Convolutional neural network layers

In this notebook, we will build the convolutional neural network layers. This will be followed by a spatial batchnorm, and then in the final notebook of this assignment, we will train a CNN to further improve the validation accuracy on CIFAR-10.

CS231n has built a solid API for building these modular frameworks and training them, and we will use their very well implemented framework as opposed to "reinventing the wheel." This includes using their Solver, various utility functions, their layer structure, and their implementation of fast CNN layers. This also includes nndl.fc_net, nndl.layers, and nndl.layer_utils. As in prior assignments, we thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class (cs231n.stanford.edu).

In [1]:
```
## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.conv_layers import *
from cs231n.data_utils import get_CIFAR10_data
from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradien
t_array
from cs231n.solver import Solver

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
  """ returns relative error """
  return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

Implementing CNN layers

Just as we implemented modular layers for fully connected networks, batch normalization, and dropout, we'll want to implement modular layers for convolutional neural networks. These layers are in `nndl/conv_layers.py` .

Convolutional forward pass

Begin by implementing a naive version of the forward pass of the CNN that uses `for` loops. This function is `conv_forward_naive` in `nndl/conv_layers.py` . Don't worry about efficiency of implementation. Later on, we provide a fast implementation of these layers. This version ought to test your understanding of convolution. In our implementation, there is a triple `for` loop.

After you implement `conv_forward_naive` , test your implementation by running the cell below.

In [2]:
```python
x_shape = (2, 3, 4, 4)
w_shape = (3, 3, 4, 4)
x = np.linspace(-0.1, 0.5, num=np.prod(x_shape)).reshape(x_shape)
w = np.linspace(-0.2, 0.3, num=np.prod(w_shape)).reshape(w_shape)
b = np.linspace(-0.1, 0.2, num=3)

conv_param = {'stride': 2, 'pad': 1}
out, _ = conv_forward_naive(x, w, b, conv_param)
correct_out = np.array([[[[-0.08759809, -0.10987781],
                           [-0.18387192, -0.2109216 ]],
                          [[ 0.21027089,  0.21661097],
                           [ 0.22847626,  0.23004637]],
                          [[ 0.50813986,  0.54309974],
                           [ 0.64082444,  0.67101435]]],
                         [[[-0.98053589, -1.03143541],
                           [-1.19128892, -1.24695841]],
                          [[ 0.69108355,  0.66880383],
                           [ 0.59480972,  0.56776003]],
                          [[ 2.36270298,  2.36904306],
                           [ 2.38090835,  2.38247847]]]])

# Compare your output to ours; difference should be around 1e-8
print('Testing conv_forward_naive')
print('difference: ', rel_error(out, correct_out))
```

```
Testing conv_forward_naive
difference:  2.2121476417505994e-08
```

Convolutional backward pass

Now, implement a naive version of the backward pass of the CNN. The function is `conv_backward_naive` in `nndl/conv_layers.py` . Don't worry about efficiency of implementation. Later on, we provide a fast implementation of these layers. This version ought to test your understanding of convolution. In our implementation, there is a quadruple `for` loop.

After you implement `conv_backward_naive` , test your implementation by running the cell below.

In [3]:
```
x = np.random.randn(4, 3, 5, 5)
w = np.random.randn(2, 3, 3, 3)
b = np.random.randn(2,)
dout = np.random.randn(4, 2, 5, 5)
conv_param = {'stride': 1, 'pad': 1}

out, cache = conv_forward_naive(x,w,b,conv_param)

dx_num = eval_numerical_gradient_array(lambda x: conv_forward_naive(x, w, b, conv
_param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_forward_naive(x, w, b, conv
_param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_forward_naive(x, w, b, conv
_param)[0], b, dout)

out, cache = conv_forward_naive(x, w, b, conv_param)
dx, dw, db = conv_backward_naive(dout, cache)

# Your errors should be around 1e-9'
print('Testing conv_backward_naive function')
print('dx error: ', rel_error(dx, dx_num))
print('dw error: ', rel_error(dw, dw_num))
print('db error: ', rel_error(db, db_num))
```

```
Testing conv_backward_naive function
dx error:  8.658306636445945e-10
dw error:  5.524593469842107e-10
db error:  2.5333554482976246e-11
```

Max pool forward pass

In this section, we will implement the forward pass of the max pool. The function is `max_pool_forward_naive` in `nndl/conv_layers.py` . Do not worry about the efficiency of implementation.

After you implement `max_pool_forward_naive` , test your implementation by running the cell below.

In [4]:
```
x_shape = (2, 3, 4, 4)
x = np.linspace(-0.3, 0.4, num=np.prod(x_shape)).reshape(x_shape)
pool_param = {'pool_width': 2, 'pool_height': 2, 'stride': 2}

out, _ = max_pool_forward_naive(x, pool_param)

correct_out = np.array([[[[-0.26315789, -0.24842105],
                          [-0.20421053, -0.18947368]],
                         [[-0.14526316, -0.13052632],
                          [-0.08631579, -0.07157895]],
                         [[-0.02736842, -0.01263158],
                          [ 0.03157895,  0.04631579]]],
                        [[[ 0.09052632,  0.10526316],
                          [ 0.14947368,  0.16421053]],
                         [[ 0.20842105,  0.22315789],
                          [ 0.26736842,  0.28210526]],
                         [[ 0.32631579,  0.34105263],
                          [ 0.38526316,  0.4       ]]]])

# Compare your output with ours. Difference should be around 1e-8.
print('Testing max_pool_forward_naive function:')
print('difference: ', rel_error(out, correct_out))
```

```
Testing max_pool_forward_naive function:
difference:  4.1666665157267834e-08
```

Max pool backward pass

In this section, you will implement the backward pass of the max pool. The function is `max_pool_backward_naive` in `nndl/conv_layers.py` . Do not worry about the efficiency of implementation.

After you implement `max_pool_backward_naive` , test your implementation by running the cell below.

In [5]:
```
x = np.random.randn(3, 2, 8, 8)
dout = np.random.randn(3, 2, 4, 4)
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

dx_num = eval_numerical_gradient_array(lambda x: max_pool_forward_naive(x, pool_p
aram)[0], x, dout)

out, cache = max_pool_forward_naive(x, pool_param)
dx = max_pool_backward_naive(dout, cache)

# Your error should be around 1e-12
print('Testing max_pool_backward_naive function:')
print('dx error: ', rel_error(dx, dx_num))
```

```
Testing max_pool_backward_naive function:
dx error:  3.2756157171995605e-12
```

Fast implementation of the CNN layers

Implementing fast versions of the CNN layers can be difficult. We will provide you with the fast layers implemented by cs231n. They are provided in `cs231n/fast_layers.py`.

The fast convolution implementation depends on a Cython extension; to compile it you need to run the following from the `cs231n` directory:

```
python setup.py build_ext --inplace
```

NOTE: The fast implementation for pooling will only perform optimally if the pooling regions are non-overlapping and tile the input. If these conditions are not met then the fast pooling implementation will not be much faster than the naive implementation.

You can compare the performance of the naive and fast versions of these layers by running the cell below.

You should see pretty drastic speedups in the implementation of these layers. On our machine, the forward pass speeds up by 17x and the backward pass speeds up by 840x. Of course, these numbers will vary from machine to machine, as well as on your precise implementation of the naive layers.

In [6]:
```
from cs231n.fast_layers import conv_forward_fast, conv_backward_fast
from time import time

x = np.random.randn(100, 3, 31, 31)
w = np.random.randn(25, 3, 3, 3)
b = np.random.randn(25,)
dout = np.random.randn(100, 25, 16, 16)
conv_param = {'stride': 2, 'pad': 1}

t0 = time()
out_naive, cache_naive = conv_forward_naive(x, w, b, conv_param)
t1 = time()
out_fast, cache_fast = conv_forward_fast(x, w, b, conv_param)
t2 = time()

print('Testing conv_forward_fast:')
print('Naive: %fs' % (t1 - t0))
print('Fast: %fs' % (t2 - t1))
print('Speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('Difference: ', rel_error(out_naive, out_fast))

t0 = time()
dx_naive, dw_naive, db_naive = conv_backward_naive(dout, cache_naive)
t1 = time()
dx_fast, dw_fast, db_fast = conv_backward_fast(dout, cache_fast)
t2 = time()

print('\nTesting conv_backward_fast:')
print('Naive: %fs' % (t1 - t0))
print('Fast: %fs' % (t2 - t1))
print('Speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('dx difference: ', rel_error(dx_naive, dx_fast))
print('dw difference: ', rel_error(dw_naive, dw_fast))
print('db difference: ', rel_error(db_naive, db_fast))
```

```
Testing conv_forward_fast:
Naive: 7.150056s
Fast: 0.035052s
Speedup: 203.986682x
Difference:  7.384927040715952e-11

Testing conv_backward_fast:
Naive: 12.509643s
Fast: 0.022429s
Speedup: 557.744393x
dx difference:  3.430777717863996e-11
dw difference:  3.791261763720037e-13
db difference:  7.637275700944883e-15
```

In [7]:
```python
from cs231n.fast_layers import max_pool_forward_fast, max_pool_backward_fast

x = np.random.randn(100, 3, 32, 32)
dout = np.random.randn(100, 3, 16, 16)
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

t0 = time()
out_naive, cache_naive = max_pool_forward_naive(x, pool_param)
t1 = time()
out_fast, cache_fast = max_pool_forward_fast(x, pool_param)
t2 = time()

print('Testing pool_forward_fast:')
print('Naive: %fs' % (t1 - t0))
print('fast: %fs' % (t2 - t1))
print('speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('difference: ', rel_error(out_naive, out_fast))

t0 = time()
dx_naive = max_pool_backward_naive(dout, cache_naive)
t1 = time()
dx_fast = max_pool_backward_fast(dout, cache_fast)
t2 = time()

print('\nTesting pool_backward_fast:')
print('Naive: %fs' % (t1 - t0))
print('speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('dx difference: ', rel_error(dx_naive, dx_fast))
```

```
Testing pool_forward_fast:
Naive: 0.513833s
fast: 0.002673s
speedup: 192.237267x
difference:  0.0

Testing pool_backward_fast:
Naive: 0.604756s
speedup: 35.790817x
dx difference:  0.0
```

Implementation of cascaded layers

We've provided the following functions in `nndl/conv_layer_utils.py` :

```
- conv_relu_forward
- conv_relu_backward
- conv_relu_pool_forward
- conv_relu_pool_backward
```

These use the fast implementations of the conv net layers. You can test them below:

In [8]:
```
from nndl.conv_layer_utils import conv_relu_pool_forward, conv_relu_pool_backward

x = np.random.randn(2, 3, 16, 16)
w = np.random.randn(3, 3, 3, 3)
b = np.random.randn(3,)
dout = np.random.randn(2, 3, 8, 8)
conv_param = {'stride': 1, 'pad': 1}
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

out, cache = conv_relu_pool_forward(x, w, b, conv_param, pool_param)
dx, dw, db = conv_relu_pool_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: conv_relu_pool_forward(x, w, b,
conv_param, pool_param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_relu_pool_forward(x, w, b,
conv_param, pool_param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_relu_pool_forward(x, w, b,
conv_param, pool_param)[0], b, dout)

print('Testing conv_relu_pool')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
Testing conv_relu_pool
dx error:  3.9035485439532574e-08
dw error:  5.191936231119018e-10
db error:  4.298503169513Ol e-10
```

In [9]:
```
from nndl.conv_layer_utils import conv_relu_forward, conv_relu_backward

x = np.random.randn(2, 3, 8, 8)
w = np.random.randn(3, 3, 3, 3)
b = np.random.randn(3,)
dout = np.random.randn(2, 3, 8, 8)
conv_param = {'stride': 1, 'pad': 1}

out, cache = conv_relu_forward(x, w, b, conv_param)
dx, dw, db = conv_relu_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: conv_relu_forward(x, w, b, conv_
param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_relu_forward(x, w, b, conv_
param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_relu_forward(x, w, b, conv_
param)[0], b, dout)

print('Testing conv_relu:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
Testing conv_relu:
dx error:  4.9743478361805337e-08
dw error:  2.1977886692100602e-10
db error:  2.869425594942413e-10
```

What next?

We saw how helpful batch normalization was for training FC nets. In the next notebook, we'll implement a batch normalization for convolutional neural networks, and then finish off by implementing a CNN to improve our validation accuracy on CIFAR-10.

In [ ]:

```python
import numpy as np
from nndl.layers import *
import pdb

"""
This code was originally written for CS 231n at Stanford University
(cs231n.stanford.edu).  It has been modified in various areas for use in the
ECE 239AS class at UCLA.  This includes the descriptions of what code to
implement as well as some slight potential changes in variable names to be
consistent with class nomenclature.  We thank Justin Johnson & Serena Yeung for
permission to use this code.  To see the original version, please visit
cs231n.stanford.edu.
"""

def conv_forward_naive(x, w, b, conv_param):
  """
  A naive implementation of the forward pass for a convolutional layer.

  The input consists of N data points, each with C channels, height H and width
  W. We convolve each input with F different filters, where each filter spans
  all C channels and has height HH and width HH.

  Input:
  - x: Input data of shape (N, C, H, W)
  - w: Filter weights of shape (F, C, HH, WW)
  - b: Biases, of shape (F,)
  - conv_param: A dictionary with the following keys:
    - 'stride': The number of pixels between adjacent receptive fields in the
      horizontal and vertical directions.
    - 'pad': The number of pixels that will be used to zero-pad the input.

  Returns a tuple of:
  - out: Output data, of shape (N, F, H', W') where H' and W' are given by
    H' = 1 + (H + 2 * pad - HH) / stride
    W' = 1 + (W + 2 * pad - WW) / stride
  - cache: (x, w, b, conv_param)
  """
  out = None
  pad = conv_param['pad']
  stride = conv_param['stride']

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the forward pass of a convolutional neural network.
  #   Store the output as 'out'.
  #   Hint: to pad the array, you can use the function np.pad.
  # ================================================================ #
  Hprime = int((x.shape[2] + 2 * pad - w.shape[2]) / stride) + 1
  Wprime = int((x.shape[3] + 2 * pad - w.shape[3]) / stride) + 1
  out = np.zeros((x.shape[0], w.shape[0], Hprime, Wprime))

  for i, dp in enumerate(x):
    padded_dp = np.pad(dp, pad_width=[(0, 0), (pad, pad), (pad, pad)], mode='cons
tant')
    for j, filter in enumerate(w):
      for xpos in range(Hprime):
        xoffset = xpos * stride
        for ypos in range(Wprime):
          yoffset = ypos * stride
          out[i, j, xpos, ypos] = np.sum(np.multiply(padded_dp[:, xoffset:xoffset
+ w.shape[2], yoffset:yoffset + w.shape[3]], filter)) + b[j]
```

```
    # ================================================================= #
    # END YOUR CODE HERE
    # ================================================================= #

    cache = (x, w, b, conv_param)
    return out, cache


def conv_backward_naive(dout, cache):
    """
    A naive implementation of the backward pass for a convolutional layer.

    Inputs:
    - dout: Upstream derivatives.
    - cache: A tuple of (x, w, b, conv_param) as in conv_forward_naive

    Returns a tuple of:
    - dx: Gradient with respect to x
    - dw: Gradient with respect to w
    - db: Gradient with respect to b
    """
    dx, dw, db = None, None, None

    N, F, out_height, out_width = dout.shape
    x, w, b, conv_param = cache

    stride, pad = [conv_param['stride'], conv_param['pad']]
    xpad = np.pad(x, ((0,0), (0,0), (pad,pad), (pad,pad)), mode='constant')
    num_filts, _, f_height, f_width = w.shape

    # ================================================================= #
    # YOUR CODE HERE:
    #   Implement the backward pass of a convolutional neural network.
    #   Calculate the gradients: dx, dw, and db.
    # ================================================================= #
    dx = np.zeros(x.shape)
    dxp = np.pad(dx, ((0,0), (0,0), (pad, pad), (pad, pad)), mode='constant')
    dw = np.zeros(w.shape)
    db = np.zeros(b.shape)
    for i, padded_dp in enumerate(xpad):
      for j, filter in enumerate(w):
        for xpos in range(dout.shape[2]):
          xoffset = xpos * stride
          for ypos in range(dout.shape[3]):
            yoffset = ypos * stride
            dw[j] += dout[i, j, xpos, ypos] * padded_dp[:, xoffset:xoffset + w.shap
e[2], yoffset: yoffset + w.shape[3]]
            dxp[i, :, xoffset:xoffset + w.shape[2], yoffset: yoffset + w.shape[3]]
+= dout[i, j, xpos, ypos] * w[j]
    db = np.sum(np.sum(np.sum(dout, axis=3), axis=2), axis=0)
    dx = dxp[:,:,pad:-pad,pad:-pad]
    # ================================================================= #
    # END YOUR CODE HERE
    # ================================================================= #

    return dx, dw, db


def max_pool_forward_naive(x, pool_param):
    """
    A naive implementation of the forward pass for a max pooling layer.
```

```
  Inputs:
  - x: Input data, of shape (N, C, H, W)
  - pool_param: dictionary with the following keys:
    - 'pool_height': The height of each pooling region
    - 'pool_width': The width of each pooling region
    - 'stride': The distance between adjacent pooling regions

  Returns a tuple of:
  - out: Output data
  - cache: (x, pool_param)
  """
  out = None

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the max pooling forward pass.
  # ================================================================ #
  Hprime = int((x.shape[2] + - pool_param['pool_height']) / pool_param['stride'])
+ 1
  Wprime = int((x.shape[3] + - pool_param['pool_width']) / pool_param['stride'])
+ 1
  out = np.zeros((x.shape[0], x.shape[1], Hprime, Wprime))

  for i, dp in enumerate(x):
    for l, layer in enumerate(dp):
      for xpos in range(Hprime):
        xoffset = xpos * pool_param['stride']
        for ypos in range(Wprime):
          yoffset = ypos * pool_param['stride']
          out[i, l, xpos, ypos] = np.amax(layer[xoffset:xoffset + pool_param['poo
l_height'], yoffset:yoffset + pool_param['pool_width']])

  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #
  cache = (x, pool_param)
  return out, cache

def max_pool_backward_naive(dout, cache):
  """
  A naive implementation of the backward pass for a max pooling layer.

  Inputs:
  - dout: Upstream derivatives
  - cache: A tuple of (x, pool_param) as in the forward pass.

  Returns:
  - dx: Gradient with respect to x
  """
  dx = None
  x, pool_param = cache
  pool_height, pool_width, stride = pool_param['pool_height'], pool_param['pool_w
idth'], pool_param['stride']

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the max pooling backward pass.
  # ================================================================ #
  dx = np.zeros(x.shape)
  for i, dp in enumerate(x):
    for l, layer in enumerate(dp):
      for xpos in range(dout.shape[2]):
```

```
        xoffset = xpos * pool_param['stride']
        for ypos in range(dout.shape[3]):
          yoffset = ypos * pool_param['stride']
          field = layer[xoffset:xoffset + pool_param['pool_height'], yoffset:yoff
set + pool_param['pool_width']]
          ixmax, iymax = np.unravel_index(np.argmax(field, axis=None), field.shap
e)

          dx[i, l, ixmax + xoffset, iymax + yoffset] = dout[i, l, xpos, ypos]
  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #

  return dx

def spatial_batchnorm_forward(x, gamma, beta, bn_param):
  """
  Computes the forward pass for spatial batch normalization.

  Inputs:
  - x: Input data of shape (N, C, H, W)
  - gamma: Scale parameter, of shape (C,)
  - beta: Shift parameter, of shape (C,)
  - bn_param: Dictionary with the following keys:
    - mode: 'train' or 'test'; required
    - eps: Constant for numeric stability
    - momentum: Constant for running mean / variance. momentum=0 means that
      old information is discarded completely at every time step, while
      momentum=1 means that new information is never incorporated. The
      default of momentum=0.9 should work well in most situations.
    - running_mean: Array of shape (D,) giving running mean of features
    - running_var Array of shape (D,) giving running variance of features

  Returns a tuple of:
  - out: Output data, of shape (N, C, H, W)
  - cache: Values needed for the backward pass
  """
  out, cache = None, None

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the spatial batchnorm forward pass.
  #
  #   You may find it useful to use the batchnorm forward pass you
  #   implemented in HW #4.
  # ================================================================ #
  # Manipulate shape
  out = np.zeros(x.shape)
  xF = np.array([x[:, j, :, :].reshape(-1) for j in range(x.shape[1])])

  bn_xFT, cache = batchnorm_forward(xF.T, gamma, beta, bn_param) # batchnorm_xfla
ttenedtranspose

  # Unmanipulate shape
  for i in range(bn_xFT.shape[1]):
    out[:, i, :, :] = bn_xFT[:, i].reshape(out.shape[0], out.shape[2], out.shape[
3])
  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #

  return out, cache
```

```python
def spatial_batchnorm_backward(dout, cache):
  """
  Computes the backward pass for spatial batch normalization.

  Inputs:
  - dout: Upstream derivatives, of shape (N, C, H, W)
  - cache: Values from the forward pass

  Returns a tuple of:
  - dx: Gradient with respect to inputs, of shape (N, C, H, W)
  - dgamma: Gradient with respect to scale parameter, of shape (C,)
  - dbeta: Gradient with respect to shift parameter, of shape (C,)
  """
  dx, dgamma, dbeta = None, None, None

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the spatial batchnorm backward pass.
  #
  #   You may find it useful to use the batchnorm forward pass you
  #   implemented in HW #4.
  # ================================================================ #
  # Manipulate shape
  dx = np.zeros(dout.shape)
  dgamma = np.zeros(dout.shape[1])
  dbeta = np.zeros(dout.shape[1])
  doutF = np.array([dout[:,j,:,:].reshape(-1) for j in range(dout.shape[1])])

  dxFT, dgamma, dbeta = batchnorm_backward(doutF.T, cache)

  # Unmanipulate shape
  for i in range(dxFT.shape[1]):
    dx[:, i, :, :] = dxFT[:, i].reshape(dx.shape[0], dx.shape[2], dx.shape[3])
  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #

  return dx, dgamma, dbeta
```

Spatial batch normalization

In fully connected networks, we performed batch normalization on the activations. To do something equivalent on CNNs, we modify batch normalization slightly.

Normally batch-normalization accepts inputs of shape `(N, D)` and produces outputs of shape `(N, D)` , where we normalize across the minibatch dimension `N` . For data coming from convolutional layers, batch normalization accepts inputs of shape `(N, C, H, W)` and produces outputs of shape `(N, C, H, W)` where the `N` dimension gives the minibatch size and the `(H, W)` dimensions give the spatial size of the feature map.

How do we calculate the spatial averages? First, notice that for the `C` feature maps we have (i.e., the layer has `C` filters) that each of these ought to have its own batch norm statistics, since each feature map may be picking out very different features in the images. However, within a feature map, we may assume that across all inputs and across all locations in the feature map, there ought to be relatively similar first and second order statistics. Hence, one way to think of spatial batch-normalization is to reshape the `(N, C, H, W)` array as an `(N*H*W, C)` array and perform batch normalization on this array.

Since spatial batch norm and batch normalization are similar, it'd be good to at this point also copy and paste our prior implemented layers from HW #4. Please copy and paste your prior implemented code from HW #4 to start this assignment. If you did not correctly implement the layers in HW #4, you may collaborate with a classmate to use their implementations from HW #4. You may also visit TA or Prof OH to correct your implementation.

You'll want to copy and paste from HW #4:

```
- layers.py for your FC network layers, as well as batchnorm and dropout.
- layer_utils.py for your combined FC network layers.
- optim.py for your optimizers.
```

Be sure to place these in the `nndl/` directory so they're imported correctly. Note, as announced in class, we will not be releasing our solutions.

If you use your prior implementations of the batchnorm, then your spatial batchnorm implementation may be very short. Our implementations of the forward and backward pass are each 6 lines of code.

CS231n has built a solid API for building these modular frameworks and training them, and we will use their very well implemented framework as opposed to "reinventing the wheel." This includes using their Solver, various utility functions, their layer structure, and their implementation of fast CNN layers. This also includes nndl.fc_net, nndl.layers, and nndl.layer_utils. As in prior assignments, we thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class (cs231n.stanford.edu).

In [1]:
```
## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.conv_layers import *
from cs231n.data_utils import get_CIFAR10_data
from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradien
t_array
from cs231n.solver import Solver

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
  """ returns relative error """
  return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

Spatial batch normalization forward pass

Implement the forward pass, `spatial_batchnorm_forward` in `nndl/conv_layers.py` . Test your implementation by running the cell below.

In [2]:
```
# Check the training-time forward pass by checking means and variances
# of features both before and after spatial batch normalization

N, C, H, W = 2, 3, 4, 5
x = 4 * np.random.randn(N, C, H, W) + 10

print('Before spatial batch normalization:')
print('  Shape: ', x.shape)
print('  Means: ', x.mean(axis=(0, 2, 3)))
print('  Stds: ', x.std(axis=(0, 2, 3)))

# Means should be close to zero and stds close to one
gamma, beta = np.ones(C), np.zeros(C)
bn_param = {'mode': 'train'}
out, _ = spatial_batchnorm_forward(x, gamma, beta, bn_param)
print('After spatial batch normalization:')
print('  Shape: ', out.shape)
print('  Means: ', out.mean(axis=(0, 2, 3)))
print('  Stds: ', out.std(axis=(0, 2, 3)))

# Means should be close to beta and stds close to gamma
gamma, beta = np.asarray([3, 4, 5]), np.asarray([6, 7, 8])
out, _ = spatial_batchnorm_forward(x, gamma, beta, bn_param)
print('After spatial batch normalization (nontrivial gamma, beta):')
print('  Shape: ', out.shape)
print('  Means: ', out.mean(axis=(0, 2, 3)))
print('  Stds: ', out.std(axis=(0, 2, 3)))
```

```
Before spatial batch normalization:
  Shape:  (2, 3, 4, 5)
  Means:  [11.34592147  9.64806625 10.49672924]
  Stds:  [4.69225273 3.4007622  3.78221833]
After spatial batch normalization:
  Shape:  (2, 3, 4, 5)
  Means:  [ 9.99200722e-17  6.66133815e-17 -3.33066907e-17]
  Stds:  [0.99999977 0.99999957 0.99999965]
After spatial batch normalization (nontrivial gamma, beta):
  Shape:  (2, 3, 4, 5)
  Means:  [6. 7. 8.]
  Stds:  [2.99999932 3.99999827 4.99999825]
```

Spatial batch normalization backward pass

Implement the backward pass, `spatial_batchnorm_backward` in `nndl/conv_layers.py` . Test your implementation by running the cell below.

In [3]:
```
N, C, H, W = 2, 3, 4, 5
x = 5 * np.random.randn(N, C, H, W) + 12
gamma = np.random.randn(C)
beta = np.random.randn(C)
dout = np.random.randn(N, C, H, W)

bn_param = {'mode': 'train'}
fx = lambda x: spatial_batchnorm_forward(x, gamma, beta, bn_param)[0]
fg = lambda a: spatial_batchnorm_forward(x, gamma, beta, bn_param)[0]
fb = lambda b: spatial_batchnorm_forward(x, gamma, beta, bn_param)[0]

dx_num = eval_numerical_gradient_array(fx, x, dout)
da_num = eval_numerical_gradient_array(fg, gamma, dout)
db_num = eval_numerical_gradient_array(fb, beta, dout)

_, cache = spatial_batchnorm_forward(x, gamma, beta, bn_param)
dx, dgamma, dbeta = spatial_batchnorm_backward(dout, cache)
print('dx error: ', rel_error(dx_num, dx))
print('dgamma error: ', rel_error(da_num, dgamma))
print('dbeta error: ', rel_error(db_num, dbeta))
```

```
dx error:  9.971961087703966e-08
dgamma error:  4.21185456225215e-12
dbeta error:  8.532218482360866e-12
```

Convolutional neural networks

In this notebook, we'll put together our convolutional layers to implement a 3-layer CNN. Then, we'll ask you to implement a CNN that can achieve > 65% validation error on CIFAR-10.

CS231n has built a solid API for building these modular frameworks and training them, and we will use their very well implemented framework as opposed to "reinventing the wheel." This includes using their Solver, various utility functions, their layer structure, and their implementation of fast CNN layers. This also includes nndl.fc_net, nndl.layers, and nndl.layer_utils. As in prior assignments, we thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class (cs231n.stanford.edu).

If you have not completed the Spatial BatchNorm Notebook, please see the following description from that notebook:

Please copy and paste your prior implemented code from HW #4 to start this assignment. If you did not correctly implement the layers in HW #4, you may collaborate with a classmate to use their layer implementations from HW #4. You may also visit TA or Prof OH to correct your implementation.

You'll want to copy and paste from HW #4:

```
- layers.py for your FC network layers, as well as batchnorm and dropout.
- layer_utils.py for your combined FC network layers.
- optim.py for your optimizers.
```

Be sure to place these in the `nndl/` directory so they're imported correctly. Note, as announced in class, we will not be releasing our solutions.

In [1]:

```
# As usual, a bit of setup

import numpy as np
import matplotlib.pyplot as plt
from nndl.cnn import *
from cs231n.data_utils import get_CIFAR10_data
from cs231n.gradient_check import eval_numerical_gradient_array, eval_numerical_g
radient
from nndl.layers import *
from nndl.conv_layers import *
from cs231n.fast_layers import *
from cs231n.solver import Solver

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
  """ returns relative error """
  return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

In [2]:
```
# Load the (preprocessed) CIFAR10 data.

data = get_CIFAR10_data()
for k in data.keys():
  print('{}: {} '.format(k, data[k].shape))
```

```
X_train: (49000, 3, 32, 32)
y_train: (49000,)
X_val: (1000, 3, 32, 32)
y_val: (1000,)
X_test: (1000, 3, 32, 32)
y_test: (1000,)
```

Three layer CNN

In this notebook, you will implement a three layer CNN. The `ThreeLayerConvNet` class is in `nndl/cnn.py` . You'll need to modify that code for this section, including the initialization, as well as the calculation of the loss and gradients. You should be able to use the building blocks you have either earlier coded or that we have provided. Be sure to use the fast layers.

The architecture of this CNN will be:

conv - relu - 2x2 max pool - affine - relu - affine - softmax

We won't use batchnorm yet. You've also done enough of these to know how to debug; use the cells below.

Note: As we are implementing several layers CNN networks. The gradient error can be expected for the `eval_numerical_gradient()` function. If your `W1 max relative error` and `W2 max relative error` are around or below 0.01, they should be acceptable. Other errors should be less than 1e-5.

In [ ]:
```
num_inputs = 2
input_dim = (3, 16, 16)
reg = 0.0
num_classes = 10
X = np.random.randn(num_inputs, *input_dim)
y = np.random.randint(num_classes, size=num_inputs)

model = ThreeLayerConvNet(num_filters=3, filter_size=3,
                          input_dim=input_dim, hidden_dim=7,
                          dtype=np.float64)
loss, grads = model.loss(X, y)
for param_name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    param_grad_num = eval_numerical_gradient(f, model.params[param_name], verbose
=False, h=1e-6)
    e = rel_error(param_grad_num, grads[param_name])
    print('{} max relative error: {}'.format(param_name, rel_error(param_grad_num
, grads[param_name])))
```

Overfit small dataset

To check your CNN implementation, let's overfit a small dataset.

In [ ]:
```
num_train = 100
small_data = {
  'X_train': data['X_train'][:num_train],
  'y_train': data['y_train'][:num_train],
  'X_val': data['X_val'],
  'y_val': data['y_val'],
}

model = ThreeLayerConvNet(weight_scale=1e-2)

solver = Solver(model, small_data,
                num_epochs=10, batch_size=50,
                update_rule='adam',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True, print_every=1)
solver.train()
```

In [ ]:
```
plt.subplot(2, 1, 1)
plt.plot(solver.loss_history, 'o')
plt.xlabel('iteration')
plt.ylabel('loss')

plt.subplot(2, 1, 2)
plt.plot(solver.train_acc_history, '-o')
plt.plot(solver.val_acc_history, '-o')
plt.legend(['train', 'val'], loc='upper left')
plt.xlabel('epoch')
plt.ylabel('accuracy')
plt.show()
```

Train the network

Now we train the 3 layer CNN on CIFAR-10 and assess its accuracy.

In [ ]:
```
model = ThreeLayerConvNet(weight_scale=0.001, hidden_dim=500, reg=0.001)

solver = Solver(model, data,
                num_epochs=1, batch_size=50,
                update_rule='adam',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True, print_every=20)
solver.train()
```

Get > 65% validation accuracy on CIFAR-10.

In the last part of the assignment, we'll now ask you to train a CNN to get better than 65% validation accuracy on CIFAR-10.

Things you should try:

- Filter size: Above we used 7x7; but VGGNet and onwards showed stacks of 3x3 filters are good.
- Number of filters: Above we used 32 filters. Do more or fewer do better?
- Batch normalization: Try adding spatial batch normalization after convolution layers and vanilla batch normalization aafter affine layers. Do your networks train faster?
- Network architecture: Can a deeper CNN do better? Consider these architectures:
  - [conv-relu-pool]xN - conv - relu - [affine]xM - [softmax or SVM]
  - [conv-relu-pool]XN - [affine]XM - [softmax or SVM]
  - [conv-relu-conv-relu-pool]xN - [affine]xM - [softmax or SVM]

Tips for training

For each network architecture that you try, you should tune the learning rate and regularization strength. When doing this there are a couple important things to keep in mind:

- If the parameters are working well, you should see improvement within a few hundred iterations
- Remember the coarse-to-fine approach for hyperparameter tuning: start by testing a large range of hyperparameters for just a few training iterations to find the combinations of parameters that are working at all.
- Once you have found some sets of parameters that seem to work, search more finely around these parameters. You may need to train for more epochs.

In [10]:
```
# ================================================================ #
# YOUR CODE HERE:
#   Implement a CNN to achieve greater than 65% validation accuracy
#   on CIFAR-10.
# ================================================================ #
model = ThreeLayerConvNet(  filter_size=3,
                            num_filters=128,
                            weight_scale=0.001,
                            hidden_dim=1024,
                            reg=0.002)

solver = Solver(model, data,
                num_epochs=20, batch_size=128,
                update_rule='adam',
                optim_config={
                  'learning_rate': 5e-4,
                },
                verbose=True, print_every=100)
solver.train()

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```

```
(Iteration 1 / 7640) loss: 2.336331
(Epoch 0 / 20) train acc: 0.098000; val_acc: 0.088000
(Iteration 101 / 7640) loss: 1.670949
(Iteration 201 / 7640) loss: 1.511465
(Iteration 301 / 7640) loss: 1.482682
(Epoch 1 / 20) train acc: 0.538000; val_acc: 0.529000
(Iteration 401 / 7640) loss: 1.593825
(Iteration 501 / 7640) loss: 1.343916
(Iteration 601 / 7640) loss: 1.368077
(Iteration 701 / 7640) loss: 1.415682
(Epoch 2 / 20) train acc: 0.645000; val_acc: 0.616000
(Iteration 801 / 7640) loss: 1.417564
(Iteration 901 / 7640) loss: 1.178932
(Iteration 1001 / 7640) loss: 1.110138
(Iteration 1101 / 7640) loss: 1.101641
(Epoch 3 / 20) train acc: 0.716000; val_acc: 0.622000
(Iteration 1201 / 7640) loss: 1.276990
(Iteration 1301 / 7640) loss: 1.147939
(Iteration 1401 / 7640) loss: 1.121312
(Iteration 1501 / 7640) loss: 1.105080
(Epoch 4 / 20) train acc: 0.680000; val_acc: 0.641000
(Iteration 1601 / 7640) loss: 1.257500
(Iteration 1701 / 7640) loss: 1.118692
(Iteration 1801 / 7640) loss: 1.162882
(Iteration 1901 / 7640) loss: 1.152434
(Epoch 5 / 20) train acc: 0.728000; val_acc: 0.647000
(Iteration 2001 / 7640) loss: 0.933882
(Iteration 2101 / 7640) loss: 1.007060
(Iteration 2201 / 7640) loss: 1.018173
(Epoch 6 / 20) train acc: 0.740000; val_acc: 0.626000
(Iteration 2301 / 7640) loss: 1.053089
(Iteration 2401 / 7640) loss: 0.852686
(Iteration 2501 / 7640) loss: 0.836276
(Iteration 2601 / 7640) loss: 0.973918
(Epoch 7 / 20) train acc: 0.772000; val_acc: 0.655000
(Iteration 2701 / 7640) loss: 0.833156
(Iteration 2801 / 7640) loss: 0.960028
(Iteration 2901 / 7640) loss: 0.869738
(Iteration 3001 / 7640) loss: 0.894365
(Epoch 8 / 20) train acc: 0.755000; val_acc: 0.638000
(Iteration 3101 / 7640) loss: 1.066237
(Iteration 3201 / 7640) loss: 0.925506
(Iteration 3301 / 7640) loss: 0.901143
(Iteration 3401 / 7640) loss: 0.895855
(Epoch 9 / 20) train acc: 0.780000; val_acc: 0.620000
(Iteration 3501 / 7640) loss: 1.014221
(Iteration 3601 / 7640) loss: 0.876468
(Iteration 3701 / 7640) loss: 0.873967
(Iteration 3801 / 7640) loss: 0.898300
(Epoch 10 / 20) train acc: 0.780000; val_acc: 0.630000
(Iteration 3901 / 7640) loss: 0.906659
(Iteration 4001 / 7640) loss: 0.734977
(Iteration 4101 / 7640) loss: 0.798377
(Iteration 4201 / 7640) loss: 0.797348
(Epoch 11 / 20) train acc: 0.824000; val_acc: 0.670000
(Iteration 4301 / 7640) loss: 0.720607
(Iteration 4401 / 7640) loss: 0.774881
(Iteration 4501 / 7640) loss: 0.938662
(Epoch 12 / 20) train acc: 0.826000; val_acc: 0.661000
(Iteration 4601 / 7640) loss: 0.838905
(Iteration 4701 / 7640) loss: 0.690243
```

```
(Iteration 4801 / 7640) loss: 0.606958
(Iteration 4901 / 7640) loss: 0.607002
(Epoch 13 / 20) train acc: 0.814000; val_acc: 0.643000
(Iteration 5001 / 7640) loss: 0.661225
(Iteration 5101 / 7640) loss: 0.810502
(Iteration 5201 / 7640) loss: 0.741615
(Iteration 5301 / 7640) loss: 0.676790
(Epoch 14 / 20) train acc: 0.822000; val_acc: 0.661000
(Iteration 5401 / 7640) loss: 0.763202
(Iteration 5501 / 7640) loss: 0.799008
(Iteration 5601 / 7640) loss: 0.720595
(Iteration 5701 / 7640) loss: 0.797355
(Epoch 15 / 20) train acc: 0.833000; val_acc: 0.655000
(Iteration 5801 / 7640) loss: 0.644746
(Iteration 5901 / 7640) loss: 0.649774
(Iteration 6001 / 7640) loss: 0.674265
(Iteration 6101 / 7640) loss: 0.644854
(Epoch 16 / 20) train acc: 0.867000; val_acc: 0.679000
(Iteration 6201 / 7640) loss: 0.759045
(Iteration 6301 / 7640) loss: 0.766314
(Iteration 6401 / 7640) loss: 0.625347
(Epoch 17 / 20) train acc: 0.866000; val_acc: 0.649000
(Iteration 6501 / 7640) loss: 0.650522
(Iteration 6601 / 7640) loss: 0.564891
(Iteration 6701 / 7640) loss: 0.890038
(Iteration 6801 / 7640) loss: 0.613686
(Epoch 18 / 20) train acc: 0.860000; val_acc: 0.650000
(Iteration 6901 / 7640) loss: 0.586294
(Iteration 7001 / 7640) loss: 0.735426
(Iteration 7101 / 7640) loss: 0.626093
(Iteration 7201 / 7640) loss: 0.606037
(Epoch 19 / 20) train acc: 0.856000; val_acc: 0.650000
(Iteration 7301 / 7640) loss: 0.767815
(Iteration 7401 / 7640) loss: 0.694980
(Iteration 7501 / 7640) loss: 0.537208
(Iteration 7601 / 7640) loss: 0.575530
(Epoch 20 / 20) train acc: 0.890000; val_acc: 0.667000
```

In [ ]:
```
import numpy as np

from nndl.layers import *
from nndl.conv_layers import *
from cs231n.fast_layers import *
from nndl.layer_utils import *
from nndl.conv_layer_utils import *

import pdb

"""
This code was originally written for CS 231n at Stanford University
(cs231n.stanford.edu).  It has been modified in various areas for use in the
ECE 239AS class at UCLA.  This includes the descriptions of what code to
implement as well as some slight potential changes in variable names to be
consistent with class nomenclature.  We thank Justin Johnson & Serena Yeung for
permission to use this code.  To see the original version, please visit
cs231n.stanford.edu.
"""

class ThreeLayerConvNet(object):
  """
  A three-layer convolutional network with the following architecture:

  conv - relu - 2x2 max pool - affine - relu - affine - softmax

  The network operates on minibatches of data that have shape (N, C, H, W)
  consisting of N images, each with height H and width W and with C input
  channels.
  """

  def __init__(self, input_dim=(3, 32, 32), num_filters=32, filter_size=7,
               hidden_dim=100, num_classes=10, weight_scale=1e-3, reg=0.0,
               dtype=np.float32, use_batchnorm=False):
    """
    Initialize a new network.

    Inputs:
    - input_dim: Tuple (C, H, W) giving size of input data
    - num_filters: Number of filters to use in the convolutional layer
    - filter_size: Size of filters to use in the convolutional layer
    - hidden_dim: Number of units to use in the fully-connected hidden layer
    - num_classes: Number of scores to produce from the final affine layer.
    - weight_scale: Scalar giving standard deviation for random initialization
      of weights.
    - reg: Scalar giving L2 regularization strength
    - dtype: numpy datatype to use for computation.
    """
    self.use_batchnorm = use_batchnorm
    self.params = {}
    self.reg = reg
    self.dtype = dtype


    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize the weights and biases of a three layer CNN. To initialize:
    #     - the biases should be initialized to zeros.
    #     - the weights should be initialized to a matrix with entries
    #         drawn from a Gaussian distribution with zero mean and
    #         standard deviation given by weight_scale.
```

```
    # ========================================================================= #
    self.params['W1'] = np.random.normal(0, weight_scale, [num_filters, input_dim
[0], filter_size, filter_size])
    self.params['b1'] = np.zeros(num_filters)
    W1_lenx = int((input_dim[1] - 2) / 2) + 1 # b/c pad is set such that conv lay
er doesn't shrink it, but there is a pool
    W1_leny = int((input_dim[2] - 2) / 2) + 1
    self.params['W2'] = np.random.normal(0, weight_scale, [W1_lenx * W1_leny * nu
m_filters, hidden_dim])
    self.params['b2'] = np.zeros(hidden_dim)
    self.params['W3'] = np.random.normal(0, weight_scale, [hidden_dim, num_classe
s])
    self.params['b3'] = np.zeros(num_classes)

    if self.use_batchnorm:
        self.bn_params = [{'mode': 'train', 'eps': 1e-5, 'momentum': 0.9} for i i
n np.arange(self.num_layers - 1)]
        self.params['gamma1'] = np.ones(input_dim[0])
        self.params['beta1'] = np.zeros(input_dim[0])
        self.params['gamma2'] = np.ones()
        self.params['beta2'] = np.zeros()
    # ========================================================================= #
    # END YOUR CODE HERE
    # ========================================================================= #

    for k, v in self.params.items():
      self.params[k] = v.astype(dtype)


  def loss(self, X, y=None):
    """
    Evaluate loss and gradient for the three-layer convolutional network.

    Input / output: Same API as TwoLayerNet in fc_net.py.
    """
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    W3, b3 = self.params['W3'], self.params['b3']

    # pass conv_param to the forward pass for the convolutional layer
    filter_size = W1.shape[2]
    conv_param = {'stride': 1, 'pad': (filter_size - 1) / 2}

    # pass pool_param to the forward pass for the max-pooling layer
    pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

    scores = None

    # ========================================================================= #
    # YOUR CODE HERE:
    #   Implement the forward pass of the three layer CNN.  Store the output
    #   scores as the variable "scores".
    # ========================================================================= #



    if not self.use_batchnorm:
        # conv - relu - 2x2 max pool - affine - relu - affine - softmax
        p1, p1_cache = conv_relu_pool_forward(X, W1, b1, conv_param, pool_param)
# a1 -> h1 -> p1
        h2, h2_cache = affine_relu_forward(p1, W2, b2) # p1 -> a2 -> h2
        scores, a3_cache = affine_forward(h2, W3, b3)
```

```
    else:
        # conv - sbn - relu - 2x2 max pool - affine - bn - relu - affine - softma
x
        a, conv_cache = conv_forward_fast(x, w, b, conv_param)
        out, cache = spatial_batchnorm_forward(x, gamma, beta, bn_param)
        s, relu_cache = relu_forward(a)
        out, pool_cache = max_pool_forward_fast(s, pool_param)
        out, cache = affine_forward(x, w, b)
        out, cache = batchnorm_forward(x, gamma, beta, bn_param)
        s, relu_cache = relu_forward(a)
        out, cache = affine_forward(x, w, b)
    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    if y is None:
      return scores

    loss, grads = 0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backward pass of the three layer CNN.  Store the grads
    #   in the grads dictionary, exactly as before (i.e., the gradient of
    #   self.params[k] will be grads[k]).  Store the loss as "loss", and
    #   don't forget to add regularization on ALL weight matrices.
    # ================================================================ #
    loss, dx = softmax_loss(scores, y)
    loss += 0.5 * self.reg * np.sum([np.sum(self.params['W{}'.format(layer + 1)]
** 2) for layer in range(3)])

    if not self.use_batchnorm:
        dl_dh2, grads['W3'], grads['b3'] = affine_backward(dx, a3_cache)
        grads['W3'] += self.reg * self.params['W3']
        dl_dp1, grads['W2'], grads['b2'] = affine_relu_backward(dl_dh2, h2_cache)
        grads['W2'] += self.reg * self.params['W2']
        dl_dx, grads['W1'], grads['b1'] = conv_relu_pool_backward(dl_dp1, p1_cach
e)
        grads['W1'] += self.reg * self.params['W1']
    else:
        pass
    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grads


pass
```