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

"""
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']

  # ================================================================ #
```

```python
# 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.
# ================================================================ #

w_in, h_in = x.shape[3], x.shape[2]
w_kernel, h_kernel = w.shape[3], w.shape[2]
num_img, num_channel, num_kernel = x.shape[0], x.shape[1], w.shape[0]

w_out = int(np.floor((w_in+2*pad-w_kernel)/stride) + 1)
h_out = int(np.floor((h_in+2*pad-h_kernel)/stride) + 1)
out = np.zeros([num_img, num_kernel, h_out, w_out])

"""
for every img input
iterate thru ouput pixels [i,j] ([vertical axis, horizontal axis])
for each pixel in output, see the corresponding input window
multiply input win with the filter(weight), take a sum over # channel
conv(1*receptive_field*#ch, #filter*filter_size(==receptive_field)*#ch) -> #filter*1 (1 #
that's how we do for ONE output pixel [i,j] for ONE img example n
"""

zp_row = np.zeros([num_img, num_channel, pad, w_in])
zp_col = np.zeros([num_img, num_channel, h_in+2*pad, pad])
# use np.concat to pad feature maps (NOT np.stack)
xpad = np.concatenate([zp_col, np.concatenate([zp_row, x, zp_row], axis=2), zp_col], \
                                               axis=3)

for n in range(num_img):
  for i in range(h_out):
    for j in range(w_out):
      xpad_n = xpad[n]  # xpad_n of shape [#ch, h_in, w_in]

      # determine the range of the input that is multiplied by the filters
      x_seg = xpad_n[:, 0+i*stride:0+i*stride+h_kernel, 0+j*stride:0+j*stride+w_kernel]

      # calculate the conv output; b is bias term
      # segment of the input * filter + b; compress over channel dim, w dim, and h dim
      # expected shape: [#filter, 1, 1]
      out[n,:,i,j] = np.sum(np.multiply(x_seg, w), axis=(1,2,3)) + b

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

```python
    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, dw, db = np.zeros(x.shape), np.zeros(w.shape), np.zeros(b.shape)

    w_in, h_in = x.shape[3], x.shape[2]
    w_kernel, h_kernel = w.shape[3], w.shape[2]
    num_img, num_channel, num_kernel = x.shape[0], x.shape[1], w.shape[0]

    w_out = int(np.floor((w_in+2*pad-w_kernel)/stride) + 1)
    h_out = int(np.floor((h_in+2*pad-h_kernel)/stride) + 1)

    # xpad shape: [N, C, h_in+zp, w_in+zp]
    # w kernel shape: [F, C, h_kernel, w_kernel]
    # out shape: [N, F, h_out, w_out]
    # b shape: [F]
```

```
# Calculate dw with numerical derivation
for n in tqdm(range(num_img)):
  for f in range(num_filts):
    for c in range(num_channel):
      for i in range(h_out):
        for j in range(w_out):
          dout_n_f = dout[n][f]   # dout_n_f: matrix of shape [h_out, w_out]; work with 1
          x_seg = xpad[n, c, \
                       0+i*stride:0+i*stride+h_kernel, 0+j*stride:0+j*stride+w_kernel]  #

          # find the corresponding receptive field x_seg in x_pad during conv (same shape
          # since during conv, each receptive field contributes to kernel, need to accumu
          # output of conv is scalar y[i,j], thus upperstream grad is scalar
          dw[f, c, :, :] += x_seg*dout_n_f[i][j]


# Calculate dx with numerical derivation
# get a container for dx_pad since we do math on x_pad
# the indexing in conv forward pass still applies; remember to trim dx_pad->dx
dx_pad = np.zeros(xpad.shape)

for n in tqdm(range(num_img)):
  for f in range(num_filts):
    for c in range(num_channel):
      for i in range(h_out):
        for j in range(w_out):
          dout_n_f = dout[n][f]   # dout_n_f: matrix of shape [h_out, w_out]; work with 1
          kernel_f_c = w[f][c]  # kernel_f_c: matrix of shape [h_kernel, w_jernel]; work w

          # find the corresponding receptive field in x_pad during conv (same shape as ke
          # since we have overlapping regions in x_pad during conv, need to accumulate gr
          # output of conv is scalar y[i,j], thus upperstream grad is scalar
          dx_pad[n, c, 0+i*stride:0+i*stride+h_kernel, 0+j*stride:0+j*stride+w_kernel] \
          += kernel_f_c*dout_n_f[i][j]

dx = dx_pad[:,:,pad:-pad,pad:-pad]

"""
Faster Method if dealing with conv 3*3 same

# Calculate dw
for n in tqdm(range(num_img)):
  for f in range(num_filts):
    for c in range(num_channel):
      for i in range(h_kernel):
        for j in range(w_kernel):
```

```
                # work with 1 layer of feature map (matrix) for 1 example
                xpad_n_c = xpad[n][c]  # xpad_n of shape [h_in, w_in],

                # backprop w.r.t to W; dL/dW has a 2Dconv relationship with dout
                x_seg = xpad_n_c[0+i*stride:0+i*stride+h_out, 0+j*stride:0+j*stride+w_out]

                # each of the nth example contributes to dw, thus accumulate dw
                # scalar value to assign; assign to 1 entry in the dw matrix
                dw[f, c, i, j] += np.sum(np.multiply(x_seg, dout[n][f]), axis=(0,1))

    # Calculate dx with conv
    for n in tqdm(range(num_img)):
      for f in range(num_filts):
        for c in range(num_channel):
          for k in range(int(h_in)):
            for l in range(int(w_in)):

                dout_n_f = dout[n][f]   # dout_n_f: matrix of shape [h_out, w_out]; work with 1
                zp_out_col = np.zeros([dout_n_f.shape[0], pad])
                zp_out_row = np.zeros([pad, 2*pad+dout_n_f.shape[1]])
                dout_n_f_zp = np.vstack([zp_out_row, np.hstack([zp_out_col, dout_n_f, zp_out_co
                kernel_f_c = w[f][c]
                kernel_flip = np.flip(kernel_f_c)  # get flipped kernel
                h_flipkernel, w_flipkernel = kernel_flip.shape[0], kernel_flip.shape[1]

                # backprop w.r.t to X; dL/dX has a 2Dconv relationship with dout
                dout_n_f_zp_seg = dout_n_f_zp[0+k*stride:0+k*stride+h_flipkernel, 0+l*stride:0+

                # each of the nth example contributes to dx, thus accumulate dx
                # scalar value to assign; assign to 1 entry in the dx matrix

                dx[n, c, k, l] += np.sum(np.multiply(dout_n_f_zp_seg, kernel_flip), axis=(0,1))
    """

    # all entries have +b contribution, which correspond to local grad 1
    db = np.sum(dout, axis=(0,2,3))

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

    return dx, dw, db


def max_pool_forward_naive(x, pool_param):
```

```python
"""
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.
# ================================================================ #

num_img, num_channel, w_in, h_in = x.shape[0], x.shape[1], x.shape[3], x.shape[2]
h_kernel, w_kernel, stride = pool_param["pool_height"], pool_param["pool_width"], \
                             pool_param["stride"]

h_mp = int(np.floor((w_in-w_kernel)/stride) + 1)
w_mp = int(np.floor((h_in-h_kernel)/stride) + 1)
out = np.zeros([num_img, num_channel, h_mp, w_mp])

for n in range(num_img):
  for c in range(num_channel):
    for i in range(h_mp):
      for j in range(w_mp):
        feature_map = x[n][c]  # feature_map of shape [h_in, w_in]

        # determine the range of the input that we need to find the max
        # do it for every example and for every channel
        map_seg = feature_map[0+i*stride:0+i*stride+h_kernel, \
                              0+j*stride:0+j*stride+w_kernel]
        out[n,c,i,j] = np.amax(map_seg)

# ================================================================ #
# 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_width'], \
                                    pool_param['stride']

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the max pooling backward pass.
  # ================================================================ #

  num_img, num_channel, w_in, h_in = x.shape[0], x.shape[1], x.shape[3], x.shape[2]
  h_kernel, w_kernel, stride = pool_param["pool_height"], pool_param["pool_width"], \
                               pool_param["stride"]
  h_mp = int(np.floor((w_in-w_kernel)/stride) + 1)
  w_mp = int(np.floor((h_in-h_kernel)/stride) + 1)
  dx = np.zeros(x.shape)

  for n in tqdm(range(num_img)):
    for c in range(num_channel):
      for i in range(h_mp):
        for j in range(w_mp):
          feature_map = x[n][c]  # feature_map of shape [h_in, w_in]
          dout_n_c = dout[n][c]  # dout_n_c of shape [h_mp, w_mp]

          # determine the range of the input that we need to find the max for backprop
          # do it for every example and for every channel
          dx_seg = feature_map[0+i*stride:0+i*stride+h_kernel, \
                               0+j*stride:0+j*stride+w_kernel]  # find the mp patch
          max_loc_i, max_loc_j = np.unravel_index(np.argmax(dx_seg), dx_seg.shape)   # find
          dx[n, c, max_loc_i+i*stride, max_loc_j+j*stride] = 1*dout_n_c[i][j]  # convert pa
                                                                               # pass grad

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

```python
    # =============================================================== #

    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.
    # =============================================================== #
    """
    In CNN:
    Spatial batch-normalization is to reshape the (N, C, H, W) array into (N*H*W, C)
    and perform batch normalization on this array.

    In: (N, C, H, W) array
    Out: (N, C, H, W) array
    """

    # N, C, H, W = x.shape
    x_permute = np.transpose(x, [0,2,3,1])  # permute axes into shape (N, H, W, C)
```

```python
    x_spatial_bn = np.reshape(x_permute, \
                              [np.prod(x_permute.shape[0:3]), x_permute.shape[-1]])  # reshape
    out_bn, cache = batchnorm_forward(x_spatial_bn, gamma, beta, bn_param)  # out_bn shape: (N
    out_permute = np.reshape(out_bn, x_permute.shape)   # out_permute shape: (N, H, W, C)
    out = np.transpose(out_permute, [0,3,1,2])  # permute axes back into shape (N, C, H, W)

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

    return out, cache


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.
    # ================================================================ #

    dout_permute = np.transpose(dout, [0,2,3,1])  # permute axes into shape (N, H, W, C)
    dout_bn = np.reshape(dout_permute, \
                         [np.prod(dout_permute.shape[0:3]), dout_permute.shape[-1]])   # dout_
    dx_bn, dgamma, dbeta = batchnorm_backward(dout_bn, cache)   # dx_bn shape: (N*H*W, C)
    dx_bn_permute = np.reshape(dx_bn, dout_permute.shape)   # out_permute shape: (N, H, W, C)
    dx = np.transpose(dx_bn_permute, [0,3,1,2])  # permute axes back into shape (N, C, H, W)

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

```
return dx, dgamma, dbeta
```

CNN-Layers

March 3, 2024

0.1 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).

[ ]:
```
## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.conv_layers import *
from utils.data_utils import get_CIFAR10_data
from utils.gradient_check import eval_numerical_gradient,␣
 ↪eval_numerical_gradient_array
from utils.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))))
```

0.2 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`.

0.2.1 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.

[ ]:
```
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
```

0.2.2 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.

```
[ ]: 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))
```

```
100%|      | 4/4 [00:00<00:00, 1333.85it/s]
100%|      | 4/4 [00:00<00:00, 1333.43it/s]
```

```
Testing conv_backward_naive function
dx error:  5.2211455605974305e-08
dw error:  2.1169348498483835e-09
db error:  2.276947320173437e-11
```

0.2.3 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.

```
[ ]: 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
```

0.2.4 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.

[ ]:
```
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_param)[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))
```

```
100%|      | 3/3 [00:00<00:00, 2998.79it/s]

Testing max_pool_backward_naive function:
dx error:  3.2756358192849457e-12
```

0.3 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.

[ ]:
```
from utils.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: 0.276013s
Fast: 0.010985s
Speedup: 25.127146x
Difference:  2.4824849014007703e-10
```

```
100%|      | 100/100 [00:08<00:00, 11.98it/s]
100%|      | 100/100 [00:09<00:00, 10.15it/s]
```

```
Testing conv_backward_fast:
Naive: 18.209036s
Fast: 0.008000s
Speedup: 2276.091015x
dx difference:  5.5252831220167364e-11
dw difference:  2.616766957786779e-12
db difference:  0.0
```

[ ]:

```
from utils.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: 0.360000s
fast: 0.008000s
speedup: 45.000626x
difference:  0.0

100%|      | 100/100 [00:00<00:00, 206.61it/s]


Testing pool_backward_fast:
Naive: 0.486016s
speedup: 48.681736x
dx difference:  0.0
```

0.4 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:

[ ]:
```
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:  5.8914366113629964e-09
dw error:  7.827423603874267e-10
db error:  1.4890948305639237e-11
```

[ ]:
```
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:  1.2388872084444232e-09
dw error:  1.1031508689606521e-09
db error:  1.7611801395017514e-11
```

0.5 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.

CNN-BatchNorm

March 4, 2024

0.1 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).

```
[ ]: ## Import and setups

     import time
     import numpy as np
     import matplotlib.pyplot as plt
     from nndl.conv_layers import *
     from utils.data_utils import get_CIFAR10_data
     from utils.gradient_check import eval_numerical_gradient,␣
      ↪eval_numerical_gradient_array
     from utils.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))))
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

0.2 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.

```
[ ]: # 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:  [10.06500229  9.76599828  9.92011893]
  Stds:  [3.63261098 3.93362063 3.21899582]
After spatial batch normalization:
  Shape:  (2, 3, 4, 5)
  Means:  [-8.46545056e-16 -3.42087469e-16 -5.38458167e-16]
  Stds:  [0.99999962 0.99999968 0.99999952]
After spatial batch normalization (nontrivial gamma, beta):
  Shape:  (2, 3, 4, 5)
  Means:  [6. 7. 8.]
  Stds:  [2.99999886 3.99999871 4.99999759]
```

0.3 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.

[ ]:
```
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)
```

```python
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:  3.4839051139063006e-07
dgamma error:  1.519950439910722e-11
dbeta error:  8.414216132785403e-12
```

[ ]:

```python

```

```
import numpy as np

from nndl.layers import *
from nndl.conv_layers import *
from utils.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.
# ================================================================ #
pad=1
stride=1
std = weight_scale
num_channel = input_dim[0]
w_in = input_dim[1]
self.params['W1'] = std * np.random.randn(num_filters, num_channel, \
                                          filter_size, filter_size)
self.params['b1'] = np.zeros(num_filters)
self.params['W2'] = std * np.random.randn(num_filters, num_filters, \
                                          filter_size, filter_size)
self.params['b2'] = np.zeros(num_filters)
self.bn_param_conv1 = {}
self.bn_param_conv2 = {}
self.bn_param_fc = {}

stride = 1
pad = int((filter_size - 1) / 2)

"""
w_conv_out = (w_in+2*pad-filter_size)/stride) + 1)
= (w_in+(filter_size - 1)- filter_size)/stride + 1
= (w_in-1)/stride + 1

-if stide==1, w_conv_out = w_in

w_mp = (w_in-pool_width)/stride) + 1)

-if stide==2, pool_width==2, w_mp = w_in/2
```

```python
    """

    w_conv = int((w_in-1)/stride + 1)
    w_mp = int(w_conv/2)
    h_mp = w_mp

    # output after mp: N*num_filters*h_mp*w_mp -> vectorize into N*(num_filters*h_mp*w_mp) -
    # FC1: N*hidden_dim
    self.params['W3'] = std * np.random.randn(num_filters*h_mp*w_mp, hidden_dim)
    self.params['b3'] = np.zeros(hidden_dim)
    self.params['W4'] = std * np.random.randn(hidden_dim, num_classes)
    self.params['b4'] = np.zeros(num_classes)

    if self.use_batchnorm == True:
      self.bn_param_conv1 = {'mode': 'train'}
      self.bn_param_conv2 = {'mode': 'train'}
      self.bn_param_fc = {'mode': 'train'}

      self.params['gamma1'] = np.ones(num_filters)
      self.params['beta1'] = np.zeros(num_filters)
      self.params['gamma2'] = np.ones(num_filters)
      self.params['beta2'] = np.zeros(num_filters)
      self.params['gamma3'] = np.ones(hidden_dim)
      self.params['beta3'] = np.zeros(hidden_dim)

    # ================================================================ #
    # 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']
    W4, b4 = self.params['W4'], self.params['b4']


    # 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}

"""
w_conv = (w_in+2*pad-filter_size)/stride) + 1)
= (w_in+(filter_size - 1)- filter_size)/stride + 1
= (w_in-1)/stride + 1

if stide==1, w_conv = w_in
"""

# pass pool_param to the forward pass for the max-pooling layer
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}
"""
w_mp = (w_in-pool_width)/stride) + 1)

if stide==2, pool_width==2
w_mp = w_in/2
"""
scores = None

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

# BN performs different between training and testing
mode = 'test' if y is None else 'train'
if self.use_batchnorm==True:
  self.bn_param_conv1["mode"] = mode
  self.bn_param_conv2["mode"] = mode
  self.bn_param_fc["mode"] = mode


# conv same - BN - relu - conv same - BN - relu - 2x2 max pool - affine - BN - relu - a
batch_size = X.shape[0]

conv1_out, cache_conv1 = conv_forward_fast(X, W1, b1, conv_param)

if self.use_batchnorm == True:
  gamma1, beta1 = self.params['gamma1'], self.params['beta1']
  bn_conv1, cache_bn_conv1 = spatial_batchnorm_forward(conv1_out, gamma1, \
                                                     beta1, self.bn_param_conv1)
  conv1_out = bn_conv1
```

```
conv_ReLU1, cache_conv_ReLU1 = relu_forward(conv1_out)
conv2_out, cache_conv2 = conv_forward_fast(conv_ReLU1, W2, b2, conv_param)

if self.use_batchnorm == True:
  gamma2, beta2 = self.params['gamma2'], self.params['beta2']
  bn_conv2, cache_bn_conv2 = spatial_batchnorm_forward(conv2_out, gamma2, \
                                                    beta2, self.bn_param_conv2)
  conv2_out = bn_conv2

conv_ReLU2, cache_conv_ReLU2 = relu_forward(conv2_out)
cnn_out, cache_cnn = max_pool_forward_fast(conv_ReLU2, pool_param)
cnn_vectorize = np.reshape(cnn_out, [batch_size, -1])

fc1_out, cache_fc1 = affine_forward(cnn_vectorize, W3, b3)

if self.use_batchnorm == True:
  gamma3, beta3 = self.params['gamma3'], self.params['beta3']
  bn_fc, cache_bn_fc = batchnorm_forward(fc1_out, gamma3, beta3, self.bn_param_fc)
  fc1_out = bn_fc

fc1_ReLU, cache_fc1_ReLU = relu_forward(fc1_out)

fc2_out, cache_fc2 = affine_forward(fc1_ReLU, W4, b4)
scores = fc2_out

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

# If in test mode:
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.
# ================================================================ #

# sf_loss - softmax - affine - relu - BN - affine - 2x2 max pool - relu - BN - conv same

sf_loss, grad_softmax = softmax_loss(scores, y)   # y as true labels
reg_loss = 0.5*self.reg*(np.linalg.norm(W1)**2 \
```

```
                         + np.linalg.norm(W2)**2\
                         + np.linalg.norm(W3)**2\
                         + np.linalg.norm(W4)**2)

loss = sf_loss + reg_loss

""" Forward Pass
# conv same - BN - relu - conv same - BN - relu - 2x2 max pool - affine - BN - relu - a
batch_size = X.shape[0]

conv1_out, cache_conv1 = conv_forward_fast(X, W1, b1, conv_param)

if self.use_batchnorm == True:
  gamma1, beta1 = self.params['gamma1'], self.params['beta1']
  bn_conv1, cache_bn_conv1 = spatial_batchnorm_forward(conv1_out, gamma1, \
                                                        beta1, self.bn_param_conv1)
  conv1_out = bn_conv1

conv_ReLU1, cache_conv_ReLU1 = relu_forward(conv1_out)
conv2_out, cache_conv2 = conv_forward_fast(conv_ReLU1, W2, b2, conv_param)

if self.use_batchnorm == True:
  gamma2, beta2 = self.params['gamma2'], self.params['beta2']
  bn_conv2, cache_bn_conv2 = spatial_batchnorm_forward(conv2_out, gamma2, \
                                                        beta2, self.bn_param_conv2)
  conv2_out = bn_conv2

conv_ReLU2, cache_conv_ReLU2 = relu_forward(conv2_out)
cnn_out, cache_cnn = max_pool_forward_fast(conv_ReLU2, pool_param)
cnn_vectorize = np.reshape(cnn_out, [batch_size, -1])

fc1_out, cache_fc1 = affine_forward(cnn_vectorize, W3, b3)

if self.use_batchnorm == True:
  gamma3, beta3 = self.params['gamma3'], self.params['beta3']
  bn_fc, cache_bn_fc = batchnorm_forward(fc1_out, gamma3, beta3, self.bn_param_fc)
  fc1_out = bn_fc

fc1_ReLU, cache_fc1_ReLU = relu_forward(fc1_out)

fc2_out, cache_fc2 = affine_forward(fc1_ReLU, W4, b4)
scores = fc2_out
"""

grad_fc1_ReLU, grad_W4, grad_b4 = affine_backward(grad_softmax, cache_fc2)
grad_fc1_out = relu_backward(grad_fc1_ReLU, cache_fc1_ReLU)
```

```
if self.use_batchnorm == True:
  grad_bn_fc, grad_gamma3, grad_beta3 = batchnorm_backward(grad_fc1_out, cache_bn_fc)
  grad_fc1_out = grad_bn_fc

grad_cnn_vectorize, grad_W3, grad_b3 = affine_backward(grad_fc1_out, cache_fc1)

# grad_cnn_vectorize shape: (N, f*h_mp*w_mp)
grad_cnn = np.reshape(grad_cnn_vectorize, cnn_out.shape)
grad_conv_ReLU2 = max_pool_backward_fast(grad_cnn, cache_cnn)
grad_conv_out2 = relu_backward(grad_conv_ReLU2, cache_conv_ReLU2)

if self.use_batchnorm == True:
  grad_bn_conv2, grad_gamma2, grad_beta2 = spatial_batchnorm_backward(grad_conv_out2, \
                                                                       cache_bn_conv2)
  grad_conv_out2 = grad_bn_conv2

grad_conv_ReLU1, grad_W2, grad_b2 = conv_backward_fast(grad_conv_out2, cache_conv2)
grad_conv_out1 = relu_backward(grad_conv_ReLU1, cache_conv_ReLU1)

if self.use_batchnorm == True:
  grad_bn_conv1, grad_gamma1, grad_beta1 = spatial_batchnorm_backward(grad_conv_out1, \
                                                                       cache_bn_conv1)
  grad_conv_out1 = grad_bn_conv1

grad_X, grad_W1, grad_b1 = conv_backward_fast(grad_conv_out1, cache_conv1)


grad_W1_reg = 0.5*self.reg*2*W1
grad_W2_reg = 0.5*self.reg*2*W2
grad_W3_reg = 0.5*self.reg*2*W3
grad_W4_reg = 0.5*self.reg*2*W4

grads['W4'] = grad_W4 + grad_W4_reg
grads['b4'] = grad_b4
grads['W3'] = grad_W3 + grad_W3_reg
grads['b3'] = grad_b3
grads['W2'] = grad_W2 + grad_W2_reg
grads['b2'] = grad_b2
grads['W1'] = grad_W1 + grad_W1_reg
grads['b1'] = grad_b1

if self.use_batchnorm == True:
  grads['gamma3'] = grad_gamma3
  grads['beta3'] = grad_beta3
  grads['gamma2'] = grad_gamma2
```

```python
    grads['beta2'] = grad_beta2
    grads['gamma1'] = grad_gamma1
    grads['beta1'] = grad_beta1

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

  return loss, grads
```

CNN

March 5, 2024

1 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.

[ ]:
```
# As usual, a bit of setup
import numpy as np
import matplotlib.pyplot as plt
from nndl.cnn import *
from utils.data_utils import get_CIFAR10_data
from utils.gradient_check import eval_numerical_gradient_array,␣
  ↪eval_numerical_gradient
from nndl.layers import *
from nndl.conv_layers import *
from utils.fast_layers import *
from utils.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))))
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

[ ]:
```
# 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,)
```

1.1 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.

```
[ ]: 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])))
```

```
W1 max relative error: 0.00021540464795099533
W2 max relative error: 0.0017648686770178317
W3 max relative error: 0.00010071024924608103
b1 max relative error: 2.948396033351636e-05
b2 max relative error: 4.570476253477182e-07
b3 max relative error: 1.7777958838980442e-09
```

1.1.1 Overfit small dataset

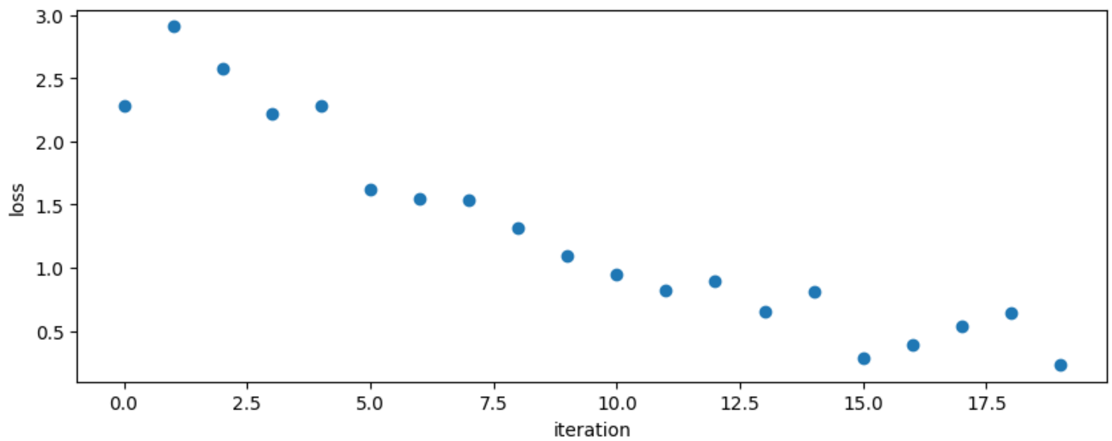
To check your CNN implementation, let's overfit a small dataset.

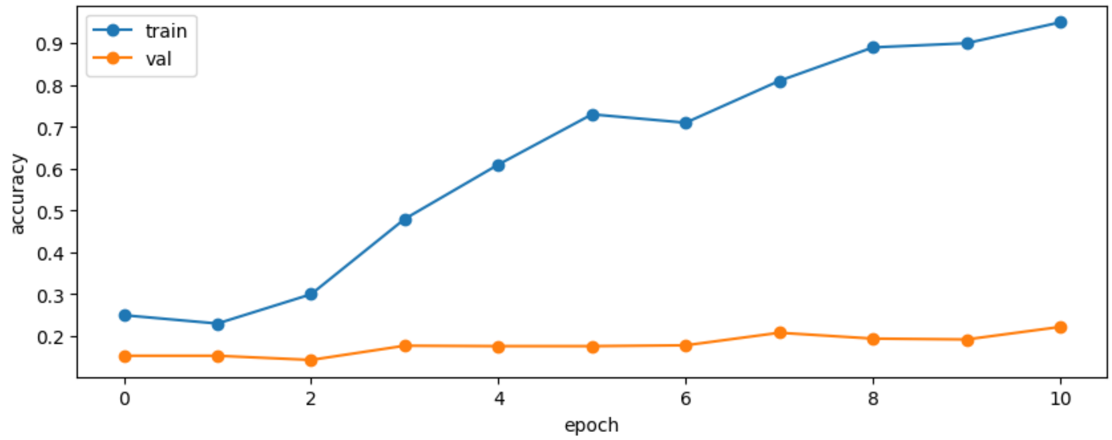
```
[ ]: num_train = 100
small_data = {
  'X_train': data['X_train'][:num_train],
  'y_train': data['y_train'][:num_train],
  'X_val': data['X_val'],
  'y_val': data['y_val'],
}

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()
```

```
(Iteration 1 / 20) loss: 2.280843
(Epoch 0 / 10) train acc: 0.250000; val_acc: 0.153000
(Iteration 2 / 20) loss: 2.909015
(Epoch 1 / 10) train acc: 0.230000; val_acc: 0.153000
(Iteration 3 / 20) loss: 2.574897
(Iteration 4 / 20) loss: 2.216059
(Epoch 2 / 10) train acc: 0.300000; val_acc: 0.143000
(Iteration 5 / 20) loss: 2.280850
(Iteration 6 / 20) loss: 1.618330
(Epoch 3 / 10) train acc: 0.480000; val_acc: 0.177000
(Iteration 7 / 20) loss: 1.550600
(Iteration 8 / 20) loss: 1.533238
(Epoch 4 / 10) train acc: 0.610000; val_acc: 0.176000
(Iteration 9 / 20) loss: 1.319464
(Iteration 10 / 20) loss: 1.092524
(Epoch 5 / 10) train acc: 0.730000; val_acc: 0.176000
(Iteration 11 / 20) loss: 0.952148
(Iteration 12 / 20) loss: 0.823965
(Epoch 6 / 10) train acc: 0.710000; val_acc: 0.178000
(Iteration 13 / 20) loss: 0.894186
(Iteration 14 / 20) loss: 0.654674
(Epoch 7 / 10) train acc: 0.810000; val_acc: 0.208000
(Iteration 15 / 20) loss: 0.812115
(Iteration 16 / 20) loss: 0.287689
(Epoch 8 / 10) train acc: 0.890000; val_acc: 0.194000
(Iteration 17 / 20) loss: 0.386654
(Iteration 18 / 20) loss: 0.537146
(Epoch 9 / 10) train acc: 0.900000; val_acc: 0.192000
(Iteration 19 / 20) loss: 0.641776
(Iteration 20 / 20) loss: 0.233192
(Epoch 10 / 10) train acc: 0.950000; val_acc: 0.222000
```

[ ]:

```
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()
```





1.2 Train the network

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

[ ]:
```
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()
```

```
(Iteration 1 / 980) loss: 2.304939
(Epoch 0 / 1) train acc: 0.117000; val_acc: 0.128000
(Iteration 21 / 980) loss: 2.183147
(Iteration 41 / 980) loss: 2.142793
(Iteration 61 / 980) loss: 1.891545
```

```
(Iteration 81 / 980) loss: 1.915329
(Iteration 101 / 980) loss: 1.781053
(Iteration 121 / 980) loss: 2.304722
(Iteration 141 / 980) loss: 1.830852
(Iteration 161 / 980) loss: 1.454334
(Iteration 181 / 980) loss: 1.887367
(Iteration 201 / 980) loss: 1.832030
(Iteration 221 / 980) loss: 1.751510
(Iteration 241 / 980) loss: 2.039439
(Iteration 261 / 980) loss: 1.837052
(Iteration 281 / 980) loss: 1.705089
(Iteration 301 / 980) loss: 1.652323
(Iteration 321 / 980) loss: 1.613009
(Iteration 341 / 980) loss: 1.644134
(Iteration 361 / 980) loss: 1.528704
(Iteration 381 / 980) loss: 1.608708
(Iteration 401 / 980) loss: 1.872454
(Iteration 421 / 980) loss: 1.678102
(Iteration 441 / 980) loss: 1.621624
(Iteration 461 / 980) loss: 1.637929
(Iteration 481 / 980) loss: 1.488162
(Iteration 501 / 980) loss: 1.794451
(Iteration 521 / 980) loss: 1.693814
(Iteration 541 / 980) loss: 1.609194
(Iteration 561 / 980) loss: 1.445277
(Iteration 581 / 980) loss: 1.670523
(Iteration 601 / 980) loss: 1.388795
(Iteration 621 / 980) loss: 1.620898
(Iteration 641 / 980) loss: 1.512084
(Iteration 661 / 980) loss: 1.521381
(Iteration 681 / 980) loss: 1.311462
(Iteration 701 / 980) loss: 2.015614
(Iteration 721 / 980) loss: 1.484359
(Iteration 741 / 980) loss: 1.610283
(Iteration 761 / 980) loss: 1.787248
(Iteration 781 / 980) loss: 1.715426
(Iteration 801 / 980) loss: 1.346103
(Iteration 821 / 980) loss: 1.622872
(Iteration 841 / 980) loss: 1.334004
(Iteration 861 / 980) loss: 1.204134
(Iteration 881 / 980) loss: 1.426710
(Iteration 901 / 980) loss: 1.547595
(Iteration 921 / 980) loss: 1.681570
(Iteration 941 / 980) loss: 1.627250
(Iteration 961 / 980) loss: 1.575646
(Epoch 1 / 1) train acc: 0.440000; val_acc: 0.449000
```

2 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.

2.0.1 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]

2.0.2 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.

[ ]:
```
model = ThreeLayerConvNet(filter_size=3, num_filters=32, weight_scale=0.001,␣
 ↪hidden_dim=500, reg=0.001, use_batchnorm=True)

solver = Solver(model, data,
                num_epochs=2, batch_size=200,
                update_rule='adam',
                optim_config={
                  'learning_rate': 5e-4,
                },
                verbose=True, print_every=50)
solver.train()
```

```
(Iteration 1 / 490) loss: 2.304937
(Epoch 0 / 2) train acc: 0.114000; val_acc: 0.105000
(Iteration 51 / 490) loss: 1.502007
(Iteration 101 / 490) loss: 1.399396
(Iteration 151 / 490) loss: 1.274546
(Iteration 201 / 490) loss: 1.148036
```

```
(Epoch 1 / 2) train acc: 0.658000; val_acc: 0.628000
(Iteration 251 / 490) loss: 1.067008
(Iteration 301 / 490) loss: 1.059585
(Iteration 351 / 490) loss: 0.912677
(Iteration 401 / 490) loss: 0.952039
(Iteration 451 / 490) loss: 0.836146
(Epoch 2 / 2) train acc: 0.722000; val_acc: 0.663000
```

[ ]:

```

```