just into two subsets. It can be done by *split_dataset()* or *split_dataset_random()*. The other one is a k-fold cross validation, in which the dataset is divided into k subsets, and k different splits are generated using each of the k subsets as a validation set and the rest as a training set. It can be done by *get_cross_validation_datasets()*.

Parameters

- **dataset** – Base dataset.
- **start** (*int*) – The first index in the interval.
- **finish** (*int*) – The next-to-the-last index in the interval.
- **order** (*sequence of ints*) – Permutation of indexes in the base dataset. If this is `None`, then the ascending order of indexes is used.

chainer.datasets.**split_dataset**(*dataset*, *split_at*, *order=None*)

Splits a dataset into two subsets.

This function creates two instances of *SubDataset*. These instances do not share any examples, and they together cover all examples of the original dataset.

Parameters

- **dataset** – Dataset to split.
- **split_at** (*int*) – Position at which the base dataset is split.
- **order** (*sequence of ints*) – Permutation of indexes in the base dataset. See the document of *SubDataset* for details.

Returns

Two *SubDataset* objects. The first subset represents the examples of indexes `order[:split_at]` while the second subset represents the examples of indexes `order[split_at:]`.

Return type tuple

chainer.datasets.**split_dataset_random**(*dataset*, *first_size*, *seed=None*)

Splits a dataset into two subsets randomly.

This function creates two instances of *SubDataset*. These instances do not share any examples, and they together cover all examples of the original dataset. The split is automatically done randomly.

Parameters

- **dataset** – Dataset to split.
- **first_size** (*int*) – Size of the first subset.
- **seed** (*int*) – Seed the generator used for the permutation of indexes. If an integer beging convertible to 32 bit unsigned integers is specified, it is guaranteed that each sample in the given dataset always belongs to a specific subset. If `None`, the permutation is changed randomly.

Returns

Two *SubDataset* objects. The first subset contains `first_size` examples randomly chosen from the dataset without replacement, and the second subset contains the rest of the dataset.

Return type tuple

chainer.datasets.**get_cross_validation_datasets**(*dataset*, *n_fold*, *order=None*)
Creates a set of training/test splits for cross validation.

This function generates `n_fold` splits of the given dataset. The first part of each split corresponds to the training dataset, while the second part to the test dataset. No pairs of test datasets share any examples, and all test datasets together cover the whole base dataset. Each test dataset contains almost same number of examples (the numbers may differ up to 1).

Parameters

- **dataset** – Dataset to split.
- **n_fold** (*int*) – Number of splits for cross validation.
- **order** (*sequence of ints*) – Order of indexes with which each split is determined. If it is `None`, then no permutation is used.

Returns List of dataset splits.

Return type list of tuples

chainer.datasets.**get_cross_validation_datasets_random**(*dataset*, *n_fold*, *seed=None*)
Creates a set of training/test splits for cross validation randomly.

This function acts almost same as `get_cross_validation_dataset()`, except automatically generating random permutation.

Parameters

- **dataset** – Dataset to split.
- **n_fold** (*int*) – Number of splits for cross validation.
- **seed** (*int*) – Seed the generator used for the permutation of indexes. If an integer beging convertible to 32 bit unsigned integers is specified, it is guaranteed that each sample in the given dataset always belongs to a specific subset. If `None`, the permutation is changed randomly.

Returns List of dataset splits.

Return type list of tuples

ImageDataset

class chainer.datasets.**ImageDataset**(*paths*, *root='.'*, *dtype=<type 'numpy.float32'>*)
Dataset of images built from a list of paths to image files.

This dataset reads an external image file on every call of the `__getitem__()` operator. The paths to the image to retrieve is given as either a list of strings or a text file that contains paths in distinct lines.

Each image is automatically converted to arrays of shape `channels, height, width`, where `channels` represents the number of channels in each pixel (e.g., 1 for grey-scale images, and 3 for RGB-color images).

Note: This dataset requires the Pillow package being installed. In order to use this dataset, install Pillow (e.g. by using the command `pip install Pillow`). Be careful to prepare appropriate libraries for image formats you want to use (e.g. libpng for PNG images, and libjpeg for JPG images).

Parameters

- **paths** (*str or list of strs*) – If it is a string, it is a path to a text file that contains paths to images in distinct lines. If it is a list of paths, the `i`-th element represents the path to the `i`-th image. In both cases, each path is a relative one from the root path given by another argument.
- **root** (*str*) – Root directory to retrieve images from.
- **dtype** – Data type of resulting image arrays.

LabeledImageDataset

class `chainer.datasets.`**LabeledImageDataset**(*pairs*, *root='.'*, *dtype=<type 'numpy.float32'>*, *label_dtype=<type 'numpy.int32'>*)

Dataset of image and label pairs built from a list of paths and labels.

This dataset reads an external image file like *ImageDataset*. The difference from *ImageDataset* is that this dataset also returns a label integer. The paths and labels are given as either a list of pairs or a text file contains paths/labels pairs in distinct lines. In the latter case, each path and corresponding label are separated by white spaces. This format is same as one used in Caffe.

Note: This dataset requires the Pillow package being installed. In order to use this dataset, install Pillow (e.g. by using the command `pip install Pillow`). Be careful to prepare appropriate libraries for image formats you want to use (e.g. libpng for PNG images, and libjpeg for JPG images).

Parameters

- **pairs** (*str or list of tuples*) – If it is a string, it is a path to a text file that contains paths to images in distinct lines. If it is a list of pairs, the `i`-th element represents a pair of the path to the `i`-th image and the corresponding label. In both cases, each path is a relative one from the root path given by another argument.
- **root** (*str*) – Root directory to retrieve images from.
- **dtype** – Data type of resulting image arrays.
- **label_dtype** – Data type of the labels.

Concrete datasets

MNIST

`chainer.datasets.`**get_mnist**(*withlabel=True*, *ndim=1*, *scale=1.0*, *dtype=<type 'numpy.float32'>*, *label_dtype=<type 'numpy.int32'>*)

Gets the MNIST dataset.

MNIST is a set of hand-written digits represented by grey-scale 28x28 images. In the original images, each pixel is represented by one-byte unsigned integer. This function scales the pixels to floating point values in the interval `[0, scale]`.

This function returns the training set and the test set of the official MNIST dataset. If `withlabel` is `True`, each dataset consists of tuples of images and labels, otherwise it only consists of images.

Parameters

- **withlabel** (*bool*) – If `True`, it returns datasets with labels. In this case, each example is a tuple of an image and a label. Otherwise, the datasets only contain images.

- **ndim** (*int*) – Number of dimensions of each image. The shape of each image is determined depending on `ndim` as follows:
  - `ndim == 1`: the shape is `(784,)`
  - `ndim == 2`: the shape is `(28, 28)`
  - `ndim == 3`: the shape is `(1, 28, 28)`
- **scale** (*float*) – Pixel value scale. If it is 1 (default), pixels are scaled to the interval `[0, 1]`.
- **dtype** – Data type of resulting image arrays.
- **label_dtype** – Data type of the labels.

Returns A tuple of two datasets. If `withlabel` is `True`, both datasets are *TupleDataset* instances. Otherwise, both datasets are arrays of images.

CIFAR10/100

`chainer.datasets.`**`get_cifar10`**(*withlabel=True*, *ndim=3*, *scale=1.0*)

Gets the CIFAR-10 dataset.

CIFAR-10 is a set of small natural images. Each example is an RGB color image of size 32x32, classified into 10 groups. In the original images, each component of pixels is represented by one-byte unsigned integer. This function scales the components to floating point values in the interval `[0, scale]`.

This function returns the training set and the test set of the official CIFAR-10 dataset. If `withlabel` is `True`, each dataset consists of tuples of images and labels, otherwise it only consists of images.

Parameters

- **withlabel** (*bool*) – If `True`, it returns datasets with labels. In this case, each example is a tuple of an image and a label. Otherwise, the datasets only contain images.
- **ndim** (*int*) – Number of dimensions of each image. The shape of each image is determined depending on ndim as follows:
  - `ndim == 1`: the shape is `(3072,)`
  - `ndim == 3`: the shape is `(3, 32, 32)`
- **scale** (*float*) – Pixel value scale. If it is 1 (default), pixels are scaled to the interval `[0, 1]`.

Returns A tuple of two datasets. If `withlabel` is `True`, both datasets are *TupleDataset* instances. Otherwise, both datasets are arrays of images.

`chainer.datasets.`**`get_cifar100`**(*withlabel=True*, *ndim=3*, *scale=1.0*)

Gets the CIFAR-100 dataset.

CIFAR-100 is a set of small natural images. Each example is an RGB color image of size 32x32, classified into 100 groups. In the original images, each component pixels is represented by one-byte unsigned integer. This function scales the components to floating point values in the interval `[0, scale]`.

This function returns the training set and the test set of the official CIFAR-100 dataset. If `withlabel` is `True`, each dataset consists of tuples of images and labels, otherwise it only consists of images.

Parameters

- **withlabel** (*bool*) – If `True`, it returns datasets with labels. In this case, each example is a tuple of an image and a label. Otherwise, the datasets only contain images.

- **ndim** (*int*) – Number of dimensions of each image. The shape of each image is determined depending on ndim as follows:
  - `ndim == 1`: the shape is `(3072,)`
  - `ndim == 3`: the shape is `(3, 32, 32)`
- **scale** (*float*) – Pixel value scale. If it is 1 (default), pixels are scaled to the interval `[0, 1]`.

Returns A tuple of two datasets. If `withlabel` is `True`, both are *TupleDataset* instances. Otherwise, both datasets are arrays of images.

Penn Tree Bank

`chainer.datasets.`**`get_ptb_words`**`()`

Gets the Penn Tree Bank dataset as long word sequences.

Penn Tree Bank is originally a corpus of English sentences with linguistic structure annotations. This function uses a variant distributed at https://github.com/tomsercu/lstm, which omits the annotation and splits the dataset into three parts: training, validation, and test.

This function returns the training, validation, and test sets, each of which is represented as a long array of word IDs. All sentences in the dataset are concatenated by End-of-Sentence mark '<eos>', which is treated as one of the vocabulary.

Returns Int32 vectors of word IDs.

Return type tuple of numpy.ndarray

See also:

Use *get_ptb_words_vocabulary()* to get the mapping between the words and word IDs.

`chainer.datasets.`**`get_ptb_words_vocabulary`**`()`

Gets the Penn Tree Bank word vocabulary.

Returns

Dictionary that maps words to corresponding word IDs. The IDs are used in the Penn Tree Bank long sequence datasets.

Return type dict

See also:

See *get_ptb_words()* for the actual datasets.

Iterator examples

Chainer provides some iterators that implement typical strategies to create mini-batches by iterating over datasets. *SerialIterator* is the simplest one, which extract mini batches in the main thread. *MultiprocessIterator* is a parallelized version of *SerialIterator*. It maintains worker subprocesses to load the next mini-batch in parallel.

SerialIterator

class chainer.iterators.**SerialIterator**(*dataset*, *batch_size*, *repeat=True*, *shuffle=True*)

Dataset iterator that serially reads the examples.

This is a simple implementation of *Iterator* that just visits each example in either the order of indexes or a shuffled order.

To avoid unintentional performance degradation, the shuffle option is set to True by default. For validation, it is better to set it to False when the underlying dataset supports fast slicing. If the order of examples has an important meaning and the updater depends on the original order, this option should be set to False.

This iterator saves -1 instead of None in snapshots since some serializers do not support None.

Parameters

- **dataset** – Dataset to iterate.
- **batch_size** (*int*) – Number of examples within each batch.
- **repeat** (*bool*) – If True, it infinitely loops over the dataset. Otherwise, it stops iteration at the end of the first epoch.
- **shuffle** (*bool*) – If True, the order of examples is shuffled at the beginning of each epoch. Otherwise, examples are extracted in the order of indexes.

MultiprocessIterator

class chainer.iterators.**MultiprocessIterator**(*dataset*, *batch_size*, *repeat=True*, *shuffle=True*, *n_processes=None*, *n_prefetch=1*, *shared_mem=None*)

Dataset iterator that loads examples in parallel.

This is an implementation of *Iterator* that loads examples with worker processes. It uses the standard multiprocessing module to parallelize the loading. The dataset is sent to the worker processes in the standard way using pickle.

Note that this iterator effectively prefetches the examples for the next batch asynchronously after the current batch is returned.

This iterator saves -1 instead of None in snapshots since some serializers do not support None.

Parameters

- **dataset** (*Dataset*) – Dataset to iterate.
- **batch_size** (*int*) – Number of examples within each batch.
- **repeat** (*bool*) – If True, it infinitely loops over the dataset. Otherwise, it stops iteration at the end of the first epoch.
- **shuffle** (*bool*) – If True, the order of examples is shuffled at the beginning of each epoch. Otherwise, examples are extracted in the order of indexes.
- **n_processes** (*int*) – Number of worker processes. The number of CPUs is used by default.
- **n_prefetch** (*int*) – Number of prefetch batches.
- **shared_mem** (*int*) – The size of using shared memory per data. If None, size is adjusted automatically.

Trainer extensions

dump_graph

chainer.training.extensions.**dump_graph**(*root_name*, *out_name='cg.dot'*, *variable_style=None*, *function_style=None*)

Returns a trainer extension to dump a computational graph.

This extension dumps a computational graph. The graph is output in DOT language.

It only dumps a graph at the first iteration by default.

Parameters

- **root_name** (*str*) – Name of the root of the computational graph. The root variable is retrieved by this name from the observation dictionary of the trainer.
- **out_name** (*str*) – Output file name.
- **variable_style** (*dict*) – Dot node style for variables. Each variable is rendered by an octagon by default.
- **function_style** (*dict*) – Dot node style for functions. Each function is rendered by a rectangular by default.

See also:

See *build_computational_graph()* for the `variable_style` and `function_style` arguments.

Evaluator

class chainer.training.extensions.**Evaluator**(*iterator*, *target*, *converter=<function concat_examples>*, *device=None*, *eval_hook=None*, *eval_func=None*)

Trainer extension to evaluate models on a validation set.

This extension evaluates the current models by a given evaluation function. It creates a *Reporter* object to store values observed in the evaluation function on each iteration. The report for all iterations are aggregated to *DictSummary*. The collected mean values are further reported to the reporter object of the trainer, where the name of each observation is prefixed by the evaluator name. See *Reporter* for details in naming rules of the reports.

Evaluator has a structure to customize similar to that of *StandardUpdater*. The main differences are:

- There are no optimizers in an evaluator. Instead, it holds links to evaluate.
- An evaluation loop function is used instead of an update function.
- Preparation routine can be customized, which is called before each evaluation. It can be used, e.g., to initialize the state of stateful recurrent networks.

There are two ways to modify the evaluation behavior besides setting a custom evaluation function. One is by setting a custom evaluation loop via the `eval_func` argument. The other is by inheriting this class and overriding the *evaluate()* method. In latter case, users have to create and handle a reporter object manually. Users also have to copy the iterators before using them, in order to reuse them at the next time of evaluation.

This extension is called at the end of each epoch by default.

Parameters

- **iterator** – Dataset iterator for the validation dataset. It can also be a dictionary of iterators. If this is just an iterator, the iterator is registered by the name `'main'`.
- **target** – Link object or a dictionary of links to evaluate. If this is just a link object, the link is registered by the name `'main'`.
- **converter** – Converter function to build input arrays. *concat_examples()* is used by default.
- **device** – Device to which the training data is sent. Negative value indicates the host memory (CPU).
- **eval_hook** – Function to prepare for each evaluation process. It is called at the beginning of the evaluation. The evaluator extension object is passed at each call.
- **eval_func** – Evaluation function called at each iteration. The target link to evaluate as a callable is used by default.

Variables

- **converter** – Converter function.
- ***device*** – Device to which the training data is sent.
- **eval_hook** – Function to prepare for each evaluation process.
- **eval_func** – Evaluation function called at each iteration.

evaluate()

Evaluates the model and returns a result dictionary.

This method runs the evaluation loop over the validation dataset. It accumulates the reported values to *DictSummary* and returns a dictionary whose values are means computed by the summary.

Users can override this method to customize the evaluation routine.

Returns

Result dictionary. This dictionary is further reported via *report()* without specifying any observer.

Return type dict

get_all_iterators()

Returns a dictionary of all iterators.

get_all_targets()

Returns a dictionary of all target links.

get_iterator(*name*)

Returns the iterator of the given name.

get_target(*name*)

Returns the target link of the given name.

ExponentialShift

class chainer.training.extensions.**ExponentialShift**(*attr*, *rate*, *init=None*, *target=None*, *optimizer=None*)

Trainer extension to exponentially shift an optimizer attribute.

This extension exponentially increases or decreases the specified attribute of the optimizer. The typical use case is an exponential decay of the learning rate.

This extension is also called before the training loop starts by default.

Parameters

- **attr** (*str*) – Name of the attribute to shift.
- **rate** (*float*) – Rate of the exponential shift. This value is multiplied to the attribute at each call.
- **init** (*float*) – Initial value of the attribute. If it is `None`, the extension extracts the attribute at the first call and uses it as the initial value.
- **target** (*float*) – Target value of the attribute. If the attribute reaches this value, the shift stops.
- **optimizer** (`Optimizer`) – Target optimizer to adjust the attribute. If it is `None`, the main optimizer of the updater is used.

LinearShift

class `chainer.training.extensions.`**LinearShift**(*attr*, *value_range*, *time_range*, *optimizer=None*)

Trainer extension to change an optimizer attribute linearly.

This extension changes an optimizer attribute from the first value to the last value linearly within a specified duration. The typical use case is warming up of the momentum coefficient.

For example, suppose that this extension is called at every iteration, and `value_range == (x, y)` and `time_range == (i, j)`. Then, this extension keeps the attribute to be `x` up to the `i`-th iteration, linearly shifts the value to `y` by the `j`-th iteration, and then keeps the value to be `y` after the `j`-th iteration.

This extension is also called before the training loop starts by default.

Parameters

- **attr** (*str*) – Name of the optimizer attribute to adjust.
- **value_range** (*tuple of float*) – The first and the last values of the attribute.
- **time_range** (*tuple of ints*) – The first and last counts of calls in which the attribute is adjusted.
- **optimizer** (`Optimizer`) – Target optimizer object. If it is None, the main optimizer of the trainer is used.

LogReport

class `chainer.training.extensions.`**LogReport**(*keys=None*, *trigger=(1*, *'epoch')*, *postprocess=None*, *log_name='log'*)

Trainer extension to output the accumulated results to a log file.

This extension accumulates the observations of the trainer to *DictSummary* at a regular interval specified by a supplied trigger, and writes them into a log file in JSON format.

There are two triggers to handle this extension. One is the trigger to invoke this extension, which is used to handle the timing of accumulating the results. It is set to `1, 'iteration'` by default. The other is the trigger to determine when to emit the result. When this trigger returns True, this extension appends the summary of accumulated values to the list of past summaries, and writes the list to the log file. Then, this extension makes a new fresh summary object which is used until the next time that the trigger fires.

It also adds some entries to each result dictionary.

- `'epoch'` and `'iteration'` are the epoch and iteration counts at the output, respectively.
- `'elapsed_time'` is the elapsed time in seconds since the training begins. The value is taken from `Trainer.elapsed_time`.

Parameters

- **keys** (*iterable of strs*) – Keys of values to accumulate. If this is None, all the values are accumulated and output to the log file.
- **trigger** – Trigger that decides when to aggregate the result and output the values. This is distinct from the trigger of this extension itself. If it is a tuple in the form `<int>, 'epoch'` or `<int>, 'iteration'`, it is passed to `IntervalTrigger`.
- **postprocess** – Callback to postprocess the result dictionaries. Each result dictionary is passed to this callback on the output. This callback can modify the result dictionaries, which are used to output to the log file.
- **log_name** (*str*) – Name of the log file under the output directory. It can be a format string: the last result dictionary is passed for the formatting. For example, users can use '{iteration}' to separate the log files for different iterations. If the log name is None, it does not output the log to any file.

log
The current list of observation dictionaries.

snapshot

chainer.training.extensions.**snapshot**(*savefun=<function save_npz>*, *filename='snapshot_iter_{.updater.iteration}'*, *trigger=(1, 'epoch')*)

Returns a trainer extension to take snapshots of the trainer.

This extension serializes the trainer object and saves it to the output directory. It is used to support resuming the training loop from the saved state.

This extension is called once for each epoch by default. The default priority is -100, which is lower than that of most built-in extensions.

Note: This extension first writes the serialized object to a temporary file and then rename it to the target file name. Thus, if the program stops right before the renaming, the temporary file might be left in the output directory.

Parameters

- **savefun** – Function to save the trainer. It takes two arguments: the output file path and the trainer object.
- **filename** (*str*) – Name of the file into which the trainer is serialized. It can be a format string, where the trainer object is passed to the `str.format()` method.
- **trigger** – Trigger that decides when to take snapshot. It can be either an already built trigger object (i.e., a callable object that accepts a trainer object and returns a bool value), or a tuple in the form `<int>, 'epoch'` or `<int>, 'iteration'`. In latter case, the tuple is passed to IntervalTrigger.

snapshot_object

chainer.training.extensions.**snapshot_object**(*target*, *filename*, *savefun=<function save_npz>*, *trigger=(1, 'epoch')*)

Returns a trainer extension to take snapshots of a given object.

This extension serializes the given object and saves it to the output directory.

This extension is called once for each epoch by default. The default priority is -100, which is lower than that of most built-in extensions.

Parameters

- **target** – Object to serialize.
- **filename** (*str*) – Name of the file into which the object is serialized. It can be a format string, where the trainer object is passed to the `str.format()` method. For example, `'snapshot_{.updater.iteration}'` is converted to `'snapshot_10000'` at the 10,000th iteration.
- **savefun** – Function to save the object. It takes two arguments: the output file path and the object to serialize.
- **trigger** – Trigger that decides when to take snapshot. It can be either an already built trigger object (i.e., a callable object that accepts a trainer object and returns a bool value), or a tuple in the form `<int>, 'epoch'` or `<int>, 'iteration'`. In latter case, the tuple is passed to IntervalTrigger.

Returns An extension function.

PlotReport

class chainer.training.extensions.**PlotReport**(*y_keys*, *x_key='iteration'*, *trigger=(1, 'epoch')*, *postprocess=None*, *file_name='plot.png'*, *marker='x'*, *grid=True*)

Trainer extension to output plots.

This extension accumulates the observations of the trainer to *DictSummary* at a regular interval specified by a supplied trigger, and plot a graph with using them.

There are two triggers to handle this extension. One is the trigger to invoke this extension, which is used to handle the timing of accumulating the results. It is set to `1, 'iteration'` by default. The other is the trigger to determine when to emit the result. When this trigger returns True, this extension appends the summary of accumulated values to the list of past summaries, and writes the list to the log file. Then, this extension makes a new fresh summary object which is used until the next time that the trigger fires.

It also adds `'epoch'` and `'iteration'` entries to each result dictionary, which are the epoch and iteration counts at the output.

Warning: If your environment needs to specify a backend of matplotlib explicitly, please call `matplotlib.use` before importing Chainer. For example:

```
import matplotlib
matplotlib.use('Agg')

import chainer
```

Then, once `chainer.training.extensions` is imported, `matplotlib.use` will have no effect.

For the details, please see here: http://matplotlib.org/faq/usage_faq.html#what-is-a-backend

Parameters

- **y_keys** (*iterable of strs*) – Keys of values regarded as y. If this is None, nothing is output to the graph.
- **x_key** (*str*) – Keys of values regarded as x. The default value is 'iteration'.
- **trigger** – Trigger that decides when to aggregate the result and output the values. This is distinct from the trigger of this extension itself. If it is a tuple in the form `<int>, 'epoch'` or `<int>, 'iteration'`, it is passed to `IntervalTrigger`.
- **postprocess** – Callback to postprocess the result dictionaries. Figure object, Axes object, and all plot data are passed to this callback in this order. This callback can modify the figure.
- **file_name** (*str*) – Name of the figure file under the output directory. It can be a format string.
- **marker** (*str*) – The marker used to plot the graph. Default is `'x'`. If `None` is given, it draws with no markers.
- **grid** (*bool*) – Set the axis grid on if True. Default is True.

PrintReport

class chainer.training.extensions.**PrintReport**(*entries*, *log_report='LogReport'*, *out=<open file '<stdout>', mode 'w'>*)

Trainer extension to print the accumulated results.

This extension uses the log accumulated by a *LogReport* extension to print specified entries of the log in a human-readable format.

Parameters

- **entries** (*list of str*) – List of keys of observations to print.
- **log_report** (*str or LogReport*) – Log report to accumulate the observations. This is either the name of a LogReport extensions registered to the trainer, or a LogReport instance to use internally.
- **out** – Stream to print the bar. Standard output is used by default.

ProgressBar

class chainer.training.extensions.**ProgressBar**(*training_length=None*, *update_interval=100*, *bar_length=50*, *out=<open file '<stdout>', mode 'w'>*)

Trainer extension to print a progress bar and recent training status.

This extension prints a progress bar at every call. It watches the current iteration and epoch to print the bar.

Parameters

- **training_length** (*tuple*) – Length of whole training. It consists of an integer and either `'epoch'` or `'iteration'`. If this value is omitted and the stop trigger of the trainer is `IntervalTrigger`, this extension uses its attributes to determine the length of the training.
- **update_interval** (*int*) – Number of iterations to skip printing the progress bar.

- **bar_length** (*int*) – Length of the progress bar in characters.
- **out** – Stream to print the bar. Standard output is used by default.

Trainer triggers

Interval

class chainer.training.triggers.**IntervalTrigger**(*period*, *unit*)
Trigger based on a fixed interval.

This trigger accepts iterations divided by a given interval. There are two ways to specify the interval: per iterations and epochs. *Iteration* means the number of updates, while *epoch* means the number of sweeps over the training dataset. Fractional values are allowed if the interval is a number of epochs; the trigger uses the *iteration* and *epoch_detail* attributes defined by the updater.

For the description of triggers, see *get_trigger()*.

Parameters

- **period** (*int or float*) – Length of the interval. Must be an integer if unit is `'iteration'`.
- **unit** (*str*) – Unit of the length specified by `period`. It must be either `'iteration'` or `'epoch'`.

ManualSchedule

class chainer.training.triggers.**ManualScheduleTrigger**(*points*, *unit*)
Trigger invoked at specified point(s) of iterations or epochs.

This trigger accepts iterations or epochs indicated by given point(s). There are two ways to specify the point(s): iteration and epoch. `iteration` means the number of updates, while `epoch` means the number of sweeps over the training dataset. Fractional values are allowed if the point is a number of epochs; the trigger uses the `iteration` and `epoch_detail` attributes defined by the updater.

Parameters

- **points** (*int, float, or list of int or float*) – time of the trigger. Must be an integer or list of integer if unit is `'iteration'`.
- **unit** (*str*) – Unit of the time specified by `points`. It must be either `'iteration'` or `'epoch'`.

Minimum and maximum values

class chainer.training.triggers.**MaxValueTrigger**(*key*, *trigger=(1*, *'epoch')*)
Trigger invoked when specific value becomes maximum.

For example you can use this trigger to take snapshot on the epoch the validation accuracy is maximum.

Parameters

- **key** (*str*) – Key of value. The trigger fires when the value associated with this key becomes maximum.

- **trigger** – Trigger that decides the comparison interval between current best value and new value. This must be a tuple in the form of `<int>, 'epoch'` or `<int>, 'iteration'` which is passed to *IntervalTrigger*.

class chainer.training.triggers.**MinValueTrigger**(*key*, *trigger=(1*, *'epoch')*)
Trigger invoked when specific value becomes minimum.

For example you can use this trigger to take snapshot on the epoch the validation loss is minimum.

Parameters

- **key** (*str*) – Key of value. The trigger fires when the value associated with this key becomes minimum.
- **trigger** – Trigger that decides the comparison interval between current best value and new value. This must be a tuple in the form of `<int>, 'epoch'` or `<int>, 'iteration'` which is passed to *IntervalTrigger*.

Caffe Reference Model Support

Caffe is a popular framework maintained by BVLC at UC Berkeley. It is widely used by computer vision communities, and aims at fast computation and easy usage without any programming. The BVLC team provides trained reference models in their Model Zoo, one of the reason why this framework gets popular.

Chainer can import the reference models and emulate the network by *Link* implementations. This functionality is provided by the *chainer.links.caffe.CaffeFunction* class.

class chainer.links.caffe.**CaffeFunction**(*model_path*)
Caffe emulator based on the model file of Caffe.

Given a protocol buffers file of a Caffe model, this class loads and emulates it on *Variable* objects. It supports the official reference models provided by BVLC.

Note: protobuf>=3.0.0 is required if you use Python 3 because protobuf 2 is not supported on Python 3.

Note: CaffeFunction ignores the following layers:

•Layers that CaffeFunction does not support (including data layers)

•Layers that have no top blobs

•Layers whose bottom blobs are incomplete (i.e., some or all of them are not given nor computed)

Warning: It does not support full compatibility against Caffe. Some layers and configurations are not implemented in Chainer yet, though the reference models provided by the BVLC team are supported except data layers.

Example

Consider we want to extract the (unnormalized) log class probability of given images using BVLC reference CaffeNet. The model can be downloaded from:

http://dl.caffe.berkeleyvision.org/bvlc_reference_caffenet.caffemodel

We want to compute the `fc8` blob from the `data` blob. It is simply written as follows:

```
# Load the model
func = CaffeFunction('path/to/bvlc_reference_caffenet.caffemodel')

# Minibatch of size 10
x_data = numpy.ndarray((10, 3, 227, 227), dtype=numpy.float32)
...  # (Fill the minibatch here)

# Forward the pre-trained net
x = Variable(x_data)
y, = func(inputs={'data': x}, outputs=['fc8'])
```

The result `y` contains the Variable corresponding to the `fc8` blob. The computational graph is memorized as a usual forward computation in Chainer, so we can run backprop through this pre-trained net.

Parameters **model_path** (*str*) – Path to the binary-proto model file of Caffe.

Variables **forwards** (*dict*) – A mapping from layer names to corresponding functions.

Visualization of Computational Graph

As neural networks get larger and complicated, it gets much harder to confirm if their architectures are constructed properly. Chainer supports visualization of computational graphs. Users can generate computational graphs by invoking *build_computational_graph()*. Generated computational graphs are dumped to specified format (Currently Dot Language is supported).

Basic usage is as follows:

```
import chainer.computational_graph as c
...
g = c.build_computational_graph(vs)
with open('path/to/output/file', 'w') as o:
    o.write(g.dump())
```

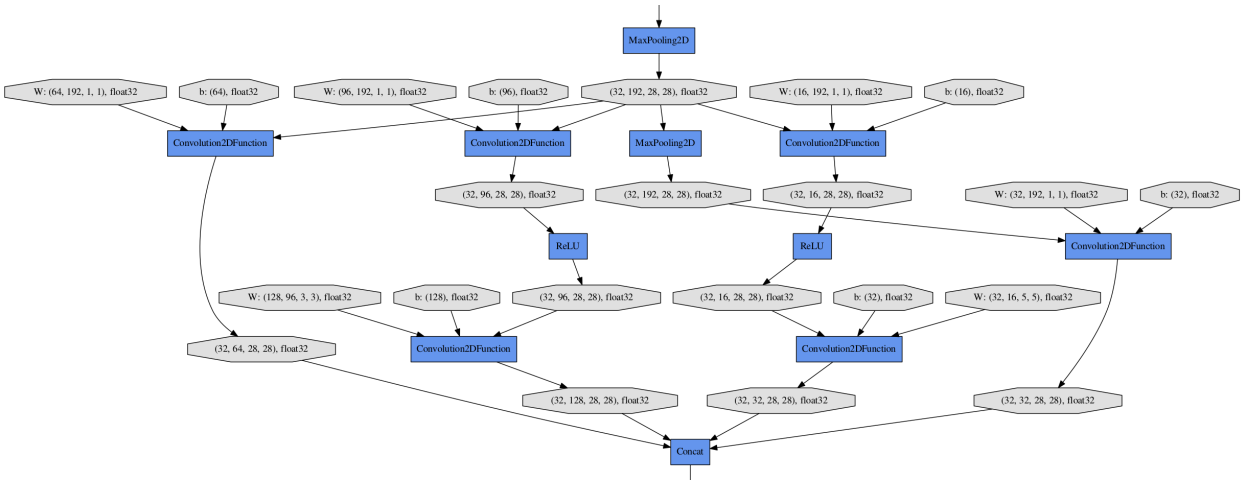
where `vs` is list of `Variable` instances and `g` is an instance of *ComputationalGraph*. This code generates the computational graph that are backward-reachable (i.e. reachable by repetition of steps backward) from at least one of `vs`.

Here is an example of (a part of) the generated graph (inception(3a) in GoogLeNet). This example is from `example/imagenet`.

`chainer.computational_graph.`**build_computational_graph**(*outputs*, *remove_split=True*, *variable_style={'shape': 'octagon', 'style': 'filled', 'fillcolor': '#E0E0E0'}*, *function_style={'shape': 'record', 'style': 'filled', 'fillcolor': '#6495ED'}*, *rankdir='TB'*, *remove_variable=False*, *show_name=True*)

Builds a graph of functions and variables backward-reachable from outputs.

Parameters



- **outputs** (*list*) – nodes from which the graph is constructed. Each element of outputs must be either `Variable` object or `Function` object.
- **remove_split** (*bool*) – It must be `True`. This argument is left for backward compatibility.
- **variable_style** (*dict*) – Dot node style for variable. Possible keys are 'shape', 'color', 'fillcolor', 'style', and etc.
- **function_style** (*dict*) – Dot node style for function.
- **rankdir** (*str*) – Direction of the graph that must be TB (top to bottom), BT (bottom to top), LR (left to right) or RL (right to left).
- **remove_variable** (*bool*) – If *True*, :class:'~chainer.Variable's are removed from the resulting computational graph. Only :class:'~chainer.Function's are shown in the output.
- **show_name** (*bool*) – If *True*, the *name* attribute of each node is added to the label of the node. Default is *True*.

Returns

A graph consisting of nodes and edges that are backward-reachable from at least one of `outputs`.

If `unchain_backward` was called in some variable in the computational graph before this function, backward step is stopped at this variable.

For example, suppose that computational graph is as follows:

```
    |--> f ---> y
x --+
    |--> g ---> z
```

Let `outputs = [y, z]`. Then the full graph is emitted.

Next, let `outputs = [y]`. Note that `z` and `g` are not backward-reachable from `y`. The resulting graph would be following:

```
x ---> f ---> y
```

See `TestGraphBuilder` for details.

Return type *ComputationalGraph*

Note: The default behavior of `ComputationalGraph` has been changed from v1.23.0, so that it ouputs the richest representation of a graph as default, namely, styles are set and names of functions and variables are shown. To reproduce the same result as previous versions (<= v1.22.0), please specify *variable_style=None*, *function_style=None*, and *show_name=False* explicitly.

class chainer.computational_graph.**ComputationalGraph**(*nodes*, *edges*, *variable_style={'shape': 'octagon', 'style': 'filled', 'fillcolor': '#E0E0E0'}*, *function_style={'shape': 'record', 'style': 'filled', 'fillcolor': '#6495ED'}*, *rankdir='TB'*, *remove_variable=False*, *show_name=True*)

Class that represents computational graph.

Note: We assume that the computational graph is directed and acyclic.

Parameters

- **nodes** (*list*) – List of nodes. Each node is either `Variable` object or `Function` object.
- **edges** (*list*) – List of edges. Each edge consists of pair of nodes.
- **variable_style** (*dict*) – Dot node style for variable.
- **function_style** (*dict*) – Dot node style for function.
- **rankdir** (*str*) – Direction of the graph that must be TB (top to bottom), BT (bottom to top), LR (left to right) or RL (right to left).
- **remove_variable** (*bool*) – If *True*, :class:‘~chainer.Variable‘s are removed from the resulting computational graph. Only :class:‘~chainer.Function‘s are shown in the output.
- **show_name** (*bool*) – If *True*, the *name* attribute of each node is added to the label of the node. Default is *True*.

Note: The default behavior of `ComputationalGraph` has been changed from v1.23.0, so that it ouputs the richest representation of a graph as default, namely, styles are set and names of functions and variables are shown. To reproduce the same result as previous versions (<= v1.22.0), please specify *variable_style=None*, *function_style=None*, and *show_name=False* explicitly.

dump(*format='dot'*)

Dumps graph as a text.

Parameters

- **format** (*str*) – The graph language name of the output.
- **it must be 'dot'.** (*Currently,*) –

Returns The graph in specified format.

Return type str

Environment variables

Here are the environment variables Chainer uses.

CHAINER_CUDNN	Set 0 to disable cuDNN in Chainer. Otherwise cuDNN is enabled automatically.
CHAINER_SEED	Default seed value of random number generators for CUDA. If it is not set, the seed value is generated from Python random module. Set an integer value in decimal format.
CHAINER_TYPE_CHECK	Set 0 to disable type checking. Otherwise type checking is enabled automatically. See *Function* for details.

CHAPTER 4

CuPy Reference Manual

This is the official documentation of CuPy, a multi-dimensional array on CUDA with a subset of NumPy interface.

CuPy Overview

CuPy is an implementation of NumPy-compatible multi-dimensional array on CUDA. CuPy consists of the core multi-dimensional array class, *cupy.ndarray*, and many functions on it. It supports a subset of numpy.ndarray interface that is enough for Chainer.

The following is a brief overview of supported subset of NumPy interface:

- Basic indexing (indexing by ints, slices, newaxes, and Ellipsis)
- Element types (dtypes): bool_, (u)int{8, 16, 32, 64}, float{16, 32, 64}
- Most of the array creation routines
- Reshaping and transposition
- All operators with broadcasting
- All Universal functions (a.k.a. ufuncs) for elementwise operations except those for complex numbers
- Dot product functions (except einsum) using cuBLAS
- Reduction along axes (sum, max, argmax, etc.)

CuPy also includes following features for performance:

- Customizable memory allocator, and a simple memory pool as an example
- User-defined elementwise kernels
- User-defined reduction kernels
- cuDNN utilities

CuPy uses on-the-fly kernel synthesis: when a kernel call is required, it compiles a kernel code optimized for the shapes and dtypes of given arguments, sends it to the GPU device, and executes the kernel. The compiled code is cached to `$(HOME)/.cupy/kernel_cache` directory (this cache path can be overwritten by setting the `CUPY_CACHE_DIR` environment variable). It may make things slower at the first kernel call, though this slow down will be resolved at the second execution. CuPy also caches the kernel code sent to GPU device within the process, which reduces the kernel transfer time on further calls.

A list of supported attributes, properties, and methods of ndarray

Memory layout

`base` `ctypes` *itemsize* *flags* *nbytes* *shape* `size` *strides*

Data type

`dtype`

Other attributes

T

Array conversion

tolist() *tofile()* *dump()* *dumps()* *astype()* *copy()* *view()* *fill()*

Shape manipulation

reshape() *transpose()* *swapaxes()* *ravel()* *squeeze()*

Item selection and manipulation

take() *diagonal()*

Calculation

max() *argmax()* *min()* *argmin()* *clip()* *trace()* *sum()* *mean()* *var()* *std()* *prod()* *dot()*

Arithmetic and comparison operations

`__lt__()` `__le__()` `__gt__()` `__ge__()` `__eq__()` `__ne__()` `__nonzero__()` `__neg__()` `__pos__()` `__abs__()` `__invert__()` `__add__()` `__sub__()` `__mul__()` `__div__()` `__truediv__()` `__floordiv__()` `__mod__()` `__divmod__()` `__pow__()` `__lshift__()` `__rshift__()` `__and__()` `__or__()` `__xor__()` `__iadd__()` `__isub__()` `__imul__()` `__idiv__()` `__itruediv__()` `__ifloordiv__()` `__imod__()` `__ipow__()` `__ilshift__()` `__irshift__()` `__iand__()` `__ior__()` `__ixor__()`

Special methods

`__copy__()` `__deepcopy__()` `__reduce__()` `__array__()` `__len__()` `__getitem__()` `__setitem__()` `__int__()` `__long__()` `__float__()` `__oct__()` `__hex__()` `__repr__()` `__str__()`

Memory transfer

get() set()

A list of supported routines of `cupy` module

Array creation routines

empty() empty_like() eye() identity() ones() ones_like() zeros() zeros_like() full() full_like()

array() asarray() ascontiguousarray() copy()

arange() linspace()

diag() diagflat()

Array manipulation routines

copyto()

reshape() ravel()

rollaxis() swapaxes() transpose()

atleast_1d() atleast_2d() atleast_3d() broadcast broadcast_arrays() broadcast_to() expand_dims() squeeze()

column_stack() concatenate() dstack() hstack() vstack()

array_split() dsplit() hsplit() split() vsplit()

roll()

Binary operations

bitwise_and bitwise_or bitwise_xor invert left_shift right_shift

Indexing routines

take() diagonal()

Input and output

load() save() savez() savez_compressed()

array_repr() array_str()

Linear algebra

dot() vdot() inner() outer() tensordot()

trace()

Logic functions

isfinite isinf isnan

logical_and logical_or logical_not logical_xor

greater greater_equal less less_equal equal not_equal

Mathematical functions

sin cos tan arcsin arccos arctan hypot arctan2 deg2rad rad2deg degrees radians

sinh cosh tanh arcsinh arccosh arctanh

rint floor ceil trunc

sum() prod()

exp expm1 exp2 log log10 log2 log1p logaddexp logaddexp2

signbit copysign ldexp frexp nextafter

add reciprocal negative multiply divide power subtract true_divide floor_divide fmod mod modf remainder

clip() sqrt square absolute sign maximum minimum fmax fmin

Sorting, searching, and counting

argmax() argmin() count_nonzero() nonzero() flatnonzero() where()

Statistics

amin() amax()

mean() var() std()

bincount()

Padding

pad()

External Functions

scatter_add()

Other

asnumpy()

Multi-Dimensional Array (ndarray)

class cupy.**ndarray**

Multi-dimensional array on a CUDA device.

This class implements a subset of methods of numpy.ndarray. The difference is that this class allocates the array content on the current GPU device.

Parameters

- **shape** (*tuple of ints*) – Length of axes.
- **dtype** – Data type. It must be an argument of numpy.dtype.
- **memptr** (cupy.cuda.MemoryPointer) – Pointer to the array content head.
- **strides** (*tuple of ints*) – The strides for axes.
- **order** (*{'C', 'F'}*) – Row-major (C-style) or column-major (Fortran-style) order.

Variables

- **base** (*None or cupy.ndarray*) – Base array from which this array is created as a view.
- **data** (cupy.cuda.MemoryPointer) – Pointer to the array content head.
- **dtype** (*numpy.dtype*) – Dtype object of element type.

  See also:

  Data type objects (dtype)
- ***size*** (*int*) – Number of elements this array holds.

  This is equivalent to product over the shape tuple.

  See also:

  numpy.ndarray.size

T

Shape-reversed view of the array.

If ndim < 2, then this is just a reference to the array itself.

argmax()

Returns the indices of the maximum along a given axis.

See also:

cupy.argmax() for full documentation, numpy.ndarray.argmax()

argmin()

Returns the indices of the minimum along a given axis.

See also:

cupy.argmin() for full documentation, numpy.ndarray.argmin()

astype()
Casts the array to given data type.

Parameters

- **dtype** – Type specifier.
- **copy** (*bool*) – If it is False and no cast happens, then this method returns the array itself. Otherwise, a copy is returned.

Returns If `copy` is False and no cast is required, then the array itself is returned. Otherwise, it returns a (possibly casted) copy of the array.

Note: This method currently does not support `order`, `casting`, and `subok` arguments.

See also:

`numpy.ndarray.astype()`

clip()
Returns an array with values limited to [a_min, a_max].

See also:

cupy.clip() for full documentation, `numpy.ndarray.clip()`

copy()
Returns a copy of the array.

Parameters **order** (*{'C', 'F'}*) – Row-major (C-style) or column-major (Fortran-style) order. This function currently does not support order 'A' and 'K'.

See also:

cupy.copy() for full documentation, `numpy.ndarray.copy()`

cstruct
C representation of the array.

This property is used for sending an array to CUDA kernels. The type of returned C structure is different for different dtypes and ndims. The definition of C type is written in `cupy/carray.cuh`.

device
CUDA device on which this array resides.

diagonal()
Returns a view of the specified diagonals.

See also:

cupy.diagonal() for full documentation, `numpy.ndarray.diagonal()`

dot()
Returns the dot product with given array.

See also:

cupy.dot() for full documentation, `numpy.ndarray.dot()`

dump()
Dumps a pickle of the array to a file.

Dumped file can be read back to *cupy.ndarray* by *cupy.load()*.

dumps()
Dumps a pickle of the array to a string.

fill()
Fills the array with a scalar value.

Parameters **value** – A scalar value to fill the array content.

See also:

numpy.ndarray.fill()

flags
Object containing memory-layout information.

It only contains c_contiguous, f_contiguous, and owndata attributes. All of these are read-only. Accessing by indexes is also supported.

See also:

numpy.ndarray.flags

flatten()
Returns a copy of the array flatten into one dimension.

It currently supports C-order only.

Returns A copy of the array with one dimension.

Return type *cupy.ndarray*

See also:

numpy.ndarray.flatten()

get()
Returns a copy of the array on host memory.

Parameters **stream** (cupy.cuda.Stream) – CUDA stream object. If it is given, the copy runs asynchronously. Otherwise, the copy is synchronous.

Returns Copy of the array on host memory.

Return type numpy.ndarray

itemsize
Size of each element in bytes.

See also:

numpy.ndarray.itemsize

max()
Returns the maximum along a given axis.

See also:

cupy.amax() for full documentation, numpy.ndarray.max()

mean()
Returns the mean along a given axis.

See also:

cupy.mean() for full documentation, numpy.ndarray.mean()

min()
Returns the minimum along a given axis.

See also:

cupy.amin() for full documentation, numpy.ndarray.min()

nbytes
Size of whole elements in bytes.

It does not count skips between elements.

See also:

numpy.ndarray.nbytes

ndim
Number of dimensions.

a.ndim is equivalent to len(a.shape).

See also:

numpy.ndarray.ndim

nonzero()
Return the indices of the elements that are non-zero.

Returned Array is containing the indices of the non-zero elements in that dimension.

Returns Indices of elements that are non-zero.

Return type tuple of arrays

See also:

numpy.nonzero()

prod()
Returns the product along a given axis.

See also:

cupy.prod() for full documentation, numpy.ndarray.prod()

ravel()
Returns an array flattened into one dimension.

See also:

cupy.ravel() for full documentation, numpy.ndarray.ravel()

reduced_view()
Returns a view of the array with minimum number of dimensions.

Parameters **dtype** – Data type specifier. If it is given, then the memory sequence is reinterpreted as the new type.

Returns A view of the array with reduced dimensions.

Return type *cupy.ndarray*

repeat()
Returns an array with repeated arrays along an axis.

See also:

cupy.repeat() for full documentation, numpy.ndarray.repeat()

reshape()
Returns an array of a different shape and the same content.

See also:

cupy.reshape() for full documentation, numpy.ndarray.reshape()

scatter_add()
Adds given values to specified elements of an array.

See also:

cupy.scatter_add() for full documentation.

set()
Copies an array on the host memory to *cupy.ndarray*.

Parameters

- **arr** (*numpy.ndarray*) – The source array on the host memory.
- **stream** (cupy.cuda.Stream) – CUDA stream object. If it is given, the copy runs asynchronously. Otherwise, the copy is synchronous.

shape
Lengths of axes.

Setter of this property involves reshaping without copy. If the array cannot be reshaped without copy, it raises an exception.

squeeze()
Returns a view with size-one axes removed.

See also:

cupy.squeeze() for full documentation, numpy.ndarray.squeeze()

std()
Returns the standard deviation along a given axis.

See also:

cupy.std() for full documentation, numpy.ndarray.std()

strides
Strides of axes in bytes.

See also:

numpy.ndarray.strides

sum()
Returns the sum along a given axis.

See also:

cupy.sum() for full documentation, numpy.ndarray.sum()

swapaxes()
Returns a view of the array with two axes swapped.

See also:

cupy.swapaxes() for full documentation, numpy.ndarray.swapaxes()

take()
Returns an array of elements at given indices along the axis.

See also:

cupy.take() for full documentation, numpy.ndarray.take()

tofile()
Writes the array to a file.

See also:

numpy.ndarray.tolist()

tolist()
Converts the array to a (possibly nested) Python list.

Returns The possibly nested Python list of array elements.

Return type list

See also:

numpy.ndarray.tolist()

trace()
Returns the sum along diagonals of the array.

See also:

cupy.trace() for full documentation, numpy.ndarray.trace()

transpose()
Returns a view of the array with axes permuted.

See also:

cupy.transpose() for full documentation, numpy.ndarray.reshape()

var()
Returns the variance along a given axis.

See also:

cupy.var() for full documentation, numpy.ndarray.var()

view()
Returns a view of the array.

Parameters **dtype** – If this is different from the data type of the array, the returned view reinterpret the memory sequence as an array of this type.

Returns A view of the array. A reference to the original array is stored at the base attribute.

Return type *cupy.ndarray*

See also:

numpy.ndarray.view()

cupy.**asnumpy**(*a*, *stream=None*)
Returns an array on the host memory from an arbitrary source array.

Parameters

- **a** – Arbitrary object that can be converted to numpy.ndarray.

- **stream** (cupy.cuda.Stream) – CUDA stream object. If it is specified, then the device-to-host copy runs asynchronously. Otherwise, the copy is synchronous. Note that if a is not a *cupy.ndarray* object, then this argument has no effect.

Returns Converted array on the host memory.

Return type numpy.ndarray

Universal Functions (ufunc)

CuPy provides universal functions (a.k.a. ufuncs) to support various elementwise operations. CuPy's ufunc supports following features of NumPy's one:

- Broadcasting
- Output type determination
- Casting rules

CuPy's ufunc currently does not provide methods such as `reduce`, `accumulate`, `reduceat`, `outer`, and `at`.

Ufunc class

class cupy.**ufunc**

Universal function.

Variables

- **name** (*str*) – The name of the universal function.
- **nin** (*int*) – Number of input arguments.
- **nout** (*int*) – Number of output arguments.
- **nargs** (*int*) – Number of all arguments.

types

A list of type signatures.

Each type signature is represented by type character codes of inputs and outputs separated by '->'.

Available ufuncs

Math operations

add subtract multiply divide logaddexp logaddexp2 true_divide floor_divide negative power remainder mod fmod absolute rint sign exp exp2 log log2 log10 expm1 log1p sqrt square reciprocal

Trigonometric functions

sin cos tan arcsin arccos arctan arctan2 hypot sinh cosh tanh arcsinh arccosh arctanh deg2rad rad2deg

Bit-twiddling functions

bitwise_and bitwise_or bitwise_xor invert left_shift right_shift

Comparison functions

greater greater_equal less less_equal not_equal equal logical_and logical_or logical_xor logical_not maximum minimum fmax fmin

Floating point values

isfinite isinf isnan signbit copysign nextafter modf ldexp frexp fmod floor ceil trunc

ufunc.at

Currently, CuPy does not support `at` for ufuncs in general. However, *cupy.scatter_add()* can substitute `add.at` as both behave identically.

Routines

The following pages describe NumPy-compatible routines. These functions cover a subset of NumPy routines.

Array Creation Routines

Basic creation routines

`cupy.`**`empty`**(*shape*, *dtype=<type 'float'>*, *order='C'*)
Returns an array without initializing the elements.

Parameters

- **shape** (*tuple of ints*) – Dimensionalities of the array.
- **dtype** – Data type specifier.
- **order** (*{'C', 'F'}*) – Row-major (C-style) or column-major (Fortran-style) order.

Returns A new array with elements not initialized.

Return type *cupy.ndarray*

See also:

`numpy.empty()`

`cupy.`**`empty_like`**(*a*, *dtype=None*)
Returns a new array with same shape and dtype of a given array.

This function currently does not support `order` and `subok` options.

Parameters

- **a** (`cupy.ndarray`) – Base array.
- **dtype** – Data type specifier. The data type of `a` is used by default.

Returns A new array with same shape and dtype of `a` with elements not initialized.

Return type *cupy.ndarray*

See also:

`numpy.empty_like()`

`cupy.eye`(*N*, *M=None*, *k=0*, *dtype=<type 'float'>*)
Returns a 2-D array with ones on the diagonals and zeros elsewhere.

Parameters

- **N** (*int*) – Number of rows.
- **M** (*int*) – Number of columns. M == N by default.
- **k** (*int*) – Index of the diagonal. Zero indicates the main diagonal, a positive index an upper diagonal, and a negative index a lower diagonal.
- **dtype** – Data type specifier.

Returns A 2-D array with given diagonals filled with ones and zeros elsewhere.

Return type *cupy.ndarray*

See also:

`numpy.eye()`

`cupy.identity`(*n*, *dtype=<type 'float'>*)
Returns a 2-D identity array.

It is equivalent to `eye(n, n, dtype)`.

Parameters

- **n** (*int*) – Number of rows and columns.
- **dtype** – Data type specifier.

Returns A 2-D identity array.

Return type *cupy.ndarray*

See also:

`numpy.identity()`

`cupy.ones`(*shape*, *dtype=<type 'float'>*)
Returns a new array of given shape and dtype, filled with ones.

This function currently does not support `order` option.

Parameters

- **shape** (*tuple of ints*) – Dimensionalities of the array.
- **dtype** – Data type specifier.

Returns An array filled with ones.

Return type *cupy.ndarray*

See also:

`numpy.ones()`

cupy.**ones_like**(*a*, *dtype=None*)
Returns an array of ones with same shape and dtype as a given array.

This function currently does not support `order` and `subok` options.

Parameters

- **a** (`cupy.ndarray`) – Base array.
- **dtype** – Data type specifier. The dtype of `a` is used by default.

Returns An array filled with ones.

Return type *cupy.ndarray*

See also:

`numpy.ones_like()`

cupy.**zeros**(*shape*, *dtype=<type 'float'>*, *order='C'*)
Returns a new array of given shape and dtype, filled with zeros.

Parameters

- **shape** (*tuple of ints*) – Dimensionalities of the array.
- **dtype** – Data type specifier.
- **order** (*{'C', 'F'}*) – Row-major (C-style) or column-major (Fortran-style) order.

Returns An array filled with ones.

Return type *cupy.ndarray*

See also:

`numpy.zeros()`

cupy.**zeros_like**(*a*, *dtype=None*)
Returns an array of zeros with same shape and dtype as a given array.

This function currently does not support `order` and `subok` options.

Parameters

- **a** (`cupy.ndarray`) – Base array.
- **dtype** – Data type specifier. The dtype of `a` is used by default.

Returns An array filled with ones.

Return type *cupy.ndarray*

See also:

`numpy.zeros_like()`

cupy.**full**(*shape*, *fill_value*, *dtype=None*)
Returns a new array of given shape and dtype, filled with a given value.

This function currently does not support `order` option.

Parameters

- **shape** (*tuple of ints*) – Dimensionalities of the array.
- **fill_value** – A scalar value to fill a new array.
- **dtype** – Data type specifier.

Returns An array filled with `fill_value`.

Return type *cupy.ndarray*

See also:

`numpy.full()`

`cupy.`**`full_like`**(*a*, *fill_value*, *dtype=None*)
Returns a full array with same shape and dtype as a given array.

This function currently does not support `order` and `subok` options.

Parameters

- **a** (`cupy.ndarray`) – Base array.
- **fill_value** – A scalar value to fill a new array.
- **dtype** – Data type specifier. The dtype of `a` is used by default.

Returns An array filled with `fill_value`.

Return type *cupy.ndarray*

See also:

`numpy.full_like()`

Creation from other data

`cupy.`**`array`**(*obj*, *dtype=None*, *copy=True*, *ndmin=0*)
Creates an array on the current device.

This function currently does not support the `order` and `subok` options.

Parameters

- **obj** – *cupy.ndarray* object or any other object that can be passed to `numpy.array()`.
- **dtype** – Data type specifier.
- **copy** (*bool*) – If `False`, this function returns `obj` if possible. Otherwise this function always returns a new array.
- **ndmin** (*int*) – Minimum number of dimensions. Ones are inserted to the head of the shape if needed.

Returns An array on the current device.

Return type *cupy.ndarray*

See also:

`numpy.array()`

`cupy.`**`asarray`**(*a*, *dtype=None*)
Converts an object to array.

This is equivalent to `array(a, dtype, copy=False)`. This function currently does not support the `order` option.

Parameters

- **a** – The source object.

- **dtype** – Data type specifier. It is inferred from the input by default.

Returns An array on the current device. If `a` is already on the device, no copy is performed.

Return type *cupy.ndarray*

See also:

`numpy.asarray()`

`cupy.`**asanyarray**(*a*, *dtype=None*)
Converts an object to array.

This is currently equivalent to *`asarray()`*, since there is no subclass of ndarray in CuPy. Note that the original `numpy.asanyarray()` returns the input array as is if it is an instance of a subtype of `numpy.ndarray`.

See also:

`cupy.asarray()`, `numpy.asanyarray()`

`cupy.`**ascontiguousarray**(*a*, *dtype=None*)
Returns a C-contiguous array.

Parameters

- **a** (`cupy.ndarray`) – Source array.
- **dtype** – Data type specifier.

Returns If no copy is required, it returns `a`. Otherwise, it returns a copy of `a`.

Return type *cupy.ndarray*

See also:

`numpy.ascontiguousarray()`

`cupy.`**copy**(**args*, ***kwargs*)

Numerical ranges

`cupy.`**arange**(*start*, *stop=None*, *step=1*, *dtype=None*)
Returns an array with evenly spaced values within a given interval.

Values are generated within the half-open interval [start, stop). The first three arguments are mapped like the `range` built-in function, i.e. start and step are optional.

Parameters

- **start** – Start of the interval.
- **stop** – End of the interval.
- **step** – Step width between each pair of consecutive values.
- **dtype** – Data type specifier. It is inferred from other arguments by default.

Returns The 1-D array of range values.

Return type *cupy.ndarray*

See also:

`numpy.arange()`

cupy.**linspace**(*start*, *stop*, *num=50*, *endpoint=True*, *retstep=False*, *dtype=None*)
Returns an array with evenly-spaced values within a given interval.

Instead of specifying the step width like *cupy.arange()*, this function requires the total number of elements specified.

Parameters

- **start** – Start of the interval.
- **stop** – End of the interval.
- **num** – Number of elements.
- **endpoint** (*bool*) – If `True`, the stop value is included as the last element. Otherwise, the stop value is omitted.
- **retstep** (*bool*) – If `True`, this function returns (array, step). Otherwise, it returns only the array.
- **dtype** – Data type specifier. It is inferred from the start and stop arguments by default.

Returns The 1-D array of ranged values.

Return type *cupy.ndarray*

cupy.**logspace**(*start*, *stop*, *num=50*, *endpoint=True*, *base=10.0*, *dtype=None*)
Returns an array with evenly-spaced values on a log-scale.

Instead of specifying the step width like *cupy.arange()*, this function requires the total number of elements specified.

Parameters

- **start** – Start of the interval.
- **stop** – End of the interval.
- **num** – Number of elements.
- **endpoint** (*bool*) – If `True`, the stop value is included as the last element. Otherwise, the stop value is omitted.
- **base** (*float*) – Base of the log space. The step sizes between the elements on a log-scale are the same as `base`.
- **dtype** – Data type specifier. It is inferred from the start and stop arguments by default.

Returns The 1-D array of ranged values.

Return type *cupy.ndarray*

cupy.**meshgrid**(**xi*, ***kwargs*)
Return coordinate matrices from coordinate vectors.

Given one-dimensional coordinate arrays x1, x2, ..., xn, this function makes N-D grids.

For one-dimensional arrays x1, x2, ..., xn with lengths `Ni = len(xi)`, this function returns `(N1, N2, N3, ..., Nn)` shaped arrays if indexing='ij' or `(N2, N1, N3, ..., Nn)` shaped arrays if indexing='xy'.

Unlike NumPy, CuPy currently only supports 1-D arrays as inputs. Also, CuPy does not support `sparse` option yet.

Parameters

- **xi** (*tuple of ndarrays*) – 1-D arrays representing the coordinates of a grid.

- **indexing** (*{'xy', 'ij'}, optional*) – Cartesian ('xy', default) or matrix ('ij') indexing of output.
- **copy** (*bool, optional*) – If `False`, a view into the original arrays are returned. Default is True.

Returns list of cupy.ndarray

See also:

`numpy.meshgrid()`

Matrix creation

`cupy.`**`diag`**(*v*, *k=0*)

Returns a diagonal or a diagonal array.

Parameters

- **v** (*array-like*) – Array or array-like object.
- **k** (*int*) – Index of diagonals. Zero indicates the main diagonal, a positive value an upper diagonal, and a negative value a lower diagonal.

Returns If `v` indicates a 1-D array, then it returns a 2-D array with the specified diagonal filled by `v`. If `v` indicates a 2-D array, then it returns the specified diagonal of `v`. In latter case, if `v` is a *cupy.ndarray* object, then its view is returned.

Return type *cupy.ndarray*

See also:

`numpy.diag()`

`cupy.`**`diagflat`**(*v*, *k=0*)

Creates a diagonal array from the flattened input.

Parameters

- **v** (*array-like*) – Array or array-like object.
- **k** (*int*) – Index of diagonals. See *cupy.diag()* for detail.

Returns A 2-D diagonal array with the diagonal copied from `v`.

Return type *cupy.ndarray*

Array Manipulation Routines

Basic manipulations

`cupy.`**`copyto`**(*dst*, *src*, *casting='same_kind'*, *where=None*)

Copies values from one array to another with broadcasting.

This function can be called for arrays on different devices. In this case, casting, `where`, and broadcasting is not supported, and an exception is raised if these are used.

Parameters

- **dst** (`cupy.ndarray`) – Target array.
- **src** (`cupy.ndarray`) – Source array.

- **casting** (*str*) – Casting rule. See numpy.can_cast() for detail.
- **where** (*cupy.ndarray of bool*) – If specified, this array acts as a mask, and an element is copied only if the corresponding element of where is True.

See also:

numpy.copyto()

Shape manipulation

cupy.**reshape**(*a*, *newshape*)

Returns an array with new shape and same elements.

It tries to return a view if possible, otherwise returns a copy.

This function currently does not support order option.

Parameters

- **a** (cupy.ndarray) – Array to be reshaped.
- **newshape** (*int or tuple of ints*) – The new shape of the array to return. If it is an integer, then it is treated as a tuple of length one. It should be compatible with a.size. One of the elements can be -1, which is automatically replaced with the appropriate value to make the shape compatible with a.size.

Returns A reshaped view of a if possible, otherwise a copy.

Return type *cupy.ndarray*

See also:

numpy.reshape()

cupy.**ravel**(*a*)

Returns a flattened array.

It tries to return a view if possible, otherwise returns a copy.

This function currently does not support order option.

Parameters **a** (cupy.ndarray) – Array to be flattened.

Returns A flattened view of a if possible, otherwise a copy.

Return type *cupy.ndarray*

See also:

numpy.ravel()

Transposition

cupy.**rollaxis**(*a*, *axis*, *start=0*)

Moves the specified axis backwards to the given place.

Parameters

- **a** (cupy.ndarray) – Array to move the axis.
- **axis** (*int*) – The axis to move.
- **start** (*int*) – The place to which the axis is moved.

Returns A view of `a` that the axis is moved to `start`.

Return type *cupy.ndarray*

See also:

`numpy.rollaxis()`

cupy.**swapaxes**(*a*, *axis1*, *axis2*)
Swaps the two axes.

Parameters

- **a** (`cupy.ndarray`) – Array to swap the axes.
- **axis1** (*int*) – The first axis to swap.
- **axis2** (*int*) – The second axis to swap.

Returns A view of `a` that the two axes are swapped.

Return type *cupy.ndarray*

See also:

`numpy.swapaxes()`

cupy.**transpose**(*a*, *axes=None*)
Permutes the dimensions of an array.

Parameters

- **a** (`cupy.ndarray`) – Array to permute the dimensions.
- **axes** (*tuple of ints*) – Permutation of the dimensions. This function reverses the shape by default.

Returns A view of `a` that the dimensions are permuted.

Return type *cupy.ndarray*

See also:

`numpy.transpose()`

Edit dimensionalities

cupy.**atleast_1d**(**arys*)
Converts arrays to arrays with dimensions >= 1.

Parameters **arys** (*tuple of arrays*) – Arrays to be converted. All arguments must be *cupy.ndarray* objects. Only zero-dimensional array is affected.

Returns If there are only one input, then it returns its converted version. Otherwise, it returns a list of converted arrays.

See also:

`numpy.atleast_1d()`

cupy.**atleast_2d**(**arys*)
Converts arrays to arrays with dimensions >= 2.

If an input array has dimensions less than two, then this function inserts new axes at the head of dimensions to make it have two dimensions.

Parameters **arys** (*tuple of arrays*) – Arrays to be converted. All arguments must be *cupy.ndarray* objects.

Returns If there are only one input, then it returns its converted version. Otherwise, it returns a list of converted arrays.

See also:

numpy.atleast_2d()

cupy.**atleast_3d**(**arys*)
Converts arrays to arrays with dimensions >= 3.

If an input array has dimensions less than three, then this function inserts new axes to make it have three dimensions. The place of the new axes are following:

- If its shape is `()`, then the shape of output is `(1, 1, 1)`.
- If its shape is `(N,)`, then the shape of output is `(1, N, 1)`.
- If its shape is `(M, N)`, then the shape of output is `(M, N, 1)`.
- Otherwise, the output is the input array itself.

Parameters **arys** (*tuple of arrays*) – Arrays to be converted. All arguments must be *cupy.ndarray* objects.

Returns If there are only one input, then it returns its converted version. Otherwise, it returns a list of converted arrays.

See also:

numpy.atleast_3d()

class cupy.**broadcast**
Object that performs broadcasting.

CuPy actually uses this class to support broadcasting in various operations. Note that this class does not provide an iterator.

Parameters **arrays** (*tuple of arrays*) – Arrays to be broadcasted.

Variables

- ***shape*** (*tuple of ints*) – The broadcasted shape.
- **nd** (*int*) – Number of dimensions of the broadcasted shape.
- ***size*** (*int*) – Total size of the broadcasted shape.
- **values** (*list of arrays*) – The broadcasted arrays.

See also:

numpy.broadcast

cupy.**broadcast_arrays**(**args*)
Broadcasts given arrays.

Parameters **args** (*tuple of arrays*) – Arrays to broadcast for each other.

Returns A list of broadcasted arrays.

Return type list

See also:

numpy.broadcast_arrays()

cupy.**broadcast_to**(*array*, *shape*)
Broadcast an array to a given shape.

Parameters

- **array** (cupy.ndarray) – Array to broadcast.
- **shape** (*tuple of int*) – The shape of the desired array.

Returns Broadcasted view.

Return type *cupy.ndarray*

See also:

numpy.broadcast_to()

cupy.**expand_dims**(*a*, *axis*)
Expands given arrays.

Parameters

- **a** (cupy.ndarray) – Array to be expanded.
- **axis** (*int*) – Position where new axis is to be inserted.

Returns

The number of dimensions is one greater than that of the input array.

Return type *cupy.ndarray*

See also:

numpy.expand_dims()

cupy.**squeeze**(*a*, *axis=None*)
Removes size-one axes from the shape of an array.

Parameters

- **a** (cupy.ndarray) – Array to be reshaped.
- **axis** (*int or tuple of ints*) – Axes to be removed. This function removes all size-one axes by default. If one of the specified axes is not of size one, an exception is raised.

Returns An array without (specified) size-one axes.

Return type *cupy.ndarray*

See also:

numpy.squeeze()

Changing kind of array

cupy.**asfortranarray**(*a*, *dtype=None*)
Return an array laid out in Fortran order in memory.

Parameters

- **a** (ndarray) – The input array.

- **dtype** (*str or dtype object, optional*) – By default, the data-type is inferred from the input data.

Returns The input *a* in Fortran, or column-major, order.

Return type *ndarray*

See also:

numpy.asfortranarray()

Joining arrays along axis

cupy.**column_stack**(*tup*)

Stacks 1-D and 2-D arrays as columns into a 2-D array.

A 1-D array is first converted to a 2-D column array. Then, the 2-D arrays are concatenated along the second axis.

Parameters **tup** (*sequence of arrays*) – 1-D or 2-D arrays to be stacked.

Returns A new 2-D array of stacked columns.

Return type *cupy.ndarray*

See also:

numpy.column_stack()

cupy.**concatenate**(*tup*, *axis=0*)

Joins arrays along an axis.

Parameters

- **tup** (*sequence of arrays*) – Arrays to be joined. All of these should have same dimensionalities except the specified axis.
- **axis** (*int*) – The axis to join arrays along.

Returns Joined array.

Return type *cupy.ndarray*

See also:

numpy.concatenate()

cupy.**vstack**(*tup*)

Stacks arrays vertically.

If an input array has one dimension, then the array is treated as a horizontal vector and stacked along the additional axis at the head. Otherwise, the array is stacked along the first axis.

Parameters **tup** (*sequence of arrays*) – Arrays to be stacked. Each array is converted by *cupy.atleast_2d()* before stacking.

Returns Stacked array.

Return type *cupy.ndarray*

See also:

numpy.dstack()

cupy.**hstack**(*tup*)
Stacks arrays horizontally.

If an input array has one dimension, then the array is treated as a horizontal vector and stacked along the first axis. Otherwise, the array is stacked along the second axis.

Parameters **tup** (*sequence of arrays*) – Arrays to be stacked.

Returns Stacked array.

Return type *cupy.ndarray*

See also:

numpy.hstack()

cupy.**dstack**(*tup*)
Stacks arrays along the third axis.

Parameters **tup** (*sequence of arrays*) – Arrays to be stacked. Each array is converted by *cupy.atleast_3d()* before stacking.

Returns Stacked array.

Return type *cupy.ndarray*

See also:

numpy.dstack()

cupy.**stack**(*tup*, *axis=0*)
Stacks arrays along a new axis.

Parameters

- **tup** (*sequence of arrays*) – Arrays to be stacked.
- **axis** (*int*) – Axis along which the arrays are stacked.

Returns Stacked array.

Return type *cupy.ndarray*

See also:

numpy.stack()

Splitting arrays along axis

cupy.**array_split**(*ary*, *indices_or_sections*, *axis=0*)
Splits an array into multiple sub arrays along a given axis.

This function is almost equivalent to *cupy.split()*. The only difference is that this function allows an integer sections that does not evenly divide the axis.

See also:

cupy.split() for more detail, numpy.array_split()

cupy.**split**(*ary*, *indices_or_sections*, *axis=0*)
Splits an array into multiple sub arrays along a given axis.

Parameters

- **ary** (cupy.ndarray) – Array to split.

- **indices_or_sections** (*int or sequence of ints*) – A value indicating how to divide the axis. If it is an integer, then is treated as the number of sections, and the axis is evenly divided. Otherwise, the integers indicate indices to split at. Note that the sequence on the device memory is not allowed.
- **axis** (*int*) – Axis along which the array is split.

Returns A list of sub arrays. Each array is a view of the corresponding input array.

See also:

`numpy.split()`

`cupy.`**vsplit**(*ary*, *indices_or_sections*)
Splits an array into multiple sub arrays along the first axis.

This is equivalent to `split` with `axis=0`.

See also:

`cupy.split()` for more detail, `numpy.dsplit()`

`cupy.`**hsplit**(*ary*, *indices_or_sections*)
Splits an array into multiple sub arrays horizontally.

This is equivalent to `split` with `axis=0` if `ary` has one dimension, and otherwise that with `axis=1`.

See also:

`cupy.split()` for more detail, `numpy.hsplit()`

`cupy.`**dsplit**(*ary*, *indices_or_sections*)
Splits an array into multiple sub arrays along the third axis.

This is equivalent to `split` with `axis=2`.

See also:

`cupy.split()` for more detail, `numpy.dsplit()`

Repeating part of arrays along axis

`cupy.`**tile**(*A*, *reps*)
Construct an array by repeating A the number of times given by reps.

Parameters

- **A** (`cupy.ndarray`) – Array to transform.
- **reps** (*int or tuple*) – The number of repeats.

Returns Transformed array with repeats.

Return type *cupy.ndarray*

See also:

`numpy.tile()`

`cupy.`**repeat**(*a*, *repeats*, *axis=None*)
Repeat arrays along an axis.

Parameters

- **a** (`cupy.ndarray`) – Array to transform.
- **repeats** (*int, list or tuple*) – The number of repeats.

- **axis** (*int*) – The axis to repeat.

Returns Transformed array with repeats.

Return type *cupy.ndarray*

See also:

`numpy.repeat()`

Rearranging elements

`cupy.`**`flip`**(*a*, *axis*)
Reverse the order of elements in an array along the given axis.

Note that `flip` function has been introduced since NumPy v1.12. The contents of this document is the same as the original one.

Parameters

- **a** (`ndarray`) – Input array.
- **axis** (*int*) – Axis in array, which entries are reversed.

Returns Output array.

Return type *ndarray*

See also:

`numpy.flip()`

`cupy.`**`fliplr`**(*a*)
Flip array in the left/right direction.

Flip the entries in each row in the left/right direction. Columns are preserved, but appear in a different order than before.

Parameters **a** (`ndarray`) – Input array.

Returns Output array.

Return type *ndarray*

See also:

`numpy.fliplr()`

`cupy.`**`flipud`**(*a*)
Flip array in the up/down direction.

Flip the entries in each column in the up/down direction. Rows are preserved, but appear in a different order than before.

Parameters **a** (`ndarray`) – Input array.

Returns Output array.

Return type *ndarray*

See also:

`numpy.flipud()`

`cupy.`**`roll`**(*a*, *shift*, *axis=None*)
Roll array elements along a given axis.

Parameters

- **a** (ndarray) – Array to be rolled.
- **shift** (*int*) – The number of places by which elements are shifted.
- **axis** (*int or None*) – The axis along which elements are shifted. If `axis` is `None`, the array is flattened before shifting, and after that it is reshaped to the original shape.

Returns Output array.

Return type *ndarray*

See also:

`numpy.roll()`

cupy.**rot90**(*a*, *k=1*, *axes=(0, 1)*)

Rotate an array by 90 degrees in the plane specified by axes.

Note that `axes` argument has been introduced since NumPy v1.12. The contents of this document is the same as the original one.

Parameters

- **a** (ndarray) – Array of two or more dimensions.
- **k** (*int*) – Number of times the array is rotated by 90 degrees.
- **axes** – (tuple of ints): The array is rotated in the plane defined by the axes. Axes must be different.

Returns Output array.

Return type *ndarray*

See also:

`numpy.rot90()`

Binary Operations

Elementwise bit operations

cupy.**bitwise_and** = <ufunc 'cupy_bitwise_and'>

Computes the bitwise AND of two arrays elementwise.

Only integer and boolean arrays are handled.

See also:

`numpy.bitwise_and`

cupy.**bitwise_or** = <ufunc 'cupy_bitwise_or'>

Computes the bitwise OR of two arrays elementwise.

Only integer and boolean arrays are handled.

See also:

`numpy.bitwise_or`

cupy.**bitwise_xor** = <ufunc 'cupy_bitwise_xor'>

Computes the bitwise XOR of two arrays elementwise.

Only integer and boolean arrays are handled.

See also:

numpy.bitwise_xor

cupy.**invert** = <ufunc 'cupy_invert'>
Computes the bitwise NOT of an array elementwise.

Only integer and boolean arrays are handled.

See also:

numpy.invert

cupy.**left_shift** = <ufunc 'cupy_left_shift'>
Shifts the bits of each integer element to the left.

Only integer arrays are handled.

See also:

numpy.left_shift

cupy.**right_shift** = <ufunc 'cupy_right_shift'>
Shifts the bits of each integer element to the right.

Only integer arrays are handled

See also:

numpy.right_shift

Indexing Routines

cupy.**take**(*a*, *indices*, *axis=None*, *out=None*)
Takes elements of an array at specified indices along an axis.

This is an implementation of "fancy indexing" at single axis.

This function does not support `mode` option.

Parameters

- **a** (cupy.ndarray) – Array to extract elements.
- **indices** (*int or array-like*) – Indices of elements that this function takes.
- **axis** (*int*) – The axis along which to select indices. The flattened input is used by default.
- **out** (cupy.ndarray) – Output array. If provided, it should be of appropriate shape and dtype.

Returns The result of fancy indexing.

Return type *cupy.ndarray*

See also:

numpy.take()

cupy.**diagonal**(*a*, *offset=0*, *axis1=0*, *axis2=1*)
Returns specified diagonals.

This function extracts the diagonals along two specified axes. The other axes are not changed. This function returns a writable view of this array as NumPy 1.10 will do.

Parameters

- **a** (cupy.ndarray) – Array from which the diagonals are taken.
- **offset** (*int*) – Index of the diagonals. Zero indicates the main diagonals, a positive value upper diagonals, and a negative value lower diagonals.
- **axis1** (*int*) – The first axis to take diagonals from.
- **axis2** (*int*) – The second axis to take diagonals from.

Returns A view of the diagonals of `a`.

Return type *cupy.ndarray*

See also:

`numpy.diagonal()`

cupy.**ix_**(**args*)

Construct an open mesh from multiple sequences.

This function takes N 1-D sequences and returns N outputs with N dimensions each, such that the shape is 1 in all but one dimension and the dimension with the non-unit shape value cycles through all N dimensions.

Using *ix_* one can quickly construct index arrays that will index the cross product. `a[cupy.ix_([1,3], [2,5])]` returns the array `[[a[1,2] a[1,5]], [a[3,2] a[3,5]]]`.

Parameters ***args** – 1-D sequences

Returns N arrays with N dimensions each, with N the number of input sequences. Together these arrays form an open mesh.

Return type tuple of ndarrays

Examples

```
>>> a = cupy.arange(10).reshape(2, 5)
>>> a
array([[0, 1, 2, 3, 4],
       [5, 6, 7, 8, 9]])
>>> ixgrid = cupy.ix_([0,1], [2,4])
>>> ixgrid
(array([[0],
       [1]]), array([[2, 4]]))
```

See also:

`numpy.ix_()`

cupy.**fill_diagonal**(*a*, *val*, *wrap=False*)

Fill the main diagonal of the given array of any dimensionality.

For an array *a* with `a.ndim > 2`, the diagonal is the list of locations with indices `a[i, i, ..., i]` all identical. This function modifies the input array in-place, it does not return a value.

Parameters

- **a** (cupy.ndarray) – The array, at least 2-D.
- **val** (*scalar*) – The value to be written on the diagonal. Its type must be compatible with that of the array a.

- **wrap** (*bool*) – If specified, the diagonal is "wrapped" after N columns. This affects only tall matrices.

Examples

```
>>> a = cupy.zeros((3, 3), int)
>>> cupy.fill_diagonal(a, 5)
>>> a
array([[5, 0, 0],
       [0, 5, 0],
       [0, 0, 5]])
```

See also:

`numpy.fill_diagonal()`

cupy.**c_ = <cupy.indexing.generate.CClass object>**

Translates slice objects to concatenation along the second axis.

This is a CuPy object that corresponds to *cupy.r_()*, which is useful because of its common occurrence. In particular, arrays will be stacked along their last axis after being upgraded to at least 2-D with 1's post-pended to the shape (column vectors made out of 1-D arrays).

For detailed documentation, see `r_()`.

This implementation is partially borrowed from NumPy's one.

Parameters **a function, so takes no parameters** (*Not*) –

Returns Joined array.

Return type *cupy.ndarray*

See also:

`numpy.c_()`

Examples

```
>>> a = cupy.array([[1, 2, 3]], dtype=np.int32)
>>> b = cupy.array([[4, 5, 6]], dtype=np.int32)
>>> cupy.c_[a, 0, 0, b]
array([[1, 2, 3, 0, 0, 4, 5, 6]], dtype=int32)
```

cupy.**r_ = <cupy.indexing.generate.RClass object>**

Translates slice objects to concatenation along the first axis.

This is a simple way to build up arrays quickly. If the index expression contains comma separated arrays, then stack them along their first axis.

This object can build up from normal CuPy arrays. Therefore, the other objects (e.g. writing strings like '2,3,4', or using imaginary numbers like [1,2,3j], or using string integers like '-1') are not implemented yet compared with NumPy.

This implementation is partially borrowed from NumPy's one.

Parameters **a function, so takes no parameters** (*Not*) –

Returns Joined array.

Return type *cupy.ndarray*

See also:

`numpy.r_()`

Examples

```
>>> a = cupy.array([1, 2, 3], dtype=np.int32)
>>> b = cupy.array([4, 5, 6], dtype=np.int32)
>>> cupy.r_[a, 0, 0, b]
array([1, 2, 3, 0, 0, 4, 5, 6], dtype=int32)
```

Input and Output

NPZ files

`cupy.`**`load`**(*file*, *mmap_mode=None*)

Loads arrays or pickled objects from `.npy`, `.npz` or pickled file.

This function just calls `numpy.load` and then sends the arrays to the current device. NPZ file is converted to NpzFile object, which defers the transfer to the time of accessing the items.

Parameters

- **`file`** (*file-like object or string*) – The file to read.
- **`mmap_mode`** (*None, 'r+', 'r', 'w+', 'c'*) – If not `None`, memory-map the file to construct an intermediate `numpy.ndarray` object and transfer it to the current device.

Returns CuPy array or NpzFile object depending on the type of the file. NpzFile object is a dictionary-like object with the context manager protocol (which enables us to use *with* statement on it).

See also:

`numpy.load()`

`cupy.`**`save`**(*file*, *arr*)

Saves an array to a binary file in `.npy` format.

Parameters

- **`file`** (*file or str*) – File or filename to save.
- **`arr`** (*array_like*) – Array to save. It should be able to feed to *cupy.asnumpy()*.

See also:

`numpy.save()`

`cupy.`**`savez`**(*file*, **args*, ***kwds*)

Saves one or more arrays into a file in uncompressed `.npz` format.

Arguments without keys are treated as arguments with automatic keys named `arr_0`, `arr_1`, etc. corresponding to the positions in the argument list. The keys of arguments are used as keys in the `.npz` file, which are used for accessing NpzFile object when the file is read by *cupy.load()* function.

Parameters

- **`file`** (*file or str*) – File or filename to save.

- ***args** – Arrays with implicit keys.
- ****kwds** – Arrays with explicit keys.

See also:

`numpy.savez()`

cupy.**savez_compressed**(*file*, **args*, ***kwds*)
Saves one or more arrays into a file in compressed `.npz` format.

It is equivalent to *`cupy.savez()`* function except the output file is compressed.

See also:

`cupy.savez()` for more detail, `numpy.savez_compressed()`

String formatting

cupy.**array_repr**(*arr*, *max_line_width=None*, *precision=None*, *suppress_small=None*)
Returns the string representation of an array.

Parameters

- **arr** (*`array_like`*) – Input array. It should be able to feed to *`cupy.asnumpy()`*.
- **max_line_width** (*`int`*) – The maximum number of line lengths.
- **precision** (*`int`*) – Floating point precision. It uses the current printing precision of NumPy.
- **suppress_small** (*`bool`*) – If `True`, very small numbers are printed as zeros

Returns The string representation of `arr`.

Return type str

See also:

`numpy.array_repr()`

cupy.**array_str**(*arr*, *max_line_width=None*, *precision=None*, *suppress_small=None*)
Returns the string representation of the content of an array.

Parameters

- **arr** (*`array_like`*) – Input array. It should be able to feed to *`cupy.asnumpy()`*.
- **max_line_width** (*`int`*) – The maximum number of line lengths.
- **precision** (*`int`*) – Floating point precision. It uses the current printing precision of NumPy.
- **suppress_small** (*`bool`*) – If `True`, very small number are printed as zeros.

See also:

`numpy.array_str()`

Linear Algebra

Matrix and vector products

cupy.**dot**(*a*, *b*, *out=None*)
Returns a dot product of two arrays.

For arrays with more than one axis, it computes the dot product along the last axis of a and the second-to-last axis of b. This is just a matrix product if the both arrays are 2-D. For 1-D arrays, it uses their unique axis as an axis to take dot product over.

Parameters

- **a** (cupy.ndarray) – The left argument.
- **b** (cupy.ndarray) – The right argument.
- **out** (cupy.ndarray) – Output array.

Returns The dot product of a and b.

Return type *cupy.ndarray*

See also:

numpy.dot()

cupy.**vdot**(*a*, *b*)
Returns the dot product of two vectors.

The input arrays are flattened into 1-D vectors and then it performs inner product of these vectors.

Parameters

- **a** (cupy.ndarray) – The first argument.
- **b** (cupy.ndarray) – The second argument.

Returns Zero-dimensional array of the dot product result.

Return type *cupy.ndarray*

See also:

numpy.vdot()

cupy.**inner**(*a*, *b*)
Returns the inner product of two arrays.

It uses the last axis of each argument to take sum product.

Parameters

- **a** (cupy.ndarray) – The first argument.
- **b** (cupy.ndarray) – The second argument.

Returns The inner product of a and b.

Return type *cupy.ndarray*

See also:

numpy.inner()

cupy.**outer**(*a*, *b*, *out=None*)
Returns the outer product of two vectors.

The input arrays are flattened into 1-D vectors and then it performs outer product of these vectors.

Parameters

- **a** (cupy.ndarray) – The first argument.
- **b** (cupy.ndarray) – The second argument.
- **out** (cupy.ndarray) – Output array.

Returns 2-D array of the outer product of a and b.

Return type *cupy.ndarray*

See also:

numpy.outer()

cupy.**tensordot**(*a*, *b*, *axes=2*)

Returns the tensor dot product of two arrays along specified axes.

This is equivalent to compute dot product along the specified axes which are treated as one axis by reshaping.

Parameters

- **a** (cupy.ndarray) – The first argument.
- **b** (cupy.ndarray) – The second argument.
- **axes** –
  - If it is an integer, then axes axes at the last of a and the first of b are used.
  - If it is a pair of sequences of integers, then these two sequences specify the list of axes for a and b. The corresponding axes are paired for sum-product.
- **out** (cupy.ndarray) – Output array.

Returns The tensor dot product of a and b along the axes specified by axes.

Return type *cupy.ndarray*

See also:

numpy.tensordot()

Decompositions

cupy.linalg.**cholesky**(*a*)

Cholesky decomposition.

Decompose a given two-dimensional square matrix into L * L.T, where L is a lower-triangular matrix and .T is a conjugate transpose operator. Note that in the current implementation a must be a real matrix, and only float32 and float64 are supported.

Parameters **a** (cupy.ndarray) – The input matrix with dimension (N, N)

See also:

numpy.linalg.cholesky()

Norms etc.

cupy.**trace**(*a*, *offset=0*, *axis1=0*, *axis2=1*, *dtype=None*, *out=None*)

Returns the sum along the diagonals of an array.

It computes the sum along the diagonals at axis1 and axis2.

Parameters

- **a** (cupy.ndarray) – Array to take trace.
- **offset** (*int*) – Index of diagonals. Zero indicates the main diagonal, a positive value an upper diagonal, and a negative value a lower diagonal.
- **axis1** (*int*) – The first axis along which the trace is taken.
- **axis2** (*int*) – The second axis along which the trace is taken.
- **dtype** – Data type specifier of the output.
- **out** (cupy.ndarray) – Output array.

Returns The trace of `a` along axes `(axis1, axis2)`.

Return type *cupy.ndarray*

See also:

`numpy.trace()`

Logic Functions

Infinities and NaNs

cupy.**isfinite = <ufunc 'cupy_isfinite'>**
Tests finiteness elementwise.

Each element of returned array is `True` only if the corresponding element of the input is finite (i.e. not an infinity nor NaN).

See also:

`numpy.isfinite`

cupy.**isinf = <ufunc 'cupy_isinf'>**
Tests if each element is the positive or negative infinity.

See also:

`numpy.isinf`

cupy.**isnan = <ufunc 'cupy_isnan'>**
Tests if each element is a NaN.

See also:

`numpy.isnan`

Logic operations

cupy.**logical_and = <ufunc 'cupy_logical_and'>**
Computes the logical AND of two arrays.

See also:

`numpy.logical_and`

cupy.**logical_or = <ufunc 'cupy_logical_or'>**
Computes the logical OR of two arrays.

See also:

`numpy.logical_or`

cupy.**logical_not** = <ufunc 'cupy_logical_not'>
Computes the logical NOT of an array.

See also:

numpy.logical_not

cupy.**logical_xor** = <ufunc 'cupy_logical_xor'>
Computes the logical XOR of two arrays.

See also:

numpy.logical_xor

Comparison operations

cupy.**greater** = <ufunc 'cupy_greater'>
Tests elementwise if `x1 > x2`.

See also:

numpy.greater

cupy.**greater_equal** = <ufunc 'cupy_greater_equal'>
Tests elementwise if `x1 >= x2`.

See also:

numpy.greater_equal

cupy.**less** = <ufunc 'cupy_less'>
Tests elementwise if `x1 < x2`.

See also:

numpy.less

cupy.**less_equal** = <ufunc 'cupy_less_equal'>
Tests elementwise if `x1 <= x2`.

See also:

numpy.less_equal

cupy.**equal** = <ufunc 'cupy_equal'>
Tests elementwise if `x1 == x2`.

See also:

numpy.equal

cupy.**not_equal** = <ufunc 'cupy_not_equal'>
Tests elementwise if `x1 != x2`.

See also:

numpy.equal

Mathematical Functions

Trigonometric functions

cupy.**sin = <ufunc 'cupy_sin'>**
Elementwise sine function.

See also:

numpy.sin

cupy.**cos = <ufunc 'cupy_cos'>**
Elementwise cosine function.

See also:

numpy.cos

cupy.**tan = <ufunc 'cupy_tan'>**
Elementwise tangent function.

See also:

numpy.tan

cupy.**arcsin = <ufunc 'cupy_arcsin'>**
Elementwise inverse-sine function (a.k.a. arcsine function).

See also:

numpy.arcsin

cupy.**arccos = <ufunc 'cupy_arccos'>**
Elementwise inverse-cosine function (a.k.a. arccosine function).

See also:

numpy.arccos

cupy.**arctan = <ufunc 'cupy_arctan'>**
Elementwise inverse-tangent function (a.k.a. arctangent function).

See also:

numpy.arctan

cupy.**hypot = <ufunc 'cupy_hypot'>**
Computes the hypoteneous of orthogonal vectors of given length.

This is equivalent to `sqrt(x1 **2 + x2 ** 2)`, while this function is more efficient.

See also:

numpy.hypot

cupy.**arctan2 = <ufunc 'cupy_arctan2'>**
Elementwise inverse-tangent of the ratio of two arrays.

See also:

numpy.arctan2

cupy.**deg2rad = <ufunc 'cupy_deg2rad'>**
Converts angles from degrees to radians elementwise.

See also:

numpy.deg2rad, numpy.radians

cupy.**rad2deg = <ufunc 'cupy_rad2deg'>**
Converts angles from radians to degrees elementwise.

See also:

numpy.rad2deg, numpy.degrees

cupy.**degrees = <ufunc 'cupy_rad2deg'>**
Converts angles from radians to degrees elementwise.

See also:

numpy.rad2deg, numpy.degrees

cupy.**radians = <ufunc 'cupy_deg2rad'>**
Converts angles from degrees to radians elementwise.

See also:

numpy.deg2rad, numpy.radians

Hyperbolic functions

cupy.**sinh = <ufunc 'cupy_sinh'>**
Elementwise hyperbolic sine function.

See also:

numpy.sinh

cupy.**cosh = <ufunc 'cupy_cosh'>**
Elementwise hyperbolic cosine function.

See also:

numpy.cosh

cupy.**tanh = <ufunc 'cupy_tanh'>**
Elementwise hyperbolic tangent function.

See also:

numpy.tanh

cupy.**arcsinh = <ufunc 'cupy_arcsinh'>**
Elementwise inverse of hyperbolic sine function.

See also:

numpy.arcsinh

cupy.**arccosh = <ufunc 'cupy_arccosh'>**
Elementwise inverse of hyperbolic cosine function.

See also:

numpy.arccosh

cupy.**arctanh = <ufunc 'cupy_arctanh'>**
Elementwise inverse of hyperbolic tangent function.

See also:

numpy.arctanh

Rounding

cupy.**rint** = <ufunc 'cupy_rint'>
Rounds each element of an array to the nearest integer.

See also:

numpy.rint

cupy.**floor** = <ufunc 'cupy_floor'>
Rounds each element of an array to its floor integer.

See also:

numpy.floor

cupy.**ceil** = <ufunc 'cupy_ceil'>
Rounds each element of an array to its ceiling integer.

See also:

numpy.ceil

cupy.**trunc** = <ufunc 'cupy_trunc'>
Rounds each element of an array towards zero.

See also:

numpy.trunc

Sums and products

cupy.**sum**(**args*, ***kwargs*)

cupy.**prod**(**args*, ***kwargs*)

Exponential and logarithm functions

cupy.**exp** = <ufunc 'cupy_exp'>
Elementwise exponential function.

See also:

numpy.exp

cupy.**expm1** = <ufunc 'cupy_expm1'>
Computes `exp(x) - 1` elementwise.

See also:

numpy.expm1

cupy.**exp2** = <ufunc 'cupy_exp2'>
Elementwise exponentiation with base 2.

See also:

numpy.exp2

cupy.**log** = <ufunc 'cupy_log'>
Elementwise natural logarithm function.

See also:

numpy.log

cupy.**log10** = <ufunc 'cupy_log10'>
Elementwise common logarithm function.

See also:

numpy.log10

cupy.**log2** = <ufunc 'cupy_log2'>
Elementwise binary logarithm function.

See also:

numpy.log2

cupy.**log1p** = <ufunc 'cupy_log1p'>
Computes `log(1 + x)` elementwise.

See also:

numpy.log1p

cupy.**logaddexp** = <ufunc 'cupy_logaddexp'>
Computes `log(exp(x1) + exp(x2))` elementwise.

See also:

numpy.logaddexp

cupy.**logaddexp2** = <ufunc 'cupy_logaddexp2'>
Computes `log2(exp2(x1) + exp2(x2))` elementwise.

See also:

numpy.logaddexp2

Floating point manipulations

cupy.**signbit** = <ufunc 'cupy_signbit'>
Tests elementwise if the sign bit is set (i.e. less than zero).

See also:

numpy.signbit

cupy.**copysign** = <ufunc 'cupy_copysign'>
Returns the first argument with the sign bit of the second elementwise.

See also:

numpy.copysign

cupy.**ldexp** = <ufunc 'cupy_ldexp'>
Computes `x1 * 2 ** x2` elementwise.

See also:

numpy.ldexp

cupy.**frexp** = <ufunc 'cupy_frexp'>
Decomposes each element to mantissa and two's exponent.

This ufunc outputs two arrays of the input dtype and the `int` dtype.

See also:

numpy.frexp

cupy.**nextafter** = <ufunc 'cupy_nextafter'>
Computes the nearest neighbor float values towards the second argument.

See also:

numpy.nextafter

Arithmetic operations

cupy.**negative** = <ufunc 'cupy_negative'>
Takes numerical negative elementwise.

See also:

numpy.negative

cupy.**add** = <ufunc 'cupy_add'>
Adds two arrays elementwise.

See also:

numpy.add

cupy.**subtract** = <ufunc 'cupy_subtract'>
Subtracts arguments elementwise.

See also:

numpy.subtract

cupy.**multiply** = <ufunc 'cupy_multiply'>
Multiplies two arrays elementwise.

See also:

numpy.multiply

cupy.**divide** = <ufunc 'cupy_divide'>
Divides arguments elementwise.

See also:

numpy.divide

cupy.**true_divide** = <ufunc 'cupy_true_divide'>
Elementwise true division (i.e. division as floating values).

See also:

numpy.true_divide

cupy.**floor_divide** = <ufunc 'cupy_floor_divide'>
Elementwise floor division (i.e. integer quotient).

See also:

numpy.floor_divide

cupy.**power** = <ufunc 'cupy_power'>
Computes `x1 ** x2` elementwise.

See also:

numpy.power

cupy.**fmod = <ufunc 'cupy_fmod'>**
Computes the remainder of C division elementwise.

See also:

numpy.fmod

cupy.**mod = <ufunc 'cupy_remainder'>**
Computes the remainder of Python division elementwise.

See also:

numpy.remainder

cupy.**remainder = <ufunc 'cupy_remainder'>**
Computes the remainder of Python division elementwise.

See also:

numpy.remainder

cupy.**modf = <ufunc 'cupy_modf'>**
Extracts the fractional and integral parts of an array elementwise.

This ufunc returns two arrays.

See also:

numpy.modf

cupy.**reciprocal = <ufunc 'cupy_reciprocal'>**
Computes `1 / x` elementwise.

See also:

numpy.reciprocal

Miscellaneous

cupy.**clip**(**args*, ***kwargs*)

cupy.**sqrt = <ufunc 'cupy_sqrt'>**

cupy.**square = <ufunc 'cupy_square'>**
Elementwise square function.

See also:

numpy.square

cupy.**absolute = <ufunc 'cupy_absolute'>**
Elementwise absolute value function.

See also:

numpy.absolute

cupy.**sign = <ufunc 'cupy_sign'>**
Elementwise sign function.

It returns -1, 0, or 1 depending on the sign of the input.

See also:

numpy.sign

cupy.**maximum** = <ufunc 'cupy_maximum'>
Takes the maximum of two arrays elementwise.

If NaN appears, it returns the NaN.

See also:

numpy.maximum

cupy.**minimum** = <ufunc 'cupy_minimum'>
Takes the minimum of two arrays elementwise.

If NaN appears, it returns the NaN.

See also:

numpy.minimum

cupy.**fmax** = <ufunc 'cupy_fmax'>
Takes the maximum of two arrays elementwise.

If NaN appears, it returns the other operand.

See also:

numpy.fmax

cupy.**fmin** = <ufunc 'cupy_fmin'>
Takes the minimum of two arrays elementwise.

If NaN appears, it returns the other operand.

See also:

numpy.fmin

Random Sampling (cupy.random)

CuPy's random number generation routines are based on cuRAND. They cover a small fraction of numpy.random.

The big difference of *cupy.random* from numpy.random is that *cupy.random* supports dtype option for most functions. This option enables us to generate float32 values directly without any space overhead.

Sample random data

cupy.random.**choice**(*a*, *size=None*, *replace=True*, *p=None*)
Returns an array of random values from a given 1-D array.

Each element of the returned array is independently sampled from a according to p or uniformly.

Parameters

- **a** (*1-D array-like or int*) – If an array-like, a random sample is generated from its elements. If an int, the random sample is generated as if a was cupy.arange(n)
- **size** (*int or tuple of ints*) – The shape of the array.
- **replace** (*boolean*) – Whether the sample is with or without replacement
- **p** (*1-D array-like*) – The probabilities associated with each entry in a. If not given the sample assumes a uniform distribution over all entries in a.

Returns

An array of a values distributed according to p or uniformly.

Return type *cupy.ndarray*

See also:

numpy.random.choice()

cupy.random.**rand**(**size*, ***kwarg*)

Returns an array of uniform random values over the interval [0, 1).

Each element of the array is uniformly distributed on the half-open interval [0, 1). All elements are identically and independently distributed (i.i.d.).

Parameters

- **size** (*tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only numpy.float32 and numpy.float64 types are allowed. The default is numpy.float64.

Returns A random array.

Return type *cupy.ndarray*

See also:

numpy.random.rand()

cupy.random.**randn**(**size*, ***kwarg*)

Returns an array of standard normal random values.

Each element of the array is normally distributed with zero mean and unit variance. All elements are identically and independently distributed (i.i.d.).

Parameters

- **size** (*tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only numpy.float32 and numpy.float64 types are allowed. The default is numpy.float64.

Returns An array of standard normal random values.

Return type *cupy.ndarray*

See also:

numpy.random.randn()

cupy.random.**randint**(*low*, *high=None*, *size=None*)

Returns a scalar or an array of integer values over [low, high).

Each element of returned values are independently sampled from uniform distribution over left-close and right-open interval [low, high).

Parameters

- **low** (*int*) – If high is not None, it is the lower bound of the interval. Otherwise, it is the **upper** bound of the interval and lower bound of the interval is set to 0.
- **high** (*int*) – Upper bound of the interval.
- **size** (*None or int or tuple of ints*) – The shape of returned value.

Returns If size is None, it is single integer sampled. If size is integer, it is the 1D-array of length size element. Otherwise, it is the array whose shape specified by size.

Return type int or cupy.ndarray of ints

cupy.random.**random_integers**(*low*, *high=None*, *size=None*)
Return a scalar or an array of integer values over `[low, high]`

Each element of returned values are independently sampled from uniform distribution over closed interval `[low, high]`.

Parameters

- **low** (*int*) – If `high` is not `None`, it is the lower bound of the interval. Otherwise, it is the **upper** bound of the interval and the lower bound is set to `1`.
- **high** (*int*) – Upper bound of the interval.
- **size** (*None or int or tuple of ints*) – The shape of returned value.

Returns If size is `None`, it is single integer sampled. If size is integer, it is the 1D-array of length `size` element. Otherwise, it is the array whose shape specified by `size`.

Return type int or cupy.ndarray of ints

cupy.random.**random_sample**(*size=None*, *dtype=<type 'float'>*)
Returns an array of random values over the interval `[0, 1)`.

This is a variant of *cupy.random.rand()*.

Parameters

- **size** (*int or tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns An array of uniformly distributed random values.

Return type *cupy.ndarray*

See also:

`numpy.random.random_sample()`

cupy.random.**random**(*size=None*, *dtype=<type 'float'>*)
Returns an array of random values over the interval `[0, 1)`.

This is a variant of *cupy.random.rand()*.

Parameters

- **size** (*int or tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns An array of uniformly distributed random values.

Return type *cupy.ndarray*

See also:

`numpy.random.random_sample()`

cupy.random.**ranf**(*size=None*, *dtype=<type 'float'>*)
Returns an array of random values over the interval `[0, 1)`.

This is a variant of *cupy.random.rand()*.

Parameters

- **size** (*int or tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns An array of uniformly distributed random values.

Return type *cupy.ndarray*

See also:

`numpy.random.random_sample()`

`cupy.random.`**sample**(*size=None*, *dtype=<type 'float'>*)

Returns an array of random values over the interval `[0, 1)`.

This is a variant of *`cupy.random.rand()`*.

Parameters

- **size** (*int or tuple of ints*) – The shape of the array.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns An array of uniformly distributed random values.

Return type *cupy.ndarray*

See also:

`numpy.random.random_sample()`

Distributions

`cupy.random.`**gumbel**(*loc=0.0*, *scale=1.0*, *size=None*, *dtype=<type 'float'>*)

Returns an array of samples drawn from a Gumbel distribution.

The samples are drawn from a Gumbel distribution with location `loc` and scale `scale`. Its probability density function is defined as

$$f(x) = \frac{1}{\eta} \exp\left\{-\frac{x-\mu}{\eta}\right\} \exp\left[-\exp\left\{-\frac{x-\mu}{\eta}\right\}\right],$$

where μ is `loc` and η is `scale`.

Parameters

- **loc** (*float*) – The location of the mode μ.
- **scale** (*float*) – The scale parameter η.
- **size** (*int or tuple of ints*) – The shape of the array. If `None`, a zero-dimensional array is generated.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns Samples drawn from the Gumbel destribution.

Return type *cupy.ndarray*

See also:

`numpy.random.gumbel()`

cupy.random.**lognormal**(*mean=0.0*, *sigma=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of samples drawn from a log normal distribution.

The samples are natural log of samples drawn from a normal distribution with mean `mean` and deviation `sigma`.

Parameters

- **mean** (*float*) – Mean of the normal distribution.
- **sigma** (*float*) – Standard deviation of the normal distribution.
- **size** (*int or tuple of ints*) – The shape of the array. If `None`, a zero-dimensional array is generated.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns Samples drawn from the log normal distribution.

Return type *cupy.ndarray*

See also:

`numpy.random.lognormal()`

cupy.random.**normal**(*loc=0.0*, *scale=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of normally distributed samples.

Parameters

- **loc** (*float or array_like of floats*) – Mean of the normal distribution.
- **scale** (*float or array_like of floats*) – Standard deviation of the normal distribution.
- **size** (*int or tuple of ints*) – The shape of the array. If `None`, a zero-dimensional array is generated.
- **dtype** – Data type specifier. Only `numpy.float32` and `numpy.float64` types are allowed.

Returns Normally distributed samples.

Return type *cupy.ndarray*

See also:

`numpy.random.normal()`

cupy.random.**standard_normal**(*size=None*, *dtype=<type 'float'>*)
Returns an array of samples drawn from the standard normal distribution.

This is a variant of *cupy.random.randn()*.

Parameters

- **size** (*int or tuple of ints*) – The shape of the array. If `None`, a zero-dimensional array is generated.
- **dtype** – Data type specifier.

Returns Samples drawn from the standard normal distribution.

Return type *cupy.ndarray*

See also:

`numpy.random.standard_normal()`

cupy.random.**uniform**(*low=0.0*, *high=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of uniformly-distributed samples over an interval.

Samples are drawn from a uniform distribution over the half-open interval `[low, high)`.

Parameters

- **low** (*float*) – Lower end of the interval.
- **high** (*float*) – Upper end of the interval.
- **size** (*int or tuple of ints*) – The shape of the array. If `None`, a zero-dimensional array is generated.
- **dtype** – Data type specifier.

Returns Samples drawn from the uniform distribution.

Return type *cupy.ndarray*

See also:

`numpy.random.uniform()`

Random number generator

cupy.random.**seed**(*seed=None*)
Resets the state of the random number generator with a seed.

This function resets the state of the global random number generator for the current device. Be careful that generators for other devices are not affected.

Parameters **seed** (*None or int*) – Seed for the random number generator. If `None`, it uses `os.urandom()` if available or `time.clock()` otherwise. Note that this function does not support seeding by an integer array.

cupy.random.**get_random_state**()
Gets the state of the random number generator for the current device.

If the state for the current device is not created yet, this function creates a new one, initializes it, and stores it as the state for the current device.

Returns The state of the random number generator for the device.

Return type *RandomState*

class cupy.random.**RandomState**(*seed=None*, *method=100*)
Portable container of a pseudo-random number generator.

An instance of this class holds the state of a random number generator. The state is available only on the device which has been current at the initialization of the instance.

Functions of *cupy.random* use global instances of this class. Different instances are used for different devices. The global state for the current device can be obtained by the *cupy.random.get_random_state()* function.

Parameters

- **seed** (*None or int*) – Seed of the random number generator. See the *seed()* method for detail.
- **method** (*int*) – Method of the random number generator. Following values are available:

```
cupy.cuda.curand.CURAND_RNG_PSEUDO_DEFAULT
cupy.cuda.curand.CURAND_RNG_XORWOW
cupy.cuda.curand.CURAND_RNG_MRG32K3A
cupy.cuda.curand.CURAND_RNG_MTGP32
cupy.cuda.curand.CURAND_RNG_MT19937
cupy.cuda.curand.CURAND_RNG_PHILOX4_32_10
```

choice(*a*, *size=None*, *replace=True*, *p=None*)
Returns an array of random values from a given 1-D array.

See also:

cupy.random.choice() for full document, numpy.random.choice()

interval(*mx*, *size*)
Generate multiple integers independently sampled uniformly from [0, mx].

Parameters

- **mx** (*int*) – Upper bound of the interval
- **size** (*None or int or tuple*) – Shape of the array or the scalar returned.

Returns If None, an *cupy.ndarray* with shape () is returned. If int, 1-D array of length size is returned. If tuple, multi-dimensional array with shape size is returned. Currently, each element of the array is numpy.int32.

Return type int or *cupy.ndarray*

lognormal(*mean=0.0*, *sigma=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of samples drawn from a log normal distribution.

See also:

cupy.random.lognormal() for full documentation, numpy.random.RandomState.lognormal()

normal(*loc=0.0*, *scale=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of normally distributed samples.

See also:

cupy.random.normal() for full documentation, numpy.random.RandomState.normal()

rand(**size*, ***kwarg*)
Returns uniform random values over the interval [0, 1).

See also:

cupy.random.rand() for full documentation, numpy.random.RandomState.rand()

randn(**size*, ***kwarg*)
Returns an array of standard normal random values.

See also:

cupy.random.randn() for full documentation, numpy.random.RandomState.randn()

random_sample(*size=None*, *dtype=<type 'float'>*)
Returns an array of random values over the interval [0, 1).

See also:

cupy.random.random_sample() for full documentation, numpy.random.RandomState.random_sample()

seed(*seed=None*)
Resets the state of the random number generator with a seed.

See also:

cupy.random.seed() for full documentation, numpy.random.RandomState.seed()

standard_normal(*size=None*, *dtype=<type 'float'>*)
Returns samples drawn from the standard normal distribution.

See also:

cupy.random.standard_normal() for full documentation, numpy.random.RandomState.standard_normal()

uniform(*low=0.0*, *high=1.0*, *size=None*, *dtype=<type 'float'>*)
Returns an array of uniformly-distributed samples over an interval.

See also:

cupy.random.uniform() for full documentation, numpy.random.RandomState.uniform()

Sorting, Searching, and Counting

cupy.**argmax**(*a*, *axis=None*, *dtype=None*, *out=None*, *keepdims=False*)
Returns the indices of the maximum along an axis.

Parameters

- **a** (cupy.ndarray) – Array to take argmax.
- **axis** (*int*) – Along which axis to find the maximum. a is flattened by default.
- **dtype** – Data type specifier.
- **out** (cupy.ndarray) – Output array.
- **keepdims** (*bool*) – If True, the axis axis is preserved as an axis of length one.

Returns The indices of the maximum of a along an axis.

Return type *cupy.ndarray*

See also:

numpy.argmax()

cupy.**argmin**(*a*, *axis=None*, *dtype=None*, *out=None*, *keepdims=False*)
Returns the indices of the minimum along an axis.

Parameters

- **a** (cupy.ndarray) – Array to take argmin.
- **axis** (*int*) – Along which axis to find the minimum. a is flattened by default.
- **dtype** – Data type specifier.
- **out** (cupy.ndarray) – Output array.
- **keepdims** (*bool*) – If True, the axis axis is preserved as an axis of length one.

Returns The indices of the minimum of a along an axis.

Return type *cupy.ndarray*

See also:

numpy.argmin()

cupy.**count_nonzero**(*a*, *axis=None*)
Counts the number of non-zero values in the array.

Parameters

- **a** (cupy.ndarray) – The array for which to count non-zeros.
- **axis** (*int or tuple, optional*) – Axis or tuple of axes along which to count non-zeros. Default is None, meaning that non-zeros will be counted along a flattened version of a

Returns

Number of non-zero values in the array along a given axis. Otherwise, the total number of non-zero values in the array is returned.

Return type int or cupy.ndarray of int

cupy.**nonzero**(*a*)
Return the indices of the elements that are non-zero.

Returns a tuple of arrays, one for each dimension of a, containing the indices of the non-zero elements in that dimension.

Parameters **a** (cupy.ndarray) – array

Returns Indices of elements that are non-zero.

Return type tuple of arrays

See also:

numpy.nonzero()

cupy.**flatnonzero**(*a*)
Return indices that are non-zero in the flattened version of a.

This is equivalent to a.ravel().nonzero()[0].

Parameters **a** (cupy.ndarray) – input array

Returns Output array, containing the indices of the elements of a.ravel() that are non-zero.

Return type *cupy.ndarray*

See also:

numpy.flatnonzero()

cupy.**where**(**args*, ***kwargs*)

Statistics

Order statistics

cupy.**amin**(**args*, ***kwargs*)

cupy.**amax**(**args*, ***kwargs*)

cupy.**nanmin**(*a*, *axis=None*, *out=None*, *keepdims=False*)
Returns the minimum of an array along an axis ignoring NaN.

When there is a slice whose elements are all NaN, a RuntimeWarning is raised and NaN is returned.

Parameters

- **a** (cupy.ndarray) – Array to take the minimum.
- **axis** (*int*) – Along which axis to take the minimum. The flattened array is used by default.
- **out** (cupy.ndarray) – Output array.
- **keepdims** (*bool*) – If True, the axis is remained as an axis of size one.

Returns The minimum of a, along the axis if specified.

Return type *cupy.ndarray*

See also:

numpy.nanmin()

cupy.**nanmax**(*a*, *axis=None*, *out=None*, *keepdims=False*)
Returns the maximum of an array along an axis ignoring NaN.

When there is a slice whose elements are all NaN, a RuntimeWarning is raised and NaN is returned.

Parameters

- **a** (cupy.ndarray) – Array to take the maximum.
- **axis** (*int*) – Along which axis to take the maximum. The flattened array is used by default.
- **out** (cupy.ndarray) – Output array.
- **keepdims** (*bool*) – If True, the axis is remained as an axis of size one.

Returns The maximum of a, along the axis if specified.

Return type *cupy.ndarray*

See also:

numpy.nanmax()

Means and variances

cupy.**mean**(*a*, *axis=None*, *dtype=None*, *out=None*, *keepdims=False*)
Returns the arithmetic mean along an axis.

Parameters

- **a** (cupy.ndarray) – Array to compute mean.
- **axis** (*int*) – Along which axis to compute mean. The flattened array is used by default.
- **dtype** – Data type specifier.
- **out** (cupy.ndarray) – Output array.
- **keepdims** (*bool*) – If True, the axis is remained as an axis of size one.

Returns The mean of the input array along the axis.

Return type *cupy.ndarray*

See also:

`numpy.mean()`

cupy.**var**(*a*, *axis=None*, *dtype=None*, *out=None*, *ddof=0*, *keepdims=False*)
Returns the variance along an axis.

Parameters

- **a** (`cupy.ndarray`) – Array to compute variance.
- **axis** (*int*) – Along which axis to compute variance. The flattened array is used by default.
- **dtype** – Data type specifier.
- **out** (`cupy.ndarray`) – Output array.
- **keepdims** (*bool*) – If `True`, the axis is remained as an axis of size one.

Returns The variance of the input array along the axis.

Return type *cupy.ndarray*

See also:

`numpy.var()`

cupy.**std**(*a*, *axis=None*, *dtype=None*, *out=None*, *ddof=0*, *keepdims=False*)
Returns the standard deviation along an axis.

Parameters

- **a** (`cupy.ndarray`) – Array to compute standard deviation.
- **axis** (*int*) – Along which axis to compute standard deviation. The flattened array is used by default.
- **dtype** – Data type specifier.
- **out** (`cupy.ndarray`) – Output array.
- **keepdims** (*bool*) – If `True`, the axis is remained as an axis of size one.

Returns The standard deviation of the input array along the axis.

Return type *cupy.ndarray*

See also:

`numpy.std()`

Histograms

cupy.**bincount**(*x*, *weights=None*, *minlength=None*)
Count number of occurrences of each value in array of non-negative ints.

Parameters

- **x** (`cupy.ndarray`) – Input array.
- **weights** (`cupy.ndarray`) – Weights array which has the same shape as `x`.
- **minlength** (*int*) – A minimum number of bins for the output array.

Returns

The result of binning the input array. The length of output is equal to `max(cupy.max(x) + 1, minlength)`.

Return type *cupy.ndarray*

See also:

`numpy.bincount()`

External Functions

`cupy.`**`scatter_add`**(*a*, *slices*, *value*)

Adds given values to specified elements of an array.

It adds `value` to the specified elements of `a`. If all of the indices target different locations, the operation of `scatter_add()` is equivalent to `a[slices] = a[slices] + value`. If there are multiple elements targeting the same location, `scatter_add()` uses all of these values for addition. On the other hand, `a[slices] = a[slices] + value` only adds the contribution from one of the indices targeting the same location.

Note that just like an array indexing, negative indices are interpreted as counting from the end of an array.

Also note that `scatter_add()` behaves identically to `numpy.add.at()`.

Example

```
>>> import numpy
>>> import cupy
>>> a = cupy.zeros((6,), dtype=numpy.float32)
>>> i = cupy.array([1, 0, 1])
>>> v = cupy.array([1., 1., 1.])
>>> cupy.scatter_add(a, i, v);
>>> a
array([ 1.,  2.,  0.,  0.,  0.,  0.], dtype=float32)
```

Parameters

- **`a`** (`ndarray`) – An array that gets added.
- **`slices`** – It is integer, slices, ellipsis, numpy.newaxis, integer array-like, boolean array-like or tuple of them. It works for slices used for `cupy.ndarray.__getitem__()` and `cupy.ndarray.__setitem__()`.
- **`v`** (*`array-like`*) – Values to increment `a` at referenced locations.

Note: It only supports types that are supported by CUDA's atomicAdd when an integer array is included in `slices`. The supported types are `numpy.float32`, `numpy.int32`, `numpy.uint32`, `numpy.uint64` and `numpy.ulonglong`.

Note: `scatter_add()` does not raise an error when indices exceed size of axes. Instead, it wraps indices.

See also:

`numpy.add.at()`.

NumPy-CuPy Generic Code Support

cupy.**get_array_module**(**args*)
Returns the array module for arguments.

This function is used to implement CPU/GPU generic code. If at least one of the arguments is a *cupy.ndarray* object, the *cupy* module is returned.

Parameters **args** – Values to determine whether NumPy or CuPy should be used.

Returns *cupy* or numpy is returned based on the types of the arguments.

Return type module

Example

A NumPy/CuPy generic function can be written as follows

```
>>> def softplus(x):
...     xp = cupy.get_array_module(x)
...     return xp.maximum(0, x) + xp.log1p(xp.exp(-abs(x)))
```

Low-Level CUDA Support

Device management

class cupy.cuda.**Device**
Object that represents a CUDA device.

This class provides some basic manipulations on CUDA devices.

It supports the context protocol. For example, the following code is an example of temporarily switching the current device:

```
with Device(0):
    do_something_on_device_0()
```

After the *with* statement gets done, the current device is reset to the original one.

Parameters **device** (*int or cupy.cuda.Device*) – Index of the device to manipulate. Be careful that the device ID (a.k.a. GPU ID) is zero origin. If it is a Device object, then its ID is used. The current device is selected by default.

Variables **id** (*int*) – ID of this device.

compute_capability
Compute capability of this device.

The capability is represented by a string containing the major index and the minor index. For example, compute capability 3.5 is represented by the string '35'.

cublas_handle
The cuBLAS handle for this device.

The same handle is used for the same device even if the Device instance itself is different.

cusolver_handle
The cuSOLVER handle for this device.

The same handle is used for the same device even if the Device instance itself is different.

synchronize()
Synchronizes the current thread to the device.

use()
Makes this device current.

If you want to switch a device temporarily, use the *with* statement.

Memory management

class cupy.cuda.**Memory**
Memory allocation on a CUDA device.

This class provides an RAII interface of the CUDA memory allocation.

Parameters

- **device** (cupy.cuda.Device) – Device whose memory the pointer refers to.
- **size** (*int*) – Size of the memory allocation in bytes.

class cupy.cuda.**MemoryPointer**
Pointer to a point on a device memory.

An instance of this class holds a reference to the original memory buffer and a pointer to a place within this buffer.

Parameters

- **mem** (Memory) – The device memory buffer.
- **offset** (*int*) – An offset from the head of the buffer to the place this pointer refers.

Variables

- ***device*** (cupy.cuda.Device) – Device whose memory the pointer refers to.
- **mem** (Memory) – The device memory buffer.
- **ptr** (*int*) – Pointer to the place within the buffer.

copy_from()
Copies a memory sequence from a (possibly different) device or host.

This function is a useful interface that selects appropriate one from *copy_from_device()* and *copy_from_host()*.

Parameters

- **mem** (*ctypes.c_void_p or* cupy.cuda.MemoryPointer) – Source memory pointer.
- **size** (*int*) – Size of the sequence in bytes.

copy_from_async()
Copies a memory sequence from an arbitrary place asynchronously.

This function is a useful interface that selects appropriate one from *copy_from_device_async()* and *copy_from_host_async()*.

Parameters

- **mem** (*ctypes.c_void_p or cupy.cuda.MemoryPointer*) – Source memory pointer.
- **size** (*int*) – Size of the sequence in bytes.
- **stream** (cupy.cuda.Stream) – CUDA stream.

copy_from_device()
Copies a memory sequence from a (possibly different) device.

Parameters

- **src** (cupy.cuda.MemoryPointer) – Source memory pointer.
- **size** (*int*) – Size of the sequence in bytes.

copy_from_device_async()
Copies a memory from a (possibly different) device asynchronously.

Parameters

- **src** (cupy.cuda.MemoryPointer) – Source memory pointer.
- **size** (*int*) – Size of the sequence in bytes.
- **stream** (cupy.cuda.Stream) – CUDA stream.

copy_from_host()
Copies a memory sequence from the host memory.

Parameters

- **mem** (*ctypes.c_void_p*) – Source memory pointer.
- **size** (*int*) – Size of the sequence in bytes.

copy_from_host_async()
Copies a memory sequence from the host memory asynchronously.

Parameters

- **mem** (*ctypes.c_void_p*) – Source memory pointer. It must be a pinned memory.
- **size** (*int*) – Size of the sequence in bytes.
- **stream** (cupy.cuda.Stream) – CUDA stream.

copy_to_host()
Copies a memory sequence to the host memory.

Parameters

- **mem** (*ctypes.c_void_p*) – Target memory pointer.
- **size** (*int*) – Size of the sequence in bytes.

copy_to_host_async()
Copies a memory sequence to the host memory asynchronously.

Parameters

- **mem** (*ctypes.c_void_p*) – Target memory pointer. It must be a pinned memory.
- **size** (*int*) – Size of the sequence in bytes.
- **stream** (cupy.cuda.Stream) – CUDA stream.

memset()
Fills a memory sequence by constant byte value.

Parameters

- **value** (*int*) – Value to fill.
- **size** (*int*) – Size of the sequence in bytes.

memset_async()
Fills a memory sequence by constant byte value asynchronously.

Parameters

- **value** (*int*) – Value to fill.
- **size** (*int*) – Size of the sequence in bytes.
- **stream** (cupy.cuda.Stream) – CUDA stream.

cupy.cuda.**alloc**()
Calls the current allocator.

Use *set_allocator()* to change the current allocator.

Parameters **size** (*int*) – Size of the memory allocation.

Returns Pointer to the allocated buffer.

Return type *MemoryPointer*

cupy.cuda.**set_allocator**()
Sets the current allocator.

Parameters **allocator** (function) – CuPy memory allocator. It must have the same interface as the *cupy.cuda.alloc()* function, which takes the buffer size as an argument and returns the device buffer of that size.

class cupy.cuda.**MemoryPool**
Memory pool for all devices on the machine.

A memory pool preserves any allocations even if they are freed by the user. Freed memory buffers are held by the memory pool as *free blocks*, and they are reused for further memory allocations of the same sizes. The allocated blocks are managed for each device, so one instance of this class can be used for multiple devices.

Note: When the allocation is skipped by reusing the pre-allocated block, it does not call cudaMalloc and therefore CPU-GPU synchronization does not occur. It makes interleaves of memory allocations and kernel invocations very fast.

Note: The memory pool holds allocated blocks without freeing as much as possible. It makes the program hold most of the device memory, which may make other CUDA programs running in parallel out-of-memory situation.

Parameters **allocator** (function) – The base CuPy memory allocator. It is used for allocating new blocks when the blocks of the required size are all in use.

free_all_blocks()
Release free blocks.

free_all_free()
Release free blocks.

malloc()
Allocates the memory, from the pool if possible.

This method can be used as a CuPy memory allocator. The simplest way to use a memory pool as the default allocator is the following code:

```
set_allocator(MemoryPool().malloc)
```

Parameters **size** (*int*) – Size of the memory buffer to allocate in bytes.

Returns Pointer to the allocated buffer.

Return type *MemoryPointer*

n_free_blocks()
Count the total number of free blocks.

Returns The total number of free blocks.

Return type int

Streams and events

class cupy.cuda.**Stream**(*null=False*, *non_blocking=False*)
CUDA stream.

This class handles the CUDA stream handle in RAII way, i.e., when an Stream instance is destroyed by the GC, its handle is also destroyed.

Parameters

- **null** (*bool*) – If True, the stream is a null stream (i.e. the default stream that synchronizes with all streams). Otherwise, a plain new stream is created.
- **non_blocking** (*bool*) – If True, the stream does not synchronize with the NULL stream.

Variables **ptr** (*cupy.cuda.runtime.Stream*) – Raw stream handle. It can be passed to the CUDA Runtime API via ctypes.

add_callback(*callback*, *arg*)
Adds a callback that is called when all queued work is done.

Parameters

- **callback** (function) – Callback function. It must take three arguments (Stream object, int error status, and user data object), and returns nothing.
- **arg** (*object*) – Argument to the callback.

done
True if all work on this stream has been done.

record(*event=None*)
Records an event on the stream.

Parameters **event** (*None or cupy.cuda.Event*) – CUDA event. If None, then a new plain event is created and used.

Returns The recorded event.

Return type *cupy.cuda.Event*

See also:

cupy.cuda.Event.record()

synchronize()
Waits for the stream completing all queued work.

wait_event(*event*)
Makes the stream wait for an event.

The future work on this stream will be done after the event.

Parameters **event** (cupy.cuda.Event) – CUDA event.

class cupy.cuda.**Event**(*block=False*, *disable_timing=False*, *interprocess=False*)
CUDA event, a synchronization point of CUDA streams.

This class handles the CUDA event handle in RAII way, i.e., when an Event instance is destroyed by the GC, its handle is also destroyed.

Parameters

- **block** (*bool*) – If True, the event blocks on the *synchronize()* method.
- **disable_timing** (*bool*) – If True, the event does not prepare the timing data.
- **interprocess** (*bool*) – If True, the event can be passed to other processes.

Variables **ptr** (*cupy.cuda.runtime.Stream*) – Raw stream handle. It can be passed to the CUDA Runtime API via ctypes.

done
True if the event is done.

record(*stream=None*)
Records the event to a stream.

Parameters **stream** (cupy.cuda.Stream) – CUDA stream to record event. The null stream is used by default.

See also:

cupy.cuda.Stream.record()

synchronize()
Synchronizes all device work to the event.

If the event is created as a blocking event, it also blocks the CPU thread until the event is done.

cupy.cuda.**get_elapsed_time**(*start_event*, *end_event*)
Gets the elapsed time between two events.

Parameters

- **start_event** (Event) – Earlier event.
- **end_event** (Event) – Later event.

Returns Elapsed time in milliseconds.

Return type float

Profiler

cupy.cuda.**profile**(**args*, ***kwds*)
Enable CUDA profiling during with statement.

This function enables profiling on entering a with statement, and disables profiling on leaving the statement.

```
>>> with cupy.cuda.profile():
...     # do something you want to measure
...     pass
```

cupy.cuda.profiler.**initialize**()
Initialize the CUDA profiler.

This function initialize the CUDA profiler. See the CUDA document for detail.

Parameters

- **config_file** (*str*) – Name of the configuration file.
- **output_file** (*str*) – Name of the coutput file.
- **output_mode** (*int*) – `cupy.cuda.profiler.cudaKeyValuePair` or `cupy.cuda.profiler.cudaCSV`.

cupy.cuda.profiler.**start**()
Enable profiling.

A user can enable CUDA profiling. When an error occurs, it raises an exception.

See the CUDA document for detail.

cupy.cuda.profiler.**stop**()
Disable profiling.

A user can disable CUDA profiling. When an error occurs, it raises an exception.

See the CUDA document for detail.

cupy.cuda.nvtx.**Mark**()
Marks an instantaneous event (marker) in the application.

Markes are used to describe events at a specific time during execution of the application.

Parameters

- **message** (*str*) – Name of a marker.
- **id_color** (*int*) – ID of color for a marker.

cupy.cuda.nvtx.**MarkC**()
Marks an instantaneous event (marker) in the application.

Markes are used to describe events at a specific time during execution of the application.

Parameters

- **message** (*str*) – Name of a marker.
- **color** (*uint32*) – Color code for a marker.

cupy.cuda.nvtx.**RangePush**()
Starts a nestead range.

Ranges are used to describe events over a time span during execution of the application. The duration of a range is defined by the corresponding pair of `RangePush*()` to `RangePop()` calls.

Parameters

- **message** (*str*) – Name of a range.
- **id_color** (*int*) – ID of color for a range.

cupy.cuda.nvtx.**RangePushC**()
Starts a nestead range.

Ranges are used to describe events over a time span during execution of the application. The duration of a range is defined by the corresponding pair of `RangePush*()` to `RangePop()` calls.

Parameters

- **message** (*str*) – Name of a range.
- **color** (*uint32*) – ARGB color for a range.

cupy.cuda.nvtx.**RangePop**()
Ends a nestead range.

Ranges are used to describe events over a time span during execution of the application. The duration of a range is defined by the corresponding pair of `RangePush*()` to `RangePop()` calls.

Kernel binary memoization

cupy.**memoize**()
Makes a function memoizing the result for each argument and device.

This decorator provides automatic memoization of the function result.

Parameters **for_each_device** (*bool*) – If `True`, it memoizes the results for each device. Otherwise, it memoizes the results only based on the arguments.

cupy.**clear_memo**()
Clears the memoized results for all functions decorated by memoize.

User-Defined Kernels

CuPy provides easy ways to define two types of CUDA kernels: elementwise kernels and reduction kernels. We first describe how to define and call elementwise kernels, and then describe how to define and call reduction kernels.

Basics of elementwise kernels

An elementwise kernel can be defined by the *ElementwiseKernel* class. The instance of this class defines a CUDA kernel which can be invoked by the `__call__` method of this instance.

A definition of an elementwise kernel consists of four parts: an input argument list, an output argument list, a loop body code, and the kernel name. For example, a kernel that computes a squared difference $f(x, y) = (x - y)^2$ is defined as follows:

```
>>> squared_diff = cupy.ElementwiseKernel(
...     'float32 x, float32 y',
...     'float32 z',
...     'z = (x - y) * (x - y)',
...     'squared_diff')
```

The argument lists consist of comma-separated argument definitions. Each argument definition consists of a *type specifier* and an *argument name*. Names of NumPy data types can be used as type specifiers.

Note: `n`, `i`, and names starting with an underscore _ are reserved for the internal use.

The above kernel can be called on either scalars or arrays with broadcasting:

```
>>> x = cupy.arange(10, dtype=np.float32).reshape(2, 5)
>>> y = cupy.arange(5, dtype=np.float32)
>>> squared_diff(x, y)
array([[  0.,   0.,   0.,   0.,   0.],
       [ 25.,  25.,  25.,  25.,  25.]], dtype=float32)
>>> squared_diff(x, 5)
array([[ 25.,  16.,   9.,   4.,   1.],
       [  0.,   1.,   4.,   9.,  16.]], dtype=float32)
```

Output arguments can be explicitly specified (next to the input arguments):

```
>>> z = cupy.empty((2, 5), dtype=np.float32)
>>> squared_diff(x, y, z)
array([[  0.,   0.,   0.,   0.,   0.],
       [ 25.,  25.,  25.,  25.,  25.]], dtype=float32)
```

Type-generic kernels

If a type specifier is one character, then it is treated as a **type placeholder**. It can be used to define a type-generic kernels. For example, the above `squared_diff` kernel can be made type-generic as follows:

```
>>> squared_diff_generic = cupy.ElementwiseKernel(
...     'T x, T y',
...     'T z',
...     'z = (x - y) * (x - y)',
...     'squared_diff_generic')
```

Type placeholders of a same character in the kernel definition indicate the same type. The actual type of these placeholders is determined by the actual argument type. The ElementwiseKernel class first checks the output arguments and then the input arguments to determine the actual type. If no output arguments are given on the kernel invocation, then only the input arguments are used to determine the type.

The type placeholder can be used in the loop body code:

```
>>> squared_diff_generic = cupy.ElementwiseKernel(
...     'T x, T y',
...     'T z',
...     '''
...         T diff = x - y;
...         z = diff * diff;
...     ''',
...     'squared_diff_generic')
```

More than one type placeholder can be used in a kernel definition. For example, the above kernel can be further made generic over multiple arguments:

```
>>> squared_diff_super_generic = cupy.ElementwiseKernel(
...     'X x, Y y',
```

```
...     'Z z',
...     'z = (x - y) * (x - y)',
...     'squared_diff_super_generic')
```

Note that this kernel requires the output argument explicitly specified, because the type `Z` cannot be automatically determined from the input arguments.

Raw argument specifiers

The ElementwiseKernel class does the indexing with broadcasting automatically, which is useful to define most elementwise computations. On the other hand, we sometimes want to write a kernel with manual indexing for some arguments. We can tell the ElementwiseKernel class to use manual indexing by adding the `raw` keyword preceding the type specifier.

We can use the special variable `i` and method `_ind.size()` for the manual indexing. `i` indicates the index within the loop. `_ind.size()` indicates total number of elements to apply the elementwise operation. Note that it represents the size **after** broadcast operation.

For example, a kernel that adds two vectors with reversing one of them can be written as follows:

```
>>> add_reverse = cupy.ElementwiseKernel(
...     'T x, raw T y', 'T z',
...     'z = x + y[_ind.size() - i - 1]',
...     'add_reverse')
```

(Note that this is an artificial example and you can write such operation just by `z = x + y[::-1]` without defining a new kernel). A raw argument can be used like an array. The indexing operator `y[_ind.size() - i - 1]` involves an indexing computation on `y`, so `y` can be arbitrarily shaped and strode.

Note that raw arguments are not involved in the broadcasting. If you want to mark all arguments as `raw`, you must specify the `size` argument on invocation, which defines the value of `_ind.size()`.

Reduction kernels

Reduction kernels can be defined by the *`ReductionKernel`* class. We can use it by defining four parts of the kernel code:

1. Identity value: This value is used for the initial value of reduction.
2. Mapping expression: It is used for the pre-processing of each element to be reduced.
3. Reduction expression: It is an operator to reduce the multiple mapped values. The special variables `a` and `b` are used for its operands.
4. Post mapping expression: It is used to transform the resulting reduced values. The special variable `a` is used as its input. Output should be written to the output parameter.

ReductionKernel class automatically inserts other code fragments that are required for an efficient and flexible reduction implementation.

For example, L2 norm along specified axes can be written as follows:

```
>>> l2norm_kernel = cupy.ReductionKernel(
...     'T x',  # input params
...     'T y',  # output params
...     'x * x',  # map
...     'a + b',  # reduce
```

```
...     'y = sqrt(a)',  # post-reduction map
...     '0',  # identity value
...     'l2norm'  # kernel name
... )
>>> x = cupy.arange(10, dtype='f').reshape(2, 5)
>>> l2norm_kernel(x, axis=1)
array([  5.47722578,  15.96871948], dtype=float32)
```

Note: `raw` specifier is restricted for usages that the axes to be reduced are put at the head of the shape. It means, if you want to use `raw` specifier for at least one argument, the `axis` argument must be `0` or a contiguous increasing sequence of integers starting from `0`, like `(0, 1)`, `(0, 1, 2)`, etc.

Reference

class `cupy.`**`ElementwiseKernel`**

User-defined elementwise kernel.

This class can be used to define an elementwise kernel with or without broadcasting.

The kernel is compiled at an invocation of the `__call__()` method, which is cached for each device. The compiled binary is also cached into a file under the `$HOME/.cupy/kernel_cache/` directory with a hashed file name. The cached binary is reused by other processes.

Parameters

- **`in_params`** (*`str`*) – Input argument list.
- **`out_params`** (*`str`*) – Output argument list.
- **`operation`** (*`str`*) – The body in the loop written in CUDA-C/C++.
- **`name`** (*`str`*) – Name of the kernel function. It should be set for readability of the performance profiling.
- **`reduce_dims`** (*`bool`*) – If `False`, the shapes of array arguments are kept within the kernel invocation. The shapes are reduced (i.e., the arrays are reshaped without copy to the minimum dimension) by default. It may make the kernel fast by reducing the index calculations.
- **`options`** (*`list`*) – Options passed to the `nvcc` command.
- **`preamble`** (*`str`*) – Fragment of the CUDA-C/C++ code that is inserted at the top of the cu file.
- **`loop_prep`** (*`str`*) – Fragment of the CUDA-C/C++ code that is inserted at the top of the kernel function definition and above the `for` loop.
- **`after_loop`** (*`str`*) – Fragment of the CUDA-C/C++ code that is inserted at the bottom of the kernel function definition.

class `cupy.`**`ReductionKernel`**

User-defined reduction kernel.

This class can be used to define a reduction kernel with or without broadcasting.

The kernel is compiled at an invocation of the `__call__()` method, which is cached for each device. The compiled binary is also cached into a file under the `$HOME/.cupy/kernel_cache/` directory with a hashed file name. The cached binary is reused by other processes.

Parameters

- **in_params** (*str*) – Input argument list.
- **out_params** (*str*) – Output argument list.
- **map_expr** (*str*) – Mapping expression for input values.
- **reduce_expr** (*str*) – Reduction expression.
- **post_map_expr** (*str*) – Mapping expression for reduced values.
- **identity** (*str*) – Identity value for starting the reduction.
- **name** (*str*) – Name of the kernel function. It should be set for readability of the performance profiling.
- **reduce_type** (*str*) – Type of values to be used for reduction. This type is used to store the special variables `a`.
- **reduce_dims** (*bool*) – If `True`, input arrays are reshaped without copy to smaller dimensions for efficiency.
- **preamble** (*str*) – Fragment of the CUDA-C/C++ code that is inserted at the top of the cu file.
- **options** (*tuple of str*) – Additional compilation options.

Testing Modules

CuPy offers testing utilities to support unit testing. They are under namespace *cupy.testing*.

Standard Assertions

The assertions have same names as NumPy's ones. The difference from NumPy is that they can accept both `numpy.ndarray` and *cupy.ndarray*.

`cupy.testing.`**assert_allclose**(*actual*, *desired*, *rtol=1e-07*, *atol=0*, *err_msg=''*, *verbose=True*)

Raises an AssertionError if objects are not equal up to desired tolerance.

Parameters

- **actual** (*numpy.ndarray or cupy.ndarray*) – The actual object to check.
- **desired** (*numpy.ndarray or cupy.ndarray*) – The desired, expected object.
- **rtol** (*float*) – Relative tolerance.
- **atol** (*float*) – Absolute tolerance.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.

See also:

`numpy.testing.assert_allclose()`

`cupy.testing.`**assert_array_almost_equal**(*x*, *y*, *decimal=6*, *err_msg=''*, *verbose=True*)

Raises an AssertionError if objects are not equal up to desired precision.

Parameters

- **x** (*numpy.ndarray or cupy.ndarray*) – The actual object to check.
- **y** (*numpy.ndarray or cupy.ndarray*) – The desired, expected object.
- **decimal** (*int*) – Desired precision.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.

See also:

`numpy.testing.assert_array_almost_equal()`

cupy.testing.**assert_array_almost_equal_nulp**(*x*, *y*, *nulp=1*)
Compare two arrays relatively to their spacing.

Parameters

- **x** (*numpy.ndarray or cupy.ndarray*) – The actual object to check.
- **y** (*numpy.ndarray or cupy.ndarray*) – The desired, expected object.
- **nulp** (*int*) – The maximum number of unit in the last place for tolerance.

See also:

`numpy.testing.assert_array_almost_equal_nulp()`

cupy.testing.**assert_array_max_ulp**(*a*, *b*, *maxulp=1*, *dtype=None*)
Check that all items of arrays differ in at most N Units in the Last Place.

Parameters

- **a** (*numpy.ndarray or cupy.ndarray*) – The actual object to check.
- **b** (*numpy.ndarray or cupy.ndarray*) – The desired, expected object.
- **maxulp** (*int*) – The maximum number of units in the last place that elements of `a` and `b` can differ.
- **dtype** (*numpy.dtype*) – Data-type to convert `a` and `b` to if given.

See also:

`numpy.testing.assert_array_max_ulp()`

cupy.testing.**assert_array_equal**(*x*, *y*, *err_msg=''*, *verbose=True*)
Raises an AssertionError if two array_like objects are not equal.

Parameters

- **x** (*numpy.ndarray or cupy.ndarray*) – The actual object to check.
- **y** (*numpy.ndarray or cupy.ndarray*) – The desired, expected object.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.

See also:

`numpy.testing.assert_array_equal()`

cupy.testing.**assert_array_list_equal**(*xlist*, *ylist*, *err_msg=''*, *verbose=True*)
Compares lists of arrays pairwise with `assert_array_equal`.

Parameters

- **x** (*array_like*) – Array of the actual objects.

- **y** (*array_like*) – Array of the desired, expected objects.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.

Each element of `x` and `y` must be either `numpy.ndarray` or *`cupy.ndarray`*. `x` and `y` must have same length. Otherwise, this function raises `AssertionError`. It compares elements of `x` and `y` pairwise with *`assert_array_equal()`* and raises error if at least one pair is not equal.

See also:

`numpy.testing.assert_array_equal()`

`cupy.testing.`**`assert_array_less`**(*x*, *y*, *err_msg=''*, *verbose=True*)

Raises an AssertionError if array_like objects are not ordered by less than.

Parameters

- **x** (*numpy.ndarray or cupy.ndarray*) – The smaller object to check.
- **y** (*numpy.ndarray or cupy.ndarray*) – The larger object to compare.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.

See also:

`numpy.testing.assert_array_less()`

NumPy-CuPy Consistency Check

The following decorators are for testing consistency between CuPy's functions and corresponding NumPy's ones.

`cupy.testing.`**`numpy_cupy_allclose`**(*rtol=1e-07*, *atol=0*, *err_msg=''*, *verbose=True*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks NumPy results and CuPy ones are close.

Parameters

- **rtol** (*float*) – Relative tolerance.
- **atol** (*float*) – Absolute tolerance
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the arrays whose values are close between `numpy` case and `cupy` case. For example, this test case checks `numpy.zeros` and `cupy.zeros` should return same value.

```
>>> import unittest
>>> from cupy import testing
>>> @testing.gpu
... class TestFoo(unittest.TestCase):
```

```
...
...     @testing.numpy_cupy_allclose()
...     def test_foo(self, xp):
...         # ...
...         # Prepare data with xp
...         # ...
...
...         xp_result = xp.zeros(10)
...         return xp_result
```

See also:

cupy.testing.assert_allclose()

cupy.testing.**numpy_cupy_array_almost_equal**(*decimal=6*, *err_msg=''*, *verbose=True*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks NumPy results and CuPy ones are almost equal.

Parameters

- **decimal** (*int*) – Desired precision.
- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the same arrays in the sense of *cupy.testing.assert_array_almost_equal()* (except the type of array module) even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_almost_equal()

cupy.testing.**numpy_cupy_array_almost_equal_nulp**(*nulp=1*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks results of NumPy and CuPy are equal w.r.t. spacing.

Parameters

- **nulp** (*int*) – The maximum number of unit in the last place for tolerance.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the same arrays in the sense of *cupy.testing.assert_array_almost_equal_nulp()* (except the type of array module) even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_almost_equal_nulp()

cupy.testing.**numpy_cupy_array_max_ulp**(*maxulp=1*, *dtype=None*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks results of NumPy and CuPy ones are equal w.r.t. ulp.

Parameters

- **maxulp** (*int*) – The maximum number of units in the last place that elements of resulting two arrays can differ.
- **dtype** (*numpy.dtype*) – Data-type to convert the resulting two array to if given.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the same arrays in the sense of *assert_array_max_ulp()* (except the type of array module) even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_max_ulp()

cupy.testing.**numpy_cupy_array_equal**(*err_msg=''*, *verbose=True*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks NumPy results and CuPy ones are equal.

Parameters

- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the same arrays in the sense of *numpy_cupy_array_equal()* (except the type of array module) even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_equal()

cupy.testing.**numpy_cupy_array_list_equal**(*err_msg=''*, *verbose=True*, *name='xp'*)

Decorator that checks the resulting lists of NumPy and CuPy's one are equal.

Parameters

- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.

Decorated test fixture is required to return the same list of arrays (except the type of array module) even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_list_equal()

cupy.testing.**numpy_cupy_array_less**(*err_msg=''*, *verbose=True*, *name='xp'*, *type_check=True*, *accept_error=False*)

Decorator that checks the CuPy result is less than NumPy result.

Parameters

- **err_msg** (*str*) – The error message to be printed in case of failure.
- **verbose** (*bool*) – If `True`, the conflicting values are appended to the error message.
- **name** (*str*) – Argument name whose value is either `numpy` or `cupy` module.
- **type_check** (*bool*) – If `True`, consistency of dtype is also checked.
- **accept_error** (*bool, Exception or tuple of Exception*) – Sepcify acceptable errors. When both NumPy test and CuPy test raises the same type of errors, and the type of the errors is specified with this argument, the errors are ignored and not raised. If it is `True` all error types are acceptable. If it is `False` no error is acceptable.

Decorated test fixture is required to return the smaller array when `xp` is `cupy` than the one when `xp` is `numpy`.

See also:

cupy.testing.assert_array_less()

cupy.testing.**numpy_cupy_raises**(*name='xp'*)

Decorator that checks the NumPy and CuPy throw same errors.

Parameters

- **name** (*str*) – Argument name whose value is either
- **or cupy module.** (*numpy*) –

Decorated test fixture is required throw same errors even if `xp` is `numpy` or `cupy`.

See also:

cupy.testing.assert_array_less()

Parameterized dtype Test

The following decorators offer the standard way for parameterized test with respect to single or the combination of dtype(s).

cupy.testing.**for_dtypes**(*dtypes*, *name='dtype'*)

Decorator for parameterized dtype test.

Parameters

- **dtypes** (*list of dtypes*) – dtypes to be tested.
- **name** (*str*) – Argument name to which specified dtypes are passed.

This decorator adds a keyword argument specified by `name` to the test fixture. Then, it runs the fixtures in parallel by passing the each element of `dtypes` to the named argument.

cupy.testing.**for_all_dtypes**(*name='dtype'*, *no_float16=False*, *no_bool=False*)

Decorator that checks the fixture with all dtypes.

Parameters

- **name** (*str*) – Argument name to which specified dtypes are passed.
- **no_float16** (*bool*) – If, True, `numpy.float16` is omitted from candidate dtypes.
- **no_bool** (*bool*) – If, True, `numpy.bool_` is omitted from candidate dtypes.

dtypes to be tested: `numpy.float16` (optional), `numpy.float32`, `numpy.float64`, `numpy.dtype('b')`, `numpy.dtype('h')`, `numpy.dtype('i')`, `numpy.dtype('l')`, `numpy.dtype('q')`, `numpy.dtype('B')`, `numpy.dtype('H')`, `numpy.dtype('I')`, `numpy.dtype('L')`, `numpy.dtype('Q')`, and `numpy.bool_` (optional).

The usage is as follows. This test fixture checks if `cPickle` successfully reconstructs *cupy.ndarray* for various dtypes. `dtype` is an argument inserted by the decorator.

```
>>> import unittest
>>> from cupy import testing
>>> @testing.gpu
... class TestNpz(unittest.TestCase):
...
...     @testing.for_all_dtypes()
...     def test_pickle(self, dtype):
...         a = testing.shaped_arange((2, 3, 4), dtype=dtype)
...         s = six.moves.cPickle.dumps(a)
...         b = six.moves.cPickle.loads(s)
...         testing.assert_array_equal(a, b)
```

Typically, we use this decorator in combination with decorators that check consistency between NumPy and CuPy like *cupy.testing.numpy_cupy_allclose()*. The following is such an example.

```
>>> import unittest
>>> from cupy import testing
>>> @testing.gpu
... class TestMean(unittest.TestCase):
...
...     @testing.for_all_dtypes()
...     @testing.numpy_cupy_allclose()
...     def test_mean_all(self, xp, dtype):
...         a = testing.shaped_arange((2, 3), xp, dtype)
...         return a.mean()
```

See also:

cupy.testing.for_dtypes()

cupy.testing.**for_float_dtypes**(*name='dtype'*, *no_float16=False*)

Decorator that checks the fixture with all float dtypes.

Parameters

- **name** (*str*) – Argument name to which specified dtypes are passed.
- **no_float16** (*bool*) – If, True, `numpy.float16` is omitted from candidate dtypes.

dtypes to be tested are `numpy.float16` (optional), `numpy.float32`, and `numpy.float64`.

See also:

cupy.testing.for_dtypes(), *cupy.testing.for_all_dtypes()*

cupy.testing.**for_signed_dtypes**(*name='dtype'*)

Decorator that checks the fixture with signed dtypes.

Parameters **name** (*str*) – Argument name to which specified dtypes are passed.

dtypes to be tested are `numpy.dtype('b')`, `numpy.dtype('h')`, `numpy.dtype('i')`, `numpy.dtype('l')`, and `numpy.dtype('q')`.

See also:

cupy.testing.for_dtypes(), *cupy.testing.for_all_dtypes()*

cupy.testing.**for_unsigned_dtypes**(*name='dtype'*)
Decorator that checks the fixture with all dtypes.

Parameters **name** (*str*) – Argument name to which specified dtypes are passed.

dtypes to be tested are `numpy.dtype('B')`, `numpy.dtype('H')`,

`numpy.dtype('I')`, `numpy.dtype('L')`, and `numpy.dtype('Q')`.

See also:

cupy.testing.for_dtypes(), *cupy.testing.for_all_dtypes()*

cupy.testing.**for_int_dtypes**(*name='dtype'*, *no_bool=False*)
Decorator that checks the fixture with integer and optionally bool dtypes.

Parameters

- **name** (*str*) – Argument name to which specified dtypes are passed.
- **no_bool** (*bool*) – If `True`, `numpy.bool_` is omitted from candidate dtypes.

dtypes to be tested are `numpy.dtype('b')`, `numpy.dtype('h')`, `numpy.dtype('i')`, `numpy.dtype('l')`, `numpy.dtype('q')`, `numpy.dtype('B')`, `numpy.dtype('H')`, `numpy.dtype('I')`, `numpy.dtype('L')`, `numpy.dtype('Q')`, and `numpy.bool_` (optional).

See also:

cupy.testing.for_dtypes(), *cupy.testing.for_all_dtypes()*

cupy.testing.**for_dtypes_combination**(*types*, *names=('dtype',)*, *full=None*)
Decorator that checks the fixture with a product set of dtypes.

Parameters

- **types** (*list of dtypes*) – dtypes to be tested.
- **names** (*list of str*) – Argument names to which dtypes are passed.
- **full** (*bool*) – If `True`, then all combinations of dtypes will be tested. Otherwise, the subset of combinations will be tested (see the description below).

Decorator adds the keyword arguments specified by `names` to the test fixture. Then, it runs the fixtures in parallel with passing (possibly a subset of) the product set of dtypes. The range of dtypes is specified by `types`.

The combination of dtypes to be tested changes depending on the option `full`. If `full` is `True`, all combinations of `types` are tested. Sometimes, such an exhaustive test can be costly. So, if `full` is `False`, only the subset of possible combinations is tested. Specifically, at first, the shuffled lists of `types` are made for each argument name in `names`. Let the lists be `D1`, `D2`, ..., `Dn` where n is the number of arguments. Then, the combinations to be tested will be `zip(D1, ..., Dn)`. If `full` is `None`, the behavior is switched by setting the environment variable `CUPY_TEST_FULL_COMBINATION=1`.

For example, let `types` be `[float16, float32, float64]` and `names` be `['a_type', 'b_type']`. If `full` is `True`, then the decorated test fixture is executed with all 2^3 patterns. On the other hand, if `full` is `False`, shuffled lists are made for `a_type` and `b_type`. Suppose the lists are `(16, 64,`

`32)` for `a_type` and `(32, 64, 16)` for `b_type` (prefixes are removed for short). Then the combinations of `(a_type, b_type)` to be tested are `(16, 32)`, `(64, 64)` and `(32, 16)`.

cupy.testing.**for_all_dtypes_combination**(*names=('dtyes',)*, *no_float16=False*, *no_bool=False*, *full=None*)

Decorator that checks the fixture with a product set of all dtypes.

Parameters

- **names** (*list of str*) – Argument names to which dtypes are passed.
- **no_float16** (*bool*) – If `True`, `numpy.float16` is omitted from candidate dtypes.
- **no_bool** (*bool*) – If `True`, `numpy.bool_` is omitted from candidate dtypes.
- **full** (*bool*) – If `True`, then all combinations of dtypes will be tested. Otherwise, the subset of combinations will be tested (see description in *cupy.testing.for_dtypes_combination()*).

See also:

cupy.testing.for_dtypes_combination()

cupy.testing.**for_signed_dtypes_combination**(*names=('dtype',)*, *full=None*)

Decorator for parameterized test w.r.t. the product set of signed dtypes.

Parameters

- **names** (*list of str*) – Argument names to which dtypes are passed.
- **full** (*bool*) – If `True`, then all combinations of dtypes will be tested. Otherwise, the subset of combinations will be tested (see description in *cupy.testing.for_dtypes_combination()*).

See also:

cupy.testing.for_dtypes_combination()

cupy.testing.**for_unsigned_dtypes_combination**(*names=('dtype',)*, *full=None*)

Decorator for parameterized test w.r.t. the product set of unsigned dtypes.

Parameters

- **names** (*list of str*) – Argument names to which dtypes are passed.
- **full** (*bool*) – If `True`, then all combinations of dtypes will be tested. Otherwise, the subset of combinations will be tested (see description in *cupy.testing.for_dtypes_combination()*).

See also:

cupy.testing.for_dtypes_combination()

cupy.testing.**for_int_dtypes_combination**(*names=('dtype',)*, *no_bool=False*, *full=None*)

Decorator for parameterized test w.r.t. the product set of int and boolean.

Parameters

- **names** (*list of str*) – Argument names to which dtypes are passed.
- **no_bool** (*bool*) – If `True`, `numpy.bool_` is omitted from candidate dtypes.
- **full** (*bool*) – If `True`, then all combinations of dtypes will be tested. Otherwise, the subset of combinations will be tested (see description in *cupy.testing.for_dtypes_combination()*).

See also:

cupy.testing.for_dtypes_combination()

Parameterized order Test

The following decorators offer the standard way to parameterize tests with orders.

cupy.testing.**for_orders**(*orders*, *name='order'*)
Decorator to parameterize tests with order.

Parameters

- **orders** (*list of orders*) – orders to be tested.
- **name** (*str*) – Argument name to which the specified order is passed.

This decorator adds a keyword argument specified by `name` to the test fixtures. Then, the fixtures run by passing each element of `orders` to the named argument.

cupy.testing.**for_CF_orders**(*name='order'*)
Decorator that checks the fixture with orders 'C' and 'F'.

Parameters **name** (*str*) – Argument name to which the specified order is passed.

See also:

cupy.testing.for_all_dtypes()

Environment variables

Here are the environment variables Chainer uses.

`CUPY_CACHE_DIR`	Path to the directory to store kernel cache. `$(HOME)/.cupy.kernel_cache` is used by default. See *CuPy Overview* for detail.

For install

These environment variables are only used during installation.

`CUDA_PATH`	Path to the directory containing CUDA. The parent of the directory containing `nvcc` is used as default. When `nvcc` is not found, `/usr/local/cuda` is used. See *Install Chainer with CUDA* for details.

Difference between CuPy and NumPy

The interface of CuPy is designed to obey that of NumPy. However, there are some differeneces.

Cast behavior from float to integer

Some casting behaviors from float to integer are not defined in C++ specification. The casting from a negative float to unsigned integer and infinity to integer is one of such eamples. The behavior of NumPy depends on your CPU architecture. This is Intel CPU result.

```
>>> np.array([-1], dtype='f').astype('I')
array([4294967295], dtype=uint32)
>>> cupy.array([-1], dtype='f').astype('I')
array([0], dtype=uint32)
```

```
>>> np.array([float('inf')], dtype='f').astype('i')
array([-2147483648], dtype=int32)
>>> cupy.array([float('inf')], dtype='f').astype('i')
array([2147483647], dtype=int32)
```

Boolean values squared

In NumPy implementation, `x ** 2` is calculated using multiplication operator as `x * x`. Because the result of the multiplication of boolean values is boolean, `True ** 2` return boolean value. However, when you use power operator with other arguments, it returns int values. If we aligned the behavior of the squared boolean values of CuPy to that of NumPy, we would have to check their values in advance of the calculation. But it would be slow because it would force CPUs to wait until the calculation on GPUs end. So we decided not to check its value.

```
>>> np.array([True]) ** 2
array([ True], dtype=bool)
>>> cupy.array([True]) ** 2
array([1])
```

Random methods support dtype argument

NumPy's random value generator does not support dtype option and it always resturns a `float32` value. We support the option in CuPy because cuRAND, which is used in CuPy, supports any types of float values.

```
>>> np.random.randn(dtype='f')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: randn() got an unexpected keyword argument 'dtype'
>>> cupy.random.randn(dtype='f')
array(0.10689262300729752, dtype=float32)
```

Out-of-bounds indices

CuPy handles out-of-bounds indices differently by default from NumPy when using integer array indexing. NumPy handles them by raising an error, but CuPy wraps around them.

```
>>> x = np.array([0, 1, 2])
>>> x[[1, 3]] = 10
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
IndexError: index 3 is out of bounds for axis 1 with size 3
>>> x = cupy.array([0, 1, 2])
>>> x[[1, 3]] = 10
>>> x
array([10, 10,  2])
```

Duplicate values in indices

CuPy's `__setitem__` behaves differently from NumPy when integer arrays reference the same location multiple times. In that case, the value that is actually stored is undefined. Here is an example of CuPy.

```
>>> a = cupy.zeros((2,))
>>> i = cupy.arange(10000) % 2
>>> v = cupy.arange(10000).astype(np.float)
>>> a[i] = v
>>> a
array([ 9150.,  9151.])
```

NumPy stores the value corresponding to the last element among elements referencing duplicate locations.

```
>>> a_cpu = np.zeros((2,))
>>> i_cpu = np.arange(10000) % 2
>>> v_cpu = np.arange(10000).astype(np.float)
>>> a_cpu[i_cpu] = v_cpu
>>> a_cpu
array([ 9998.,  9999.])
```

CHAPTER 5

Chainer Contribution Guide

This is a guide for all contributions to Chainer. The development of Chainer is running on the official repository at GitHub. Anyone that wants to register an issue or to send a pull request should read through this document.

Classification of Contributions

There are several ways to contribute to Chainer community:

1. Registering an issue
2. Sending a pull request (PR)
3. Sending a question to Chainer User Group
4. Open-sourcing an external example
5. Writing a post about Chainer

This document mainly focuses on 1 and 2, though other contributions are also appreciated.

Release and Milestone

We are using GitHub Flow as our basic working process. In particular, we are using the master branch for our development, and releases are made as tags.

Releases are classified into three groups: major, minor, and revision. This classification is based on following criteria:

- **Major update** contains disruptive changes that break the backward compatibility.
- **Minor update** contains additions and extensions to the APIs keeping the supported backward compatibility.
- **Revision update** contains improvements on the API implementations without changing any API specification.

The release classification is reflected into the version number x.y.z, where x, y, and z corresponds to major, minor, and revision updates, respectively.

We set a milestone for an upcoming release. The milestone is of name 'vX.Y.Z', where the version number represents a revision release at the outset. If at least one *feature* PR is merged in the period, we rename the milestone to represent a minor release (see the next section for the PR types).

See also *API Compatibility Policy*.

Issues and PRs

Issues and PRs are classified into following categories:

- **Bug**: bug reports (issues) and bug fixes (PRs)
- **Enhancement**: implementation improvements without breaking the interface
- **Feature**: feature requests (issues) and their implementations (PRs)
- **NoCompat**: disrupts backward compatibility
- **Test**: test fixes and updates
- **Document**: document fixes and improvements
- **Example**: fixes and improvements on the examples
- **Install**: fixes installation script
- **Contribution-Welcome**: issues that we request for contribution (only issues are categorized to this)
- **Other**: other issues and PRs

Issues and PRs are labeled by these categories. This classification is often reflected into its corresponding release category: Feature issues/PRs are contained into minor/major releases and NoCompat issues/PRs are contained into major releases, while other issues/PRs can be contained into any releases including revision ones.

On registering an issue, write precise explanations on what you want Chainer to be. Bug reports must include necessary and sufficient conditions to reproduce the bugs. Feature requests must include **what** you want to do (and **why** you want to do, if needed). You can contain your thoughts on **how** to realize it into the feature requests, though **what** part is most important for discussions.

Warning: If you have a question on usages of Chainer, it is highly recommended to send a post to Chainer User Group instead of the issue tracker. The issue tracker is not a place to share knowledge on practices. We may redirect question issues to Chainer User Group.

If you can write code to fix an issue, send a PR to the master branch. Before writing your code for PRs, read through the *Coding Guidelines*. The description of any PR must contain a precise explanation of **what** and **how** you want to do; it is the first documentation of your code for developers, a very important part of your PR.

Once you send a PR, it is automatically tested on Travis CI for Linux and Mac OS X, and on AppVeyor for Windows. Your PR need to pass at least the test for Linux on Travis CI. After the automatic test passes, some of the core developers will start reviewing your code. Note that this automatic PR test only includes CPU tests.

Note: We are also running continuous integration with GPU tests for the master branch. Since this service is running on our internal server, we do not use it for automatic PR tests to keep the server secure.

Even if your code is not complete, you can send a pull request as a *work-in-progress PR* by putting the `[WIP]` prefix to the PR title. If you write a precise explanation about the PR, core developers and other contributors can join the discussion about how to proceed the PR.

Coding Guidelines

We use PEP8 and a part of OpenStack Style Guidelines related to general coding style as our basic style guidelines.

To check your code, use `autopep8` and `flake8` command installed by `hacking` package:

```
$ pip install autopep8 hacking
$ autopep8 --global-config .pep8 path/to/your/code.py
$ flake8 path/to/your/code.py
```

To check Cython code, use `.flake8.cython` configuration file:

```
$ flake8 --config=.flake8.cython path/to/your/cython/code.pyx
```

The `autopep8` supports automatically correct Python code to conform to the PEP 8 style guide:

```
$ autopep8 --in-place --global-config .pep8 path/to/your/code.py
```

The `flake8` command lets you know the part of your code not obeying our style guidelines. Before sending a pull request, be sure to check that your code passes the `flake8` checking.

Note that `flake8` command is not perfect. It does not check some of the style guidelines. Here is a (not-complete) list of the rules that `flake8` cannot check.

- Relative imports are prohibited. [H304]
- Importing non-module symbols is prohibited.
- Import statements must be organized into three parts: standard libraries, third-party libraries, and internal imports. [H306]

In addition, we restrict the usage of *shortcut symbols* in our code base. They are symbols imported by packages and sub-packages of `chainer`. For example, `chainer.Variable` is a shortcut of `chainer.variable.Variable`. **It is not allowed to use such shortcuts in the ‘‘chainer‘‘ library implementation**. Note that you can still use them in `tests` and `examples` directories. Also note that you should use shortcut names of CuPy APIs in Chainer implementation.

Once you send a pull request, your coding style is automatically checked by Travis-CI. The reviewing process starts after the check passes.

The CuPy is designed based on NumPy's API design. CuPy's source code and documents contain the original NumPy ones. Please note the followings when writing the document.

- In order to identify overlapping parts, it is preferable to add some remarks that this document is just copied or altered from the original one. It is also preferable to briefly explain the specification of the function in a short paragraph, and refer to the corresponding function in NumPy so that users can read the detailed document. However, it is possible to include a complete copy of the document with such a remark if users cannot summarize in such a way.
- If a function in CuPy only implements a limited amount of features in the original one, users should explicitly describe only what is implemented in the document.

Testing Guidelines

Testing is one of the most important part of your code. You must test your code by unit tests following our testing guidelines. Note that we are using the nose package and the mock package for testing, so install nose and mock before writing your code:

```
$ pip install nose mock
```

In order to run unit tests at the repository root, you first have to build Cython files in place by running the following command:

```
$ python setup.py develop
```

Once the Cython modules are built, you can run unit tests simply by running `nosetests` command at the repository root:

```
$ nosetests
```

It requires CUDA by default. In order to run unit tests that do not require CUDA, pass `--attr='!gpu'` option to the `nosetests` command:

```
$ nosetests path/to/your/test.py --attr='!gpu'
```

Some GPU tests involve multiple GPUs. If you want to run GPU tests with insufficient number of GPUs, specify the number of available GPUs by `--eval-attr='gpu<N'` where `N` is a concrete integer. For example, if you have only one GPU, launch `nosetests` by the following command to skip multi-GPU tests:

```
$ nosetests path/to/gpu/test.py --eval-attr='gpu<2'
```

Some tests spend too much time. If you want to skip such tests, pass `--attr='!slow'` option to the `nosetests` command:

```
$ nosetests path/to/your/test.py --attr='!slow'
```

Tests are put into the `tests/chainer_tests`, `tests/cupy_tests` and `tests/install_tests` directories. These have the same structure as that of `chainer`, `cupy` and `install` directories, respectively. In order to enable test runner to find test scripts correctly, we are using special naming convention for the test subdirectories and the test scripts.

- The name of each subdirectory of `tests` must end with the `_tests` suffix.
- The name of each test script must start with the `test_` prefix.

Following this naming convention, you can run all the tests by just typing `nosetests` at the repository root:

```
$ nosetests
```

Or you can also specify a root directory to search test scripts from:

```
$ nosetests tests/chainer_tests  # to just run tests of Chainer
$ nosetests tests/cupy_tests     # to just run tests of CuPy
$ nosetests tests/install_tests  # to just run tests of installation modules
```

If you modify the code related to existing unit tests, you must run appropriate commands.

Note: CuPy tests include type-exhaustive test functions which take long time to execute. If you are running tests on a multi-core machine, you can parallelize the tests by following options:

```
$ nosetests --processes=12 --process-timeout=1000 tests/cupy_tests
```

The magic numbers can be modified for your usage. Note that some tests require many CUDA compilations, which require a bit long time. Without the `process-timeout` option, the timeout is set shorter, causing timeout failures

for many test cases.

There are many examples of unit tests under the `tests` directory. They simply use the `unittest` package of the standard library.

Even if your patch includes GPU-related code, your tests should not fail without GPU capability. Test functions that require CUDA must be tagged by the `chainer.testing.attr.gpu` decorator (or `cupy.testing.attr.gpu` for testing CuPy APIs):

```
import unittest
from chainer.testing import attr

class TestMyFunc(unittest.TestCase):
    ...

    @attr.gpu
    def test_my_gpu_func(self):
        ...
```

The functions tagged by the `gpu` decorator are skipped if `--attr='!gpu'` is given. We also have the `chainer.testing.attr.cudnn` decorator to let `nosetests` know that the test depends on cuDNN.

The test functions decorated by `gpu` must not depend on multiple GPUs. In order to write tests for multiple GPUs, use `chainer.testing.attr.multi_gpu()` or `cupy.testing.attr.multi_gpu()` decorators instead:

```
import unittest
from chainer.testing import attr

class TestMyFunc(unittest.TestCase):
    ...

    @attr.multi_gpu(2)  # specify the number of required GPUs here
    def test_my_two_gpu_func(self):
        ...
```

If your test requires too much time, add `chainer.testing.attr.slow` decorator. The test functions decorated by `slow` are skipped if `--attr='!slow'` is given:

```
import unittest
from chainer.testing import attr

class TestMyFunc(unittest.TestCase):
    ...

    @attr.slow
    def test_my_slow_func(self):
        ...
```

Note: If you want to specify more than two attributes, separate them with a comma such as `--attr='!gpu,!slow'`. See detail in the document of nose.

Once you send a pull request, your code is automatically tested by Travis-CI **with –attr='!gpu,!slow' option**. Since Travis-CI does not support CUDA, we cannot check your CUDA-related code automatically. The reviewing process starts after the test passes. Note that reviewers will test your code without the option to check CUDA-related code.

Note: Some of numerically unstable tests might cause errors irrelevant to your changes. In such a case, we ignore the failures and go on to the review process, so do not worry about it.

CHAPTER 6

API Compatibility Policy

This document expresses the design policy on compatibilities of Chainer APIs. Development team should obey this policy on deciding to add, extend, and change APIs and their behaviors.

This document is written for both users and developers. Users can decide the level of dependencies on Chainer's implementations in their codes based on this document. Developers should read through this document before creating pull requests that contain changes on the interface. Note that this document may contain ambiguities on the level of supported compatibilities.

Targeted Versions

This policy is applied to Chainer of versions v1.5.1 and higher. Note that this policy is not applied to Chainer of lower versions.

Versioning and Backward Compatibilities

The updates of Chainer are classified into three levels: major, minor, and revision. These types have distinct levels of backward compatibilities.

- **Major update** contains disruptive changes that break the backward compatibility.
- **Minor update** contains addition and extension to the APIs keeping the supported backward compatibility.
- **Revision update** contains improvements on the API implementations without changing any API specifications.

Note that we do not support full backward compatibility, which is almost infeasible for Python-based APIs, since there is no way to completely hide the implementation details.

Processes to Break Backward Compatibilities

Deprecation, Dropping, and Its Preparation

Any APIs may be *deprecated* at some minor updates. In such a case, the deprecation note is added to the API documentation, and the API implementation is changed to fire deprecation warning (if possible). There should be another way to reimplement the same things previously written with the deprecated APIs.

Any APIs may be marked as *to be dropped in the future*. In such a case, the dropping is stated in the documentation with the major version number on which the API is planned to be dropped, and the API implementation is changed to fire the future warning (if possible).

The actual dropping should be done through the following steps:

- Make the API deprecated. At this point, users should not need the deprecated API in their new application codes.
- After that, mark the API as *to be dropped in the future*. It must be done in the minor update different from that of the deprecation.
- At the major version announced in the above update, drop the API.

Consequently, it takes at least two minor versions to drop any APIs after the first deprecation.

API Changes and Its Preparation

Any APIs may be marked as *to be changed in the future* for changes without backward compatibility. In such a case, the change is stated in the documentation with the version number on which the API is planned to be changed, and the API implementation is changed to fire the future warning on the certain usages.

The actual change should be done in the following steps:

- Announce that the API will be changed in the future. At this point, the actual version of change need not be accurate.
- After the announcement, mark the API as *to be changed in the future* with version number of planned changes. At this point, users should not use the marked API in their new application codes.
- At the major update announced in the above update, change the API.

Experimental APIs

Thanks to many contributors, we have introduced many new features to Chainer.

However, we have sometimes released new features only to later notice that their APIs are not appropriate. The objective of experimental APIs is to avoid such issues by allowing the developer to mark any newly added API as experimental.

Any newly added API can be marked as *experimental*. Any API that is not experimental is called *stable* in this document.

Note: Undocumented behaviors are not considered as APIs. So they are not experimental nor stable. The treatment of undocumented behaviors are descibed in *Undocumented behaviors* section.

Chainer can change the interfaces and documents of experimental APIs at **any** version up. This change is not considered as a break of backward compatibility. Chainer can promote an experimental API to become stable at any **minor** or **major** version up. Once experimental APIs become stable, they cannot revert to experimental again.

When users use experimental APIs for the first time, warnings are raised once for each experimental API, unless users explicitly disable the emission of the warnings in advance.

See the document of *chainer.utils.experimental()* how developers mark APIs as experimental and how users enable or disable the warnings practically.

Note: It is up to developers if APIs should be annotated as experimental or not. We recommend to make the APIs experimental if they implement large modules or make a decision from several design choices.

Supported Backward Compatibility

This section defines backward compatibilities that minor updates must maintain.

Documented Interface

Chainer has the official API documentation. Many applications can be written based on the documented features. We support backward compatibilities of documented features. In other words, codes only based on the documented features run correctly with minor/revision-updated versions.

Developers are encouraged to use apparent names for objects of implementation details. For example, attributes outside of the documented APIs should have one or more underscores at the prefix of their names.

Undocumented behaviors

Behaviors of Chainer implementation not stated in the documentation are undefined. Undocumented behaviors are not guaranteed to be stable between different minor/revision versions.

Minor update may contain changes to undocumented behaviors. For example, suppose an API X is added at the minor update. In the previous version, attempts to use X cause AttributeError. This behavior is not stated in the documentation, so this is undefined. Thus, adding the API X in minor version is permissible.

Revision update may also contain changes to undefined behaviors. Typical example is a bug fix. Another example is an improvement on implementation, which may change the internal object structures not shown in the documentation. As a consequence, **even revision updates do not support compatibility of pickling, unless the full layout of pickled objects is clearly documented.**

Documentation Error

Compatibility is basically determined based on the documentation, though it sometimes contains errors. It may make the APIs confusing to assume the documentation always stronger than the implementations. We therefore may fix the documentation errors in any updates that may break the compatibility in regard to the documentation.

Note: Developers MUST NOT fix the documentation and implementation of the same functionality at the same time in revision updates as "bug fix". Such a change completely breaks the backward compatibility. If you want to fix the

bugs in both sides, first fix the documentation to fit it into the implementation, and start the API changing procedure described above.

Object Attributes and Properties

Object attributes and properties are sometimes replaced by each other at minor updates. It does not break the user codes, except the codes depend on how the attributes and properties are implemented.

Functions and Methods

Methods may be replaced by callable attributes keeping the compatibility of parameters and return values in minor updates. It does not break the user codes, except the codes depend on how the methods and callable attributes are implemented.

Exceptions and Warnings

The specifications of raising exceptions are considered as a part of standard backward compatibilities. No exception is raised in the future versions with correct usages that the documentation allows, unless the API changing process is completed.

On the other hand, warnings may be added at any minor updates for any APIs. It means minor updates do not keep backward compatibility of warnings.

Model Format Compatibility

Objects serialized by official serializers that Chainer provides are correctly loaded with the higher (future) versions. They might not be correctly loaded with Chainer of the lower versions.

Note: Current serialization APIs do not support versioning (at least in v1.6.1). It prevents us from introducing changes in the layout of objects that support serialization. We are discussing about introducing versioning in serialization APIs.

Installation Compatibility

The installation process is another concern of compatibilities. We support environmental compatibilities in the following ways.

- Any changes of dependent libraries that force modifications on the existing environments must be done in major updates. Such changes include following cases:
  - dropping supported versions of dependent libraries (e.g. dropping cuDNN v2)
  - adding new mandatory dependencies (e.g. adding h5py to setup_requires)
- Supporting optional packages/libraries may be done in minor updates (e.g. supporting h5py in optional features).

Note: The installation compatibility does not guarantee that all the features of Chainer correctly run on supported environments. It may contain bugs that only occurs in certain environments. Such bugs should be fixed in some updates.

CHAPTER 7

Tips and FAQs

It takes too long time to compile a computational graph. Can I skip it?

Chainer does not compile computational graphs, so you cannot skip it, or, I mean, you have already skipped it :).

It seems you have actually seen on-the-fly compilations of CUDA kernels. CuPy compiles kernels on demand to make kernels optimized to the number of dimensions and element types of input arguments. Pre-compilation is not available, because we have to compile an exponential number of kernels to support all CuPy functionalities. This restriction is unavoidable because Python cannot call CUDA/C++ template functions in generic way. Note that every framework using CUDA require compilation at some point; the difference between other statically-compiled frameworks (such as cutorch) and Chainer is whether a kernel is compiled at installation or at the first use.

These compilations should run only at the first use of the kernels. The compiled binaries are cached to the `$(HOME)/.cupy/kernel_cache` directory by default. If you see that compilations run every time you run the same script, then the caching is failed. Please check that the directory is kept as is between multiple executions of the script. If your home directory is not suited to caching the kernels (e.g. in case that it uses NFS), change the kernel caching directory by setting the `CUPY_CACHE_DIR` environment variable to an appropriate path. See *CuPy Overview* for more details.

mnist example does not converge in CPU mode on Mac OS X

Many users reported that mnist example does not work correctly on Mac OS X. We are suspecting it is caused by vecLib, that is a default BLAS library installed on Mac OS X.

Note: Mac OS X is not officially supported. I mean it is not tested continuously on our test server.

We recommend to use other BLAS libraries such as OpenBLAS. We empirically found that it fixes this problem. It is necessary to reinstall NumPy to use replaced BLAS library. Here is an instruction to install NumPy with OpenBLAS using Homebrew.

```
$ brew tap homebrew/science
$ brew install openblas
$ brew install numpy --with-openblas
```

If you want to install NumPy with pip, use site.cfg file.

You can check if NumPy uses OpenBLAS with `numpy.show_config` method. Check if *blas_opt_info* refers to *openblas*.

```
>>> import numpy
>>> numpy.show_config()
lapack_opt_info:
    libraries = ['openblas', 'openblas']
    library_dirs = ['/usr/local/opt/openblas/lib']
    define_macros = [('HAVE_CBLAS', None)]
    language = c
blas_opt_info:
    libraries = ['openblas', 'openblas']
    library_dirs = ['/usr/local/opt/openblas/lib']
    define_macros = [('HAVE_CBLAS', None)]
    language = c
openblas_info:
    libraries = ['openblas', 'openblas']
    library_dirs = ['/usr/local/opt/openblas/lib']
    define_macros = [('HAVE_CBLAS', None)]
    language = c
openblas_lapack_info:
    libraries = ['openblas', 'openblas']
    library_dirs = ['/usr/local/opt/openblas/lib']
    define_macros = [('HAVE_CBLAS', None)]
    language = c
blas_mkl_info:
    NOT AVAILABLE
```

See detail about this problem in issue #704.

CHAPTER 8

Comparison with Other Frameworks

A table for quick comparison

This table compares Chainer with other popular deep learning frameworks. We hope it helps you to choose an appropriate framework for the demand.

Note: This chart may be out-dated, since the developers of Chainer do not perfectly follow the latest development status of each framework. Please report us if you find an out-dated cell. Requests for new comparison axes are also welcome.

		Chainer	Theano-based	Torch7	Caffe
Specs	Scripting	Python	Python	LuaJIT	Python
	Net definition language	Python	Python	LuaJIT	Protocol Buffers
	Define-by-Run scheme	Y			
	CPU Array backend	NumPy	NumPy	Tensor	
	GPU Array backend	CuPy	CudaNdarray[1]	CudaTensor	
NNs	Reverse-mode AD	Y	Y	Y	Y
	Basic RNN support	Y	Y	Y (`nnx`)	#2033
	Variable-length loops	Y	Y (`scan`)		
	Stateful RNNs[2]	Y	Y	Y[6]	
	Per-batch architectures	Y			
Perf	CUDA support	Y	Y	Y	Y
	cuDNN support	Y	Y	Y (`cudnn.torch`)	Y
	FFT-based convolution		Y	Y (`fbcunn`)	#544
	CPU/GPU generic coding[3]	Y	[4]	Y	
	Multi GPU (data parallel)	Y	Y[7]	Y (`fbcunn`)	Y
	Multi GPU (model parallel)	Y	Y[8]	Y (`fbcunn`)	
Misc	Type checking	Y	Y	Y	N/A
	Model serialization	Y	Y (`pickle`)	Y	Y
	Caffe reference model	Y	[5]	Y (`loadcaffe`)	Y

Benchmarks

We are preparing for the benchmarks.

[1] They are also developing libgpuarray

[2] Stateful RNN is a type of RNN implementation that maintains states in the loops. It should enable us to use the states arbitrarily to update them.

[6] Also available in the Torch RNN package

[3] This row shows whether each array API supports unified codes for CPU and GPU.

[4] The array backend of Theano does not have compatible interface with NumPy, though most users write code on Theano variables, which is generic for CPU and GPU.

[7] Via Platoon

[8] Experimental as May 2016

[5] Depending on the frameworks.

CHAPTER 9

License

Copyright (c) 2015 Preferred Infrastructure, Inc.

Copyright (c) 2015 Preferred Networks, Inc.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

CuPy

The CuPy is designed based on NumPy's API. CuPy's source code and documents contain the original NumPy ones.

Copyright (c) 2005-2016, NumPy Developers.

All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: * Redistributions of source code must retain the above copyright

notice, this list of conditions and the following disclaimer.

- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

- Neither the name of the NumPy Developers nor the names of any contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

CHAPTER 10

Indices and tables

- genindex
- modindex
- search

Bibliography

[Graves2006] Alex Graves, Santiago Fernandez, Faustino Gomez, Jurgen Schmidhuber, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

[Graves2012] Alex Graves, Supervised Sequence Labelling with Recurrent Neural Networks

Python Module Index

C

chainer, 173
chainer.computational_graph, 189
chainer.cuda, 69
chainer.dataset, 58
chainer.datasets, 173
chainer.function, 168
chainer.function_hooks, 170
chainer.functions, 82
chainer.gradient_check, 79
chainer.initializer, 170
chainer.initializers, 171
chainer.iterators, 179
chainer.links, 140
chainer.links.caffe, 188
chainer.serializers, 166
chainer.testing, 81
chainer.training, 62
chainer.training.extensions, 180
chainer.training.triggers, 187
chainer.utils, 76
chainer.utils.type_check, 77
cupy, 203
cupy.random, 235
cupy.testing, 258

Index

A

absolute (in module cupy), 234
AbstractSerializer (class in chainer), 58
accumulate_grads() (chainer.Optimizer method), 53
accuracy() (in module chainer.functions), 110
AdaDelta (class in chainer.optimizers), 165
AdaGrad (class in chainer.optimizers), 165
Adam (class in chainer.optimizers), 165
add (in module cupy), 233
add() (chainer.DictSummary method), 76
add() (chainer.Summary method), 75
add_callback() (cupy.cuda.Stream method), 251
add_hook() (chainer.Function method), 45
add_hook() (chainer.Optimizer method), 53
add_link() (chainer.Chain method), 52
add_link() (chainer.ChainList method), 52
add_observer() (chainer.Reporter method), 74
add_observers() (chainer.Reporter method), 74
add_param() (chainer.Link method), 49
add_persistent() (chainer.Link method), 49
add_uninitialized_param() (chainer.Link method), 49
addgrad() (chainer.Variable method), 42
addgrads() (chainer.Link method), 49
aggregate_flags() (in module chainer.flag), 43
alloc() (in module cupy.cuda), 250
amax() (in module cupy), 243
amin() (in module cupy), 243
arange() (in module cupy), 208
arccos (in module cupy), 229
arccos() (in module chainer.functions), 121
arccosh (in module cupy), 230
arcsin (in module cupy), 229
arcsin() (in module chainer.functions), 121
arcsinh (in module cupy), 230
arctan (in module cupy), 229
arctan() (in module chainer.functions), 121
arctan2 (in module cupy), 229
arctanh (in module cupy), 230
argmax() (chainer.links.CRF1d method), 158
argmax() (cupy.ndarray method), 197
argmax() (in module chainer.functions), 122
argmax() (in module cupy), 242
argmax_crf1d() (in module chainer.functions), 115
argmin() (cupy.ndarray method), 197
argmin() (in module chainer.functions), 122
argmin() (in module cupy), 242
array() (in module cupy), 207
array_repr() (in module cupy), 224
array_split() (in module cupy), 216
array_str() (in module cupy), 224
asanyarray() (in module cupy), 208
asarray() (in module cupy), 207
ascontiguousarray() (in module cupy), 208
asfortranarray() (in module cupy), 214
asnumpy() (in module cupy), 202
assert_allclose() (in module chainer.testing), 81
assert_allclose() (in module cupy.testing), 258
assert_array_almost_equal() (in module cupy.testing), 258
assert_array_almost_equal_nulp() (in module cupy.testing), 259
assert_array_equal() (in module cupy.testing), 259
assert_array_less() (in module cupy.testing), 260
assert_array_list_equal() (in module cupy.testing), 259
assert_array_max_ulp() (in module cupy.testing), 259
astype() (cupy.ndarray method), 197
atleast_1d() (in module cupy), 212
atleast_2d() (in module cupy), 212
atleast_3d() (in module cupy), 213
AUTO (in module chainer), 43
average() (in module chainer.functions), 122
average_pooling_2d() (in module chainer.functions), 134
average_pooling_nd() (in module chainer.functions), 135

B

backward() (chainer.Function method), 45
backward() (chainer.Variable method), 42
backward_cpu() (chainer.Function method), 45
backward_gpu() (chainer.Function method), 45

backward_postprocess() (chainer.function.FunctionHook method), 169
backward_preprocess() (chainer.function.FunctionHook method), 169
batch_inv() (in module chainer.functions), 122
batch_l2_norm_squared() (in module chainer.functions), 123
batch_matmul() (in module chainer.functions), 123
batch_normalization() (in module chainer.functions), 132
BatchNormalization (class in chainer.links), 155
bernoulli_nll() (in module chainer.functions), 112
Bias (class in chainer.links), 141
bias() (in module chainer.functions), 123
Bilinear (class in chainer.links), 141
bilinear() (in module chainer.functions), 102
binary_accuracy() (in module chainer.functions), 110
BinaryHierarchicalSoftmax (class in chainer.links), 156
bincount() (in module cupy), 245
bitwise_and (in module cupy), 219
bitwise_or (in module cupy), 219
bitwise_xor (in module cupy), 219
black_out() (in module chainer.functions), 112
BlackOut (class in chainer.links), 157
broadcast (class in cupy), 213
broadcast() (in module chainer.functions), 92
broadcast_arrays() (in module cupy), 213
broadcast_to() (in module chainer.functions), 92
broadcast_to() (in module cupy), 214
build_computational_graph() (in module chainer.computational_graph), 189

C

c_ (in module cupy), 222
cache_or_load_file() (in module chainer.dataset), 61
cached_download() (in module chainer.dataset), 61
CaffeFunction (class in chainer.links.caffe), 188
call_hooks() (chainer.GradientMethod method), 56
call_hooks() (chainer.Optimizer method), 54
cast() (in module chainer.functions), 93
ceil (in module cupy), 231
ceil() (in module chainer.functions), 124
Chain (class in chainer), 51
chainer (module), 41, 173
chainer.computational_graph (module), 189
chainer.cuda (module), 69
chainer.dataset (module), 58
chainer.datasets (module), 173
chainer.function (module), 168
chainer.function_hooks (module), 170
chainer.functions (module), 82
chainer.gradient_check (module), 79
chainer.initializer (module), 170
chainer.initializers (module), 171
chainer.iterators (module), 179
chainer.links (module), 140
chainer.links.caffe (module), 188
chainer.serializers (module), 166
chainer.testing (module), 81
chainer.training (module), 62
chainer.training.extensions (module), 180
chainer.training.triggers (module), 187
chainer.utils (module), 73, 76, 278
chainer.utils.type_check (module), 77
ChainList (class in chainer), 52
check_backward() (in module chainer.gradient_check), 79
check_type_forward() (chainer.Function method), 46
children() (chainer.Link method), 50
choice() (cupy.random.RandomState method), 241
choice() (in module cupy.random), 235
cholesky() (in module cupy.linalg), 226
classification_summary() (in module chainer.functions), 111
Classifier (class in chainer.links), 159
clear_memo() (in module chainer.cuda), 72
clear_memo() (in module cupy), 254
cleargrad() (chainer.Variable method), 42
cleargrads() (chainer.Link method), 50
clip() (cupy.ndarray method), 198
clip() (in module chainer.functions), 124
clip() (in module cupy), 234
clip_grads() (chainer.Optimizer method), 54
clipped_relu() (in module chainer.functions), 83
collect_parameters() (chainer.FunctionSet method), 69
column_stack() (in module cupy), 215
ComputationalGraph (class in chainer.computational_graph), 191
compute_capability (cupy.cuda.Device attribute), 247
compute_grads_norm() (chainer.Optimizer method), 54
compute_mean() (chainer.DictSummary method), 76
compute_mean() (chainer.Summary method), 75
concat() (in module chainer.functions), 93
concat_examples() (in module chainer.dataset), 60
concatenate() (in module cupy), 215
connect_trainer() (chainer.training.Updater method), 64
connectionist_temporal_classification() (in module chainer.functions), 113
Constant (class in chainer.initializers), 171
contrastive() (in module chainer.functions), 113
convert_caffemodel_to_npz() (chainer.links.model.vision.resnet.ResNetLayers class method), 162
convert_caffemodel_to_npz() (chainer.links.VGG16Layers class method), 161
Convolution2D (class in chainer.links), 142
convolution_2d() (in module chainer.functions), 103
convolution_nd() (in module chainer.functions), 104

ConvolutionND (class in chainer.links), 143
copy() (chainer.Link method), 50
copy() (cupy.ndarray method), 198
copy() (in module chainer.cuda), 70
copy() (in module chainer.functions), 94
copy() (in module cupy), 208
copy_from() (cupy.cuda.MemoryPointer method), 248
copy_from_async() (cupy.cuda.MemoryPointer method), 248
copy_from_device() (cupy.cuda.MemoryPointer method), 249
copy_from_device_async() (cupy.cuda.MemoryPointer method), 249
copy_from_host() (cupy.cuda.MemoryPointer method), 249
copy_from_host_async() (cupy.cuda.MemoryPointer method), 249
copy_parameters_from() (chainer.FunctionSet method), 69
copy_to_host() (cupy.cuda.MemoryPointer method), 249
copy_to_host_async() (cupy.cuda.MemoryPointer method), 249
copydata() (chainer.Variable method), 42
copyparams() (chainer.Link method), 50
copysign (in module cupy), 232
copyto() (in module cupy), 210
cos (in module cupy), 229
cos() (in module chainer.functions), 124
cosh (in module cupy), 230
cosh() (in module chainer.functions), 125
count_nonzero() (in module cupy), 243
create_huffman_tree() (chainer.links.BinaryHierarchicalSoftmax static method), 157
crelu() (in module chainer.functions), 83
CRF1d (class in chainer.links), 157
crf1d() (in module chainer.functions), 114
cross_covariance() (in module chainer.functions), 116
cstruct (cupy.ndarray attribute), 198
cublas_handle (cupy.cuda.Device attribute), 247
cupy (module), 193, 203, 204, 257
cupy.random (module), 235
cupy.testing (module), 258
cusolver_handle (cupy.cuda.Device attribute), 247

D

DatasetMixin (class in chainer.dataset), 59
debug_print() (chainer.Variable method), 42
DebugMode (class in chainer), 68
Deconvolution2D (class in chainer.links), 143
deconvolution_2d() (in module chainer.functions), 105
deconvolution_nd() (in module chainer.functions), 105
DeconvolutionND (class in chainer.links), 144
default_name (chainer.training.Extension attribute), 67
deg2rad (in module cupy), 229
degrees (in module cupy), 230
delete_hook() (chainer.Function method), 46
depth2space() (in module chainer.functions), 94
Deserializer (class in chainer), 58
Device (class in cupy.cuda), 247
device (cupy.ndarray attribute), 198
diag() (in module cupy), 210
diagflat() (in module cupy), 210
diagonal() (cupy.ndarray method), 198
diagonal() (in module cupy), 220
DictDataset (class in chainer.datasets), 174
DictionarySerializer (class in chainer.serializers), 166
DictSummary (class in chainer), 76
dilated_convolution_2d() (in module chainer.functions), 106
DilatedConvolution2D (class in chainer.links), 145
divide (in module cupy), 233
done (cupy.cuda.Event attribute), 252
done (cupy.cuda.Stream attribute), 251
dot() (cupy.ndarray method), 198
dot() (in module cupy), 225
dropout() (in module chainer.functions), 131
dsplit() (in module cupy), 217
dstack() (in module chainer.functions), 95
dstack() (in module cupy), 216
dump() (chainer.computational_graph.ComputationalGraph method), 191
dump() (cupy.ndarray method), 198
dump_graph() (in module chainer.training.extensions), 181
dumps() (cupy.ndarray method), 198

E

elapsed_time (chainer.training.Trainer attribute), 63
elementwise() (in module chainer.cuda), 72
ElementwiseKernel (class in cupy), 257
elu() (in module chainer.functions), 84
embed_id() (in module chainer.functions), 107
EmbedID (class in chainer.links), 146
empty() (in module cupy), 204
empty_like() (in module cupy), 204
equal (in module cupy), 228
eval() (chainer.utils.type_check.Expr method), 78
evaluate() (chainer.training.extensions.Evaluator method), 182
Evaluator (class in chainer.training.extensions), 181
Event (class in cupy.cuda), 252
exp (in module cupy), 231
exp() (in module chainer.functions), 125
exp2 (in module cupy), 231
expand_dims() (in module chainer.functions), 95
expand_dims() (in module cupy), 214
expect() (in module chainer.utils.type_check), 78
experimental() (in module chainer.utils), 76

expm1 (in module cupy), 231
ExponentialShift (class in chainer.training.extensions), 182
Expr (class in chainer.utils.type_check), 77
extend() (chainer.training.Trainer method), 63
Extension (class in chainer.training), 66
extract() (chainer.links.model.vision.resnet.ResNetLayers method), 162
extract() (chainer.links.VGG16Layers method), 161
eye() (in module cupy), 205

F

fill() (cupy.ndarray method), 199
fill_diagonal() (in module cupy), 221
finalize() (chainer.dataset.Iterator method), 59
finalize() (chainer.training.Extension method), 67
finalize() (chainer.training.Updater method), 65
fixed_batch_normalization() (in module chainer.functions), 133
Flag (class in chainer), 43
flags (cupy.ndarray attribute), 199
flatnonzero() (in module cupy), 243
flatten() (cupy.ndarray method), 199
flatten() (in module chainer.functions), 95
flip() (in module cupy), 218
fliplr() (in module chainer.functions), 95
fliplr() (in module cupy), 218
flipud() (in module chainer.functions), 96
flipud() (in module cupy), 218
floor (in module cupy), 231
floor() (in module chainer.functions), 125
floor_divide (in module cupy), 233
fmax (in module cupy), 235
fmin (in module cupy), 235
fmod (in module cupy), 233
fmod() (in module chainer.functions), 125
for_all_dtypes() (in module cupy.testing), 263
for_all_dtypes_combination() (in module cupy.testing), 266
for_CF_orders() (in module cupy.testing), 267
for_dtypes() (in module cupy.testing), 263
for_dtypes_combination() (in module cupy.testing), 265
for_float_dtypes() (in module cupy.testing), 264
for_int_dtypes() (in module cupy.testing), 265
for_int_dtypes_combination() (in module cupy.testing), 266
for_orders() (in module cupy.testing), 267
for_signed_dtypes() (in module cupy.testing), 264
for_signed_dtypes_combination() (in module cupy.testing), 266
for_unsigned_dtypes() (in module cupy.testing), 265
for_unsigned_dtypes_combination() (in module cupy.testing), 266
force_backprop_mode() (in module chainer), 47
forget() (in module chainer.functions), 139
forward() (chainer.Function method), 46
forward_cpu() (chainer.Function method), 46
forward_gpu() (chainer.Function method), 46
forward_postprocess() (chainer.function.FunctionHook method), 169
forward_preprocess() (chainer.function.FunctionHook method), 169
free_all_blocks() (cupy.cuda.MemoryPool method), 250
free_all_free() (cupy.cuda.MemoryPool method), 250
frexp (in module cupy), 232
full() (in module cupy), 206
full_like() (in module cupy), 207
Function (class in chainer), 44
FunctionHook (class in chainer.function), 168
FunctionSet (class in chainer), 68

G

gaussian() (in module chainer.functions), 131
gaussian_kl_divergence() (in module chainer.functions), 116
gaussian_nll() (in module chainer.functions), 117
get() (cupy.ndarray method), 199
get_all_iterators() (chainer.training.extensions.Evaluator method), 182
get_all_optimizers() (chainer.training.Updater method), 65
get_all_targets() (chainer.training.extensions.Evaluator method), 182
get_array_module() (in module chainer.cuda), 72
get_array_module() (in module cupy), 247
get_cifar10() (in module chainer.datasets), 178
get_cifar100() (in module chainer.datasets), 178
get_cross_validation_datasets() (in module chainer.datasets), 176
get_cross_validation_datasets_random() (in module chainer.datasets), 176
get_current_reporter() (in module chainer), 74
get_dataset_root() (in module chainer.dataset), 61
get_device() (in module chainer.cuda), 69
get_device_from_array() (in module chainer.cuda), 70
get_device_from_id() (in module chainer.cuda), 70
get_elapsed_time() (in module cupy.cuda), 252
get_example() (chainer.dataset.DatasetMixin method), 59
get_extension() (chainer.training.Trainer method), 64
get_item() (in module chainer.functions), 96
get_iterator() (chainer.training.extensions.Evaluator method), 182
get_iterator() (chainer.training.StandardUpdater method), 66
get_max_workspace_size() (in module chainer.cuda), 72
get_mnist() (in module chainer.datasets), 177
get_optimizer() (chainer.training.Updater method), 65
get_ptb_words() (in module chainer.datasets), 179

get_ptb_words_vocabulary() (in module chainer.datasets), 179
get_random_state() (in module cupy.random), 240
get_target() (chainer.training.extensions.Evaluator method), 182
get_trigger() (in module chainer.training), 67
GlorotNormal (class in chainer.initializers), 171
GlorotUniform (class in chainer.initializers), 173
GradientClipping (class in chainer.optimizer), 57
GradientMethod (class in chainer), 56
GradientNoise (class in chainer.optimizer), 57
gradients (chainer.FunctionSet attribute), 69
greater (in module cupy), 228
greater_equal (in module cupy), 228
GRU (class in chainer.links), 146
gumbel() (in module cupy.random), 238

H

hard_sigmoid() (in module chainer.functions), 84
has_uninitialized_params (chainer.Link attribute), 50
HDF5Deserializer (class in chainer.serializers), 167
HDF5Serializer (class in chainer.serializers), 167
HeNormal (class in chainer.initializers), 172
HeUniform (class in chainer.initializers), 173
Highway (class in chainer.links), 147
hinge() (in module chainer.functions), 117
hsplit() (in module cupy), 217
hstack() (in module chainer.functions), 96
hstack() (in module cupy), 215
huber_loss() (in module chainer.functions), 118
hypot (in module cupy), 229

I

Identity (class in chainer.initializers), 171
identity() (in module chainer.functions), 125
identity() (in module cupy), 205
ImageDataset (class in chainer.datasets), 176
Inception (class in chainer.links), 148
InceptionBN (class in chainer.links), 149
init_state() (chainer.Optimizer method), 54
init_state_cpu() (chainer.Optimizer method), 54
init_state_gpu() (chainer.Optimizer method), 54
init_weight() (in module chainer), 173
initialize() (in module cupy.cuda.profiler), 253
Initializer (class in chainer.initializer), 171
inner() (in module cupy), 225
interval() (cupy.random.RandomState method), 241
IntervalTrigger (class in chainer.training.triggers), 187
inv() (in module chainer.functions), 126
invert (in module cupy), 220
is_debug() (in module chainer), 68
isfinite (in module cupy), 227
isinf (in module cupy), 227
isnan (in module cupy), 227
itemsize (cupy.ndarray attribute), 199
Iterator (class in chainer.dataset), 59
ix_() (in module cupy), 221

L

label (chainer.Function attribute), 47
label (chainer.Variable attribute), 42
LabeledImageDataset (class in chainer.datasets), 177
Lasso (class in chainer.optimizer), 57
LayerNormalization (class in chainer.links), 156
ldexp (in module cupy), 232
leaky_relu() (in module chainer.functions), 85
LeCunUniform (class in chainer.initializers), 172
left_shift (in module cupy), 220
less (in module cupy), 228
less_equal (in module cupy), 228
Linear (class in chainer.links), 149
linear() (in module chainer.functions), 108
linear_interpolate() (in module chainer.functions), 126
LinearShift (class in chainer.training.extensions), 183
Link (class in chainer), 48
links() (chainer.Link method), 50
linspace() (in module cupy), 208
load() (chainer.Deserializer method), 58
load() (in module cupy), 223
load_hdf5() (in module chainer.serializers), 167
load_npz() (in module chainer.serializers), 167
local_function_hooks (chainer.Function attribute), 47
local_response_normalization() (in module chainer.functions), 133
log (chainer.training.extensions.LogReport attribute), 184
log (in module cupy), 231
log() (in module chainer.functions), 126
log10 (in module cupy), 232
log10() (in module chainer.functions), 126
log1p (in module cupy), 232
log1p() (in module chainer.functions), 126
log2 (in module cupy), 232
log2() (in module chainer.functions), 127
log_softmax() (in module chainer.functions), 86
logaddexp (in module cupy), 232
logaddexp2 (in module cupy), 232
logical_and (in module cupy), 227
logical_not (in module cupy), 227
logical_or (in module cupy), 227
logical_xor (in module cupy), 228
lognormal() (cupy.random.RandomState method), 241
lognormal() (in module cupy.random), 238
LogReport (class in chainer.training.extensions), 183
logspace() (in module cupy), 209
logsumexp() (in module chainer.functions), 127
LSTM (class in chainer.links), 150
lstm() (in module chainer.functions), 86

M

make_extension() (in module chainer.training), 67
make_statistics() (chainer.DictSummary method), 76
make_statistics() (chainer.Summary method), 76
malloc() (cupy.cuda.MemoryPool method), 250
ManualScheduleTrigger (class in chainer.training.triggers), 187
Mark() (in module cupy.cuda.nvtx), 253
MarkC() (in module cupy.cuda.nvtx), 253
matmul() (in module chainer.functions), 127
max() (cupy.ndarray method), 199
max() (in module chainer.functions), 128
max_pooling_2d() (in module chainer.functions), 135
max_pooling_nd() (in module chainer.functions), 136
maximum (in module cupy), 234
maximum() (in module chainer.functions), 128
Maxout (class in chainer.links), 158
maxout() (in module chainer.functions), 87
MaxValueTrigger (class in chainer.training.triggers), 187
mean() (cupy.ndarray method), 199
mean() (in module cupy), 244
mean_absolute_error() (in module chainer.functions), 118
mean_squared_error() (in module chainer.functions), 118
memoize() (in module chainer.cuda), 71
memoize() (in module cupy), 254
Memory (class in cupy.cuda), 248
MemoryPointer (class in cupy.cuda), 248
MemoryPool (class in cupy.cuda), 250
memset() (cupy.cuda.MemoryPointer method), 249
memset_async() (cupy.cuda.MemoryPointer method), 250
meshgrid() (in module cupy), 209
min() (cupy.ndarray method), 199
min() (in module chainer.functions), 128
minimum (in module cupy), 235
minimum() (in module chainer.functions), 128
MinValueTrigger (class in chainer.training.triggers), 188
MLPConvolution2D (class in chainer.links), 151
mod (in module cupy), 234
modf (in module cupy), 234
MomentumSGD (class in chainer.optimizers), 165
multiply (in module cupy), 233
MultiprocessIterator (class in chainer.iterators), 180

N

n_free_blocks() (cupy.cuda.MemoryPool method), 251
n_step_lstm() (in module chainer.functions), 109
namedlinks() (chainer.Link method), 50
namedparams() (chainer.Link method), 50
nanmax() (in module cupy), 244
nanmin() (in module cupy), 243
nbytes (cupy.ndarray attribute), 200
ndarray (class in cupy), 197
ndim (cupy.ndarray attribute), 200
negative (in module cupy), 233
negative_sampling() (in module chainer.functions), 118
NegativeSampling (class in chainer.links), 159
NesterovAG (class in chainer.optimizers), 165
new_epoch() (chainer.Optimizer method), 55
next() (chainer.dataset.Iterator method), 60
nextafter (in module cupy), 233
no_backprop_mode() (in module chainer), 47
nonzero() (cupy.ndarray method), 200
nonzero() (in module cupy), 243
Normal (class in chainer.initializers), 171
normal() (cupy.random.RandomState method), 241
normal() (in module cupy.random), 239
normalize() (in module chainer.functions), 134
not_equal (in module cupy), 228
NpzDeserializer (class in chainer.serializers), 166
NStepLSTM (class in chainer.links), 152
numerical_grad() (in module chainer.gradient_check), 80
numpy_cupy_allclose() (in module cupy.testing), 260
numpy_cupy_array_almost_equal() (in module cupy.testing), 261
numpy_cupy_array_almost_equal_nulp() (in module cupy.testing), 261
numpy_cupy_array_equal() (in module cupy.testing), 262
numpy_cupy_array_less() (in module cupy.testing), 263
numpy_cupy_array_list_equal() (in module cupy.testing), 262
numpy_cupy_array_max_ulp() (in module cupy.testing), 262
numpy_cupy_raises() (in module cupy.testing), 263

O

OFF (in module chainer), 43
ON (in module chainer), 43
One() (in module chainer.initializers), 171
ones() (in module cupy), 205
ones_like() (in module cupy), 205
Optimizer (class in chainer), 53
Orthogonal (class in chainer.initializers), 172
outer() (in module cupy), 225

P

pad() (in module chainer.functions), 96
ParallelUpdater (class in chainer.training), 66
Parameter (class in chainer.links), 165
parameters (chainer.FunctionSet attribute), 69
params() (chainer.Link method), 51
permutate() (in module chainer.functions), 97
PlotReport (class in chainer.training.extensions), 185
power (in module cupy), 233
predict() (chainer.links.model.vision.resnet.ResNetLayers method), 163
predict() (chainer.links.VGG16Layers method), 161
PReLU (class in chainer.links), 158

prelu() (in module chainer.functions), 88
prepare() (chainer.Optimizer method), 55
prepare() (in module chainer.links.model.vision.resnet), 164
prepare() (in module chainer.links.model.vision.vgg), 161
PrintHook (class in chainer.function_hooks), 170
PrintReport (class in chainer.training.extensions), 186
prod() (cupy.ndarray method), 200
prod() (in module cupy), 231
profile() (in module cupy.cuda), 253
ProgressBar (class in chainer.training.extensions), 186

R

r2_score() (in module chainer.functions), 111
r_ (in module cupy), 222
rad2deg (in module cupy), 229
radians (in module cupy), 230
rand() (cupy.random.RandomState method), 241
rand() (in module cupy.random), 236
randint() (in module cupy.random), 236
randn() (cupy.random.RandomState method), 241
randn() (in module cupy.random), 236
random() (in module cupy.random), 237
random_integers() (in module cupy.random), 237
random_sample() (cupy.random.RandomState method), 241
random_sample() (in module cupy.random), 237
RandomState (class in cupy.random), 240
ranf() (in module cupy.random), 237
RangePop() (in module cupy.cuda.nvtx), 254
RangePush() (in module cupy.cuda.nvtx), 253
RangePushC() (in module cupy.cuda.nvtx), 254
ravel() (cupy.ndarray method), 200
ravel() (in module cupy), 211
reallocate_cleared_grads() (chainer.GradientMethod method), 56
reciprocal (in module cupy), 234
record() (cupy.cuda.Event method), 252
record() (cupy.cuda.Stream method), 251
reduce() (in module chainer.cuda), 72
reduced_view() (cupy.ndarray method), 200
ReductionKernel (class in cupy), 257
relu() (in module chainer.functions), 89
remainder (in module cupy), 234
remove_hook() (chainer.Optimizer method), 55
repeat() (cupy.ndarray method), 200
repeat() (in module cupy), 217
report() (chainer.Reporter method), 74
report() (in module chainer), 74
report_scope() (in module chainer), 75
Reporter (class in chainer), 73
reset_state() (chainer.links.LSTM method), 151
reset_state() (chainer.links.StatefulPeepholeLSTM method), 154
reshape() (chainer.Variable method), 42
reshape() (cupy.ndarray method), 200
reshape() (in module chainer.functions), 97
reshape() (in module cupy), 211
resize_images() (in module chainer.functions), 98
ResNet101Layers (class in chainer.links), 163
ResNet152Layers (class in chainer.links), 164
ResNet50Layers (class in chainer.links), 163
ResNetLayers (class in chainer.links.model.vision.resnet), 162
right_shift (in module cupy), 220
rint (in module cupy), 231
RMSprop (class in chainer.optimizers), 165
RMSpropGraves (class in chainer.optimizers), 165
roi_pooling_2d() (in module chainer.functions), 136
roll() (in module cupy), 218
rollaxis() (in module chainer.functions), 98
rollaxis() (in module cupy), 211
rot90() (in module cupy), 219
rsqrt() (in module chainer.functions), 129
run() (chainer.training.Trainer method), 64

S

sample() (chainer.utils.WalkerAlias method), 73
sample() (in module cupy.random), 238
save() (chainer.Serializer method), 58
save() (in module cupy), 223
save_hdf5() (in module chainer.serializers), 167
save_npz() (in module chainer.serializers), 166
savez() (in module cupy), 223
savez_compressed() (in module cupy), 224
Scale (class in chainer.links), 152
scale() (in module chainer.functions), 129
scatter_add() (cupy.ndarray method), 201
scatter_add() (in module cupy), 246
scope() (chainer.Reporter method), 74
seed() (cupy.random.RandomState method), 241
seed() (in module cupy.random), 240
select_item() (in module chainer.functions), 98
separate() (in module chainer.functions), 99
SerialIterator (class in chainer.iterators), 180
serialize() (chainer.dataset.Iterator method), 60
serialize() (chainer.Link method), 51
serialize() (chainer.Optimizer method), 55
serialize() (chainer.training.Extension method), 67
serialize() (chainer.training.Updater method), 65
Serializer (class in chainer), 58
set() (cupy.ndarray method), 201
set_allocator() (in module cupy.cuda), 250
set_creator() (chainer.Variable method), 42
set_dataset_root() (in module chainer.dataset), 61
set_debug() (in module chainer), 68
set_max_workspace_size() (in module chainer.cuda), 72
set_state() (chainer.links.LSTM method), 151

setup() (chainer.Optimizer method), 55
SGD (class in chainer.optimizers), 166
shape (cupy.ndarray attribute), 201
sigmoid() (in module chainer.functions), 89
sigmoid_cross_entropy() (in module chainer.functions), 119
sign (in module cupy), 234
signbit (in module cupy), 232
sin (in module cupy), 229
sin() (in module chainer.functions), 129
sinh (in module cupy), 230
sinh() (in module chainer.functions), 130
size() (chainer.utils.type_check.TypeInfoTuple method), 78
slstm() (in module chainer.functions), 90
SMORMS3 (class in chainer.optimizers), 166
snapshot() (in module chainer.training.extensions), 184
snapshot_object() (in module chainer.training.extensions), 185
softmax() (in module chainer.functions), 91
softmax_cross_entropy() (in module chainer.functions), 119
softplus() (in module chainer.functions), 91
space2depth() (in module chainer.functions), 99
spatial_pyramid_pooling_2d() (in module chainer.functions), 137
split() (in module cupy), 216
split_axis() (in module chainer.functions), 99
split_dataset() (in module chainer.datasets), 175
split_dataset_random() (in module chainer.datasets), 175
sqrt (in module cupy), 234
sqrt() (in module chainer.functions), 130
square (in module cupy), 234
square() (in module chainer.functions), 130
squared_difference() (in module chainer.functions), 130
squeeze() (cupy.ndarray method), 201
squeeze() (in module chainer.functions), 100
squeeze() (in module cupy), 214
stack() (in module chainer.functions), 100
stack() (in module cupy), 216
standard_normal() (cupy.random.RandomState method), 242
standard_normal() (in module cupy.random), 239
StandardUpdater (class in chainer.training), 65
start() (in module cupy.cuda.profiler), 253
start_finetuning() (chainer.links.BatchNormalization method), 156
StatefulGRU (class in chainer.links), 153
StatefulPeepholeLSTM (class in chainer.links), 153
StatelessLSTM (class in chainer.links), 154
std() (cupy.ndarray method), 201
std() (in module cupy), 245
stop() (in module cupy.cuda.profiler), 253
Stream (class in cupy.cuda), 251
strides (cupy.ndarray attribute), 201
SubDataset (class in chainer.datasets), 174
subtract (in module cupy), 233
sum() (cupy.ndarray method), 201
sum() (in module chainer.functions), 131
sum() (in module cupy), 231
Summary (class in chainer), 75
swapaxes() (cupy.ndarray method), 201
swapaxes() (in module chainer.functions), 100
swapaxes() (in module cupy), 212
synchronize() (cupy.cuda.Device method), 248
synchronize() (cupy.cuda.Event method), 252
synchronize() (cupy.cuda.Stream method), 252

T

T (cupy.ndarray attribute), 197
take() (cupy.ndarray method), 201
take() (in module cupy), 220
tan (in module cupy), 229
tan() (in module chainer.functions), 131
tanh (in module cupy), 230
tanh() (in module chainer.functions), 91
tensordot() (in module cupy), 226
tile() (in module chainer.functions), 101
tile() (in module cupy), 217
TimerHook (class in chainer.function_hooks), 170
to_cpu() (chainer.Link method), 51
to_cpu() (chainer.Variable method), 42
to_cpu() (in module chainer.cuda), 71
to_device() (in module chainer.dataset), 60
to_gpu() (chainer.Link method), 51
to_gpu() (chainer.utils.WalkerAlias method), 73
to_gpu() (chainer.Variable method), 43
to_gpu() (in module chainer.cuda), 71
tofile() (cupy.ndarray method), 202
tolist() (cupy.ndarray method), 202
total_time() (chainer.function_hooks.TimerHook method), 170
trace() (cupy.ndarray method), 202
trace() (in module cupy), 226
Trainer (class in chainer.training), 62
transpose() (chainer.Variable method), 43
transpose() (cupy.ndarray method), 202
transpose() (in module chainer.functions), 101
transpose() (in module cupy), 212
transpose_sequence() (in module chainer.functions), 101
triplet() (in module chainer.functions), 120
true_divide (in module cupy), 233
trunc (in module cupy), 231
TupleDataset (class in chainer.datasets), 174
TypeInfo (class in chainer.utils.type_check), 78
TypeInfoTuple (class in chainer.utils.type_check), 78
types (cupy.ufunc attribute), 203

U

ufunc (class in cupy), 203
unary_math_function_unittest() (in module chainer.testing), 81
unchain() (chainer.Function method), 47
unchain_backward() (chainer.Variable method), 43
Uniform (class in chainer.initializers), 172
uniform() (cupy.random.RandomState method), 242
uniform() (in module cupy.random), 239
unpooling_2d() (in module chainer.functions), 138
update() (chainer.GradientMethod method), 56
update() (chainer.Optimizer method), 55
update() (chainer.training.Updater method), 65
update_one() (chainer.GradientMethod method), 56
update_one_cpu() (chainer.GradientMethod method), 56
update_one_gpu() (chainer.GradientMethod method), 57
Updater (class in chainer.training), 64
upsampling_2d() (in module chainer.functions), 138
use() (cupy.cuda.Device method), 248
use_cleargrads() (chainer.GradientMethod method), 57

V

var() (cupy.ndarray method), 202
var() (in module cupy), 245
Variable (class in chainer), 41
vdot() (in module cupy), 225
VGG16Layers (class in chainer.links), 160
view() (cupy.ndarray method), 202
vsplit() (in module cupy), 217
vstack() (in module chainer.functions), 101
vstack() (in module cupy), 215

W

wait_event() (cupy.cuda.Stream method), 252
WalkerAlias (class in chainer.utils), 73
weight_decay() (chainer.Optimizer method), 55
WeightDecay (class in chainer.optimizer), 57
where() (in module chainer.functions), 102
where() (in module cupy), 243

X

xp (chainer.Link attribute), 51

Z

Zero() (in module chainer.initializers), 171
zero_grads() (chainer.Optimizer method), 56
zerograd() (chainer.Variable method), 43
zerograds() (chainer.Link method), 51
zeros() (in module cupy), 206
zeros_like() (in module cupy), 206