# Optimization

February 19, 2024

## 0.1 Optimization for Fully Connected Networks

In this notebook, we will implement different optimization rules for gradient descent. We have provided starter code; however, you will need to copy and paste your code from your implementation of the modular fully connected nets in HW #3 to build upon this.

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, and their layer structure. 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.fc_net 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
```

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

## 0.2 Building upon your HW #3 implementation

Copy and paste the following functions from your HW #3 implementation of a modular FC net:

- `affine_forward` in `nndl/layers.py`
- `affine_backward` in `nndl/layers.py`
- `relu_forward` in `nndl/layers.py`
- `relu_backward` in `nndl/layers.py`
- `affine_relu_forward` in `nndl/layer_utils.py`
- `affine_relu_backward` in `nndl/layer_utils.py`
- The `FullyConnectedNet` class in `nndl/fc_net.py`

### 0.2.1 Test all functions you copy and pasted

[ ]:
```
from nndl.layer_tests import *

affine_forward_test(); print('\n')
affine_backward_test(); print('\n')
relu_forward_test(); print('\n')
relu_backward_test(); print('\n')
affine_relu_test(); print('\n')
fc_net_test()
```

```
If affine_forward function is working, difference should be less than 1e-9:
difference: 9.769849468192957e-10


If affine_backward is working, error should be less than 1e-9::
dx error: 3.870669242367056e-10
dw error: 7.55251788880157e-11
db error: 1.0559846140201009e-10


If relu_forward function is working, difference should be around 1e-8:
```

```
difference: 4.999999798022158e-08


If relu_forward function is working, error should be less than 1e-9:
dx error: 3.2756290267730722e-12


If affine_relu_forward and affine_relu_backward are working, error should be
less than 1e-9::
dx error: 3.0538569541965475e-10
dw error: 1.6914953492795634e-10
db error: 2.547999424127119e-11


Running check with reg = 0
Initial loss: 2.3044244177593463
W1 relative error: 1.6913217723944727e-05
W2 relative error: 1.6425215214059696e-07
W3 relative error: 4.440049895524679e-07
b1 relative error: 8.956773802496012e-09
b2 relative error: 1.6373257502162923e-09
b3 relative error: 1.52408848384591e-10
Running check with reg = 3.14
Initial loss: 6.735019105225316
W1 relative error: 1.7015461836071512e-08
W2 relative error: 6.471376929322069e-08
W3 relative error: 1.3769114981367238e-08
b1 relative error: 5.428556316741595e-08
b2 relative error: 8.265024587950067e-09
b3 relative error: 1.4788551783306754e-10
```

# 1 Training a larger model

In general, proceeding with vanilla stochastic gradient descent to optimize models may be fraught with problems and limitations, as discussed in class. Thus, we implement optimizers that improve on SGD.

## 1.1 SGD + momentum

In the following section, implement SGD with momentum. Read the `nndl/optim.py` API, which is provided by CS231n, and be sure you understand it. After, implement `sgd_momentum` in `nndl/optim.py`. Test your implementation of `sgd_momentum` by running the cell below.

[ ]:
```
from nndl.optim import sgd_momentum

N, D = 4, 5
w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
```

```
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)
v = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-3, 'velocity': v}
next_w, _ = sgd_momentum(w, dw, config=config)

expected_next_w = np.asarray([
  [ 0.1406,      0.20738947,  0.27417895,  0.34096842,  0.40775789],
  [ 0.47454737,  0.54133684,  0.60812632,  0.67491579,  0.74170526],
  [ 0.80849474,  0.87528421,  0.94207368,  1.00886316,  1.07565263],
  [ 1.14244211,  1.20923158,  1.27602105,  1.34281053,  1.4096    ]])
expected_velocity = np.asarray([
  [ 0.5406,      0.55475789,  0.56891579, 0.58307368,  0.59723158],
  [ 0.61138947,  0.62554737,  0.63970526,  0.65386316,  0.66802105],
  [ 0.68217895,  0.69633684,  0.71049474,  0.72465263,  0.73881053],
  [ 0.75296842,  0.76712632,  0.78128421,  0.79544211,  0.8096    ]])

print('next_w error: {}'.format(rel_error(next_w, expected_next_w)))
print('velocity error: {}'.format(rel_error(expected_velocity,␣
 ↪config['velocity'])))
```

```
next_w error: 8.882347033505819e-09
velocity error: 4.269287743278663e-09
```

### 1.2 SGD + Nesterov momentum

Implement `sgd_nesterov_momentum` in `ndl/optim.py`.

[ ]:
```
from nndl.optim import sgd_nesterov_momentum

N, D = 4, 5
w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)
v = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-3, 'velocity': v}
next_w, _ = sgd_nesterov_momentum(w, dw, config=config)

expected_next_w = np.asarray([
  [0.08714,      0.15246105,  0.21778211,  0.28310316,  0.34842421],
  [0.41374526,   0.47906632,  0.54438737,  0.60970842,  0.67502947],
  [0.74035053,   0.80567158,  0.87099263,  0.93631368,  1.00163474],
  [1.06695579,   1.13227684,  1.19759789,  1.26291895,  1.32824   ]])
expected_velocity = np.asarray([
  [ 0.5406,      0.55475789,  0.56891579,  0.58307368,  0.59723158],
  [ 0.61138947,  0.62554737,  0.63970526,  0.65386316,  0.66802105],
  [ 0.68217895,  0.69633684,  0.71049474,  0.72465263,  0.73881053],
  [ 0.75296842,  0.76712632,  0.78128421,  0.79544211,  0.8096    ]])
```

```
print('next_w error: {}'.format(rel_error(next_w, expected_next_w)))
print('velocity error: {}'.format(rel_error(expected_velocity,␣
 ↪config['velocity'])))
```

```
next_w error: 1.0875186845081027e-08
velocity error: 4.269287743278663e-09
```

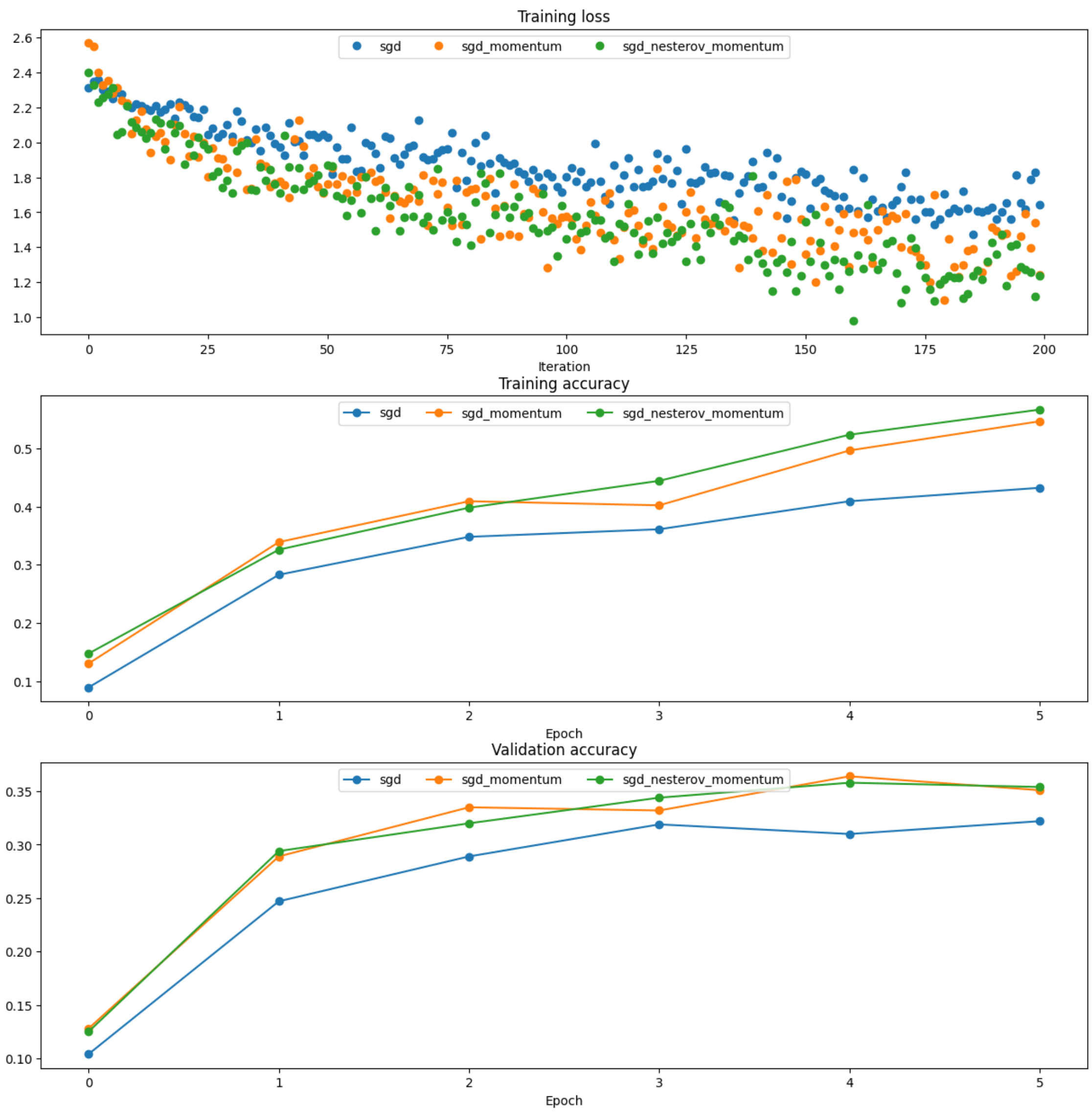
### 1.3 Evaluating SGD, SGD+Momentum, and SGD+NesterovMomentum

Run the following cell to train a 6 layer FC net with SGD, SGD+momentum, and SGD+Nesterov momentum. You should see that SGD+momentum achieves a better loss than SGD, and that SGD+Nesterov momentum achieves a slightly better loss (and training accuracy) than SGD+momentum.

[ ]:
```
num_train = 4000
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'],
}

solvers = {}

for update_rule in ['sgd', 'sgd_momentum', 'sgd_nesterov_momentum']:
  print('Optimizing with {}'.format(update_rule))
  model = FullyConnectedNet([100, 100, 100, 100, 100], weight_scale=5e-2)

  solver = Solver(model, small_data,
                  num_epochs=5, batch_size=100,
                  update_rule=update_rule,
                  optim_config={
                    'learning_rate': 1e-2,
                  },
                  verbose=False)
  solvers[update_rule] = solver
  solver.train()
  print

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')
```

```python
plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

for update_rule, solver in solvers.items():
  plt.subplot(3, 1, 1)
  plt.plot(solver.loss_history, 'o', label=update_rule)

  plt.subplot(3, 1, 2)
  plt.plot(solver.train_acc_history, '-o', label=update_rule)

  plt.subplot(3, 1, 3)
  plt.plot(solver.val_acc_history, '-o', label=update_rule)

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

```
Optimizing with sgd
Optimizing with sgd_momentum
Optimizing with sgd_nesterov_momentum
```




## 1.4 RMSProp

Now we go to techniques that adapt the gradient. Implement `rmsprop` in `nndl/optim.py`. Test your implementation by running the cell below.

[ ]:
```
from nndl.optim import rmsprop

N, D = 4, 5
w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)
a = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-2, 'a': a}
```

```python
next_w, _ = rmsprop(w, dw, config=config)

expected_next_w = np.asarray([
  [-0.39223849, -0.34037513, -0.28849239, -0.23659121, -0.18467247],
  [-0.132737,   -0.08078555, -0.02881884,  0.02316247,  0.07515774],
  [ 0.12716641,  0.17918792,  0.23122175,  0.28326742,  0.33532447],
  [ 0.38739248,  0.43947102,  0.49155973,  0.54365823,  0.59576619]])
expected_cache = np.asarray([
  [ 0.5976,      0.6126277,   0.6277108,   0.64284931,  0.65804321],
  [ 0.67329252,  0.68859723,  0.70395734,  0.71937285,  0.73484377],
  [ 0.75037008,  0.7659518,   0.78158892,  0.79728144,  0.81302936],
  [ 0.82883269,  0.84469141,  0.86060554,  0.87657507,  0.8926    ]])

print('next_w error: {}'.format(rel_error(expected_next_w, next_w)))
print('cache error: {}'.format(rel_error(expected_cache, config['a'])))
```

```
next_w error: 9.524687511038133e-08
cache error: 2.6477955807156126e-09
```

### 1.5 Adaptive moments

Now, implement `adam` in `nndl/optim.py`. Test your implementation by running the cell below.

[ ]:
```python
# Test Adam implementation; you should see errors around 1e-7 or less
from nndl.optim import adam

N, D = 4, 5
w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)
v = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)
a = np.linspace(0.7, 0.5, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-2, 'v': v, 'a': a, 't': 5}
next_w, _ = adam(w, dw, config=config)

expected_next_w = np.asarray([
  [-0.40094747, -0.34836187, -0.29577703, -0.24319299, -0.19060977],
  [-0.1380274,  -0.08544591, -0.03286534,  0.01971428,  0.0722929],
  [ 0.1248705,   0.17744702,  0.23002243,  0.28259667,  0.33516969],
  [ 0.38774145,  0.44031188,  0.49288093,  0.54544852,  0.59801459]])
expected_a = np.asarray([
  [ 0.69966,     0.68908382,  0.67851319,  0.66794809,  0.65738853,],
  [ 0.64683452,  0.63628604,  0.6257431,   0.61520571,  0.60467385,],
  [ 0.59414753,  0.58362676,  0.57311152,  0.56260183,  0.55209767,],
  [ 0.54159906,  0.53110598,  0.52061845,  0.51013645,  0.49966,   ]])
expected_v = np.asarray([
  [ 0.48,        0.49947368,  0.51894737,  0.53842105,  0.55789474],
  [ 0.57736842,  0.59684211,  0.61631579,  0.63578947,  0.65526316],
```

```
 [ 0.67473684,  0.69421053,  0.71368421,  0.73315789,  0.75263158],
 [ 0.77210526,  0.79157895,  0.81105263,  0.83052632,  0.85      ]])

print('next_w error: {}'.format(rel_error(expected_next_w, next_w)))
print('a error: {}'.format(rel_error(expected_a, config['a'])))
print('v error: {}'.format(rel_error(expected_v, config['v'])))
```

```
next_w error: 1.1395691798535431e-07
a error: 4.208314038113071e-09
v error: 4.214963193114416e-09
```

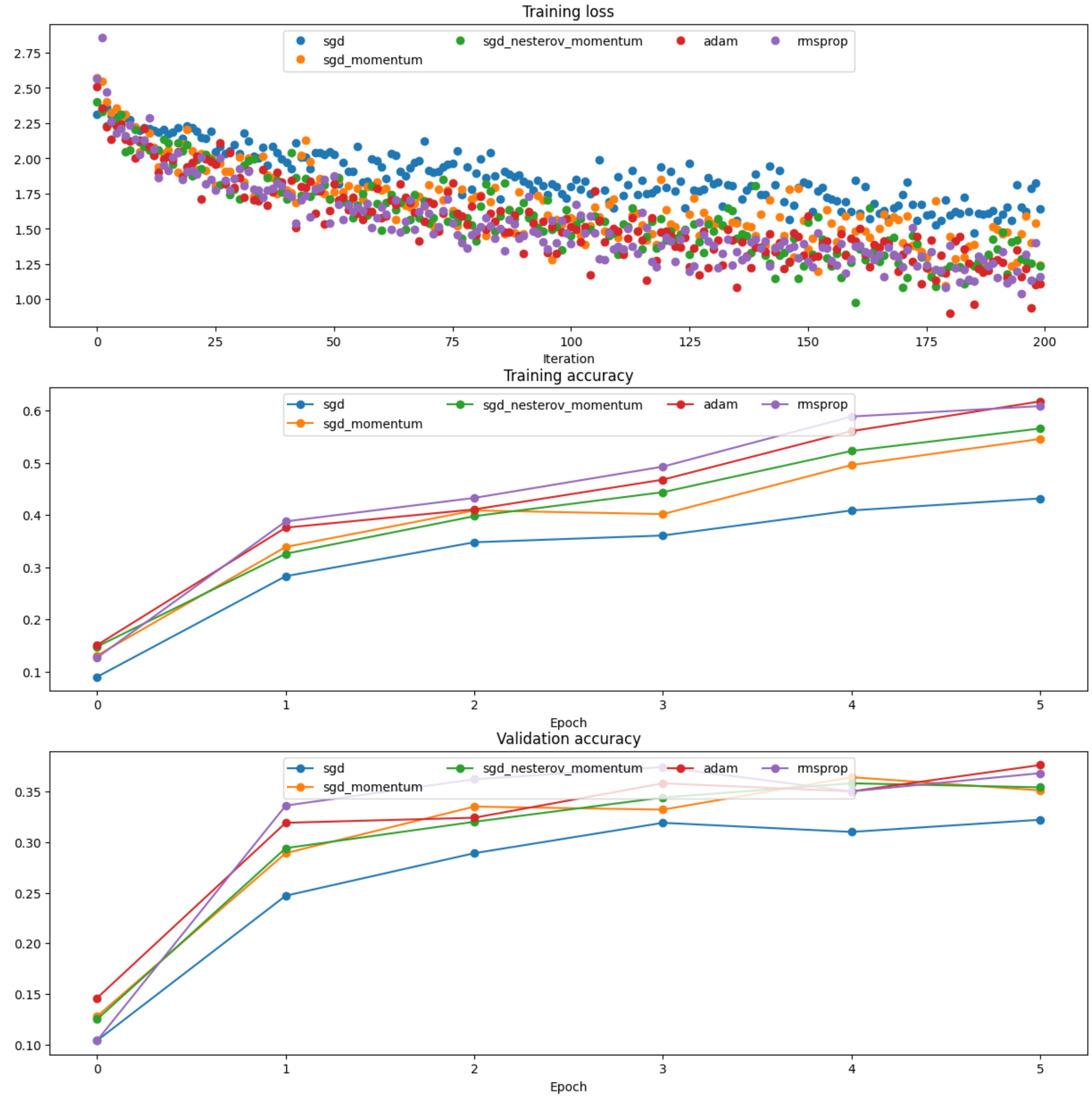
### 1.6 Comparing SGD, SGD+NesterovMomentum, RMSProp, and Adam

The following code will compare optimization with SGD, Momentum, Nesterov Momentum, RMSProp and Adam. In our code, we find that RMSProp, Adam, and SGD + Nesterov Momentum achieve approximately the same training error after a few training epochs.

[ ]:
```
learning_rates = {'rmsprop': 2e-4, 'adam': 1e-3}

for update_rule in ['adam', 'rmsprop']:
  print('Optimizing with {}'.format(update_rule))
  model = FullyConnectedNet([100, 100, 100, 100, 100], weight_scale=5e-2)

  solver = Solver(model, small_data,
                  num_epochs=5, batch_size=100,
                  update_rule=update_rule,
                  optim_config={
                    'learning_rate': learning_rates[update_rule]
                  },
                  verbose=False)
  solvers[update_rule] = solver
  solver.train()
  print

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

for update_rule, solver in solvers.items():
  plt.subplot(3, 1, 1)
```

```
  plt.plot(solver.loss_history, 'o', label=update_rule)

  plt.subplot(3, 1, 2)
  plt.plot(solver.train_acc_history, '-o', label=update_rule)

  plt.subplot(3, 1, 3)
  plt.plot(solver.val_acc_history, '-o', label=update_rule)

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

```
Optimizing with adam
Optimizing with rmsprop
```




## 1.7 Easier optimization

In the following cell, we'll train a 4 layer neural network having 500 units in each hidden layer with the different optimizers, and find that it is far easier to get up to 50+% performance on CIFAR-10. After we implement batchnorm and dropout, we'll ask you to get 55+% on CIFAR-10.

[ ]:
```
optimizer = 'adam'
best_model = None

layer_dims = [500, 500, 500]
weight_scale = 0.01
learning_rate = 1e-3
lr_decay = 0.9
```

```python
model = FullyConnectedNet(layer_dims, weight_scale=weight_scale,
                          use_batchnorm=True)

solver = Solver(model, data,
                num_epochs=10, batch_size=100,
                update_rule=optimizer,
                optim_config={
                  'learning_rate': learning_rate,
                },
                lr_decay=lr_decay,
                verbose=True, print_every=50)
solver.train()
```

```
(Iteration 1 / 4900) loss: 2.311873
(Epoch 0 / 10) train acc: 0.166000; val_acc: 0.146000
(Iteration 51 / 4900) loss: 2.026356
(Iteration 101 / 4900) loss: 1.794955
(Iteration 151 / 4900) loss: 1.733056
(Iteration 201 / 4900) loss: 1.491804
(Iteration 251 / 4900) loss: 1.862906
(Iteration 301 / 4900) loss: 1.690605
(Iteration 351 / 4900) loss: 1.847450
(Iteration 401 / 4900) loss: 1.664694
(Iteration 451 / 4900) loss: 1.732783
(Epoch 1 / 10) train acc: 0.422000; val_acc: 0.429000
(Iteration 501 / 4900) loss: 1.564308
(Iteration 551 / 4900) loss: 1.477041
(Iteration 601 / 4900) loss: 1.373680
(Iteration 651 / 4900) loss: 1.631575
(Iteration 701 / 4900) loss: 1.574071
(Iteration 751 / 4900) loss: 1.645297
(Iteration 801 / 4900) loss: 1.539631
(Iteration 851 / 4900) loss: 1.395206
(Iteration 901 / 4900) loss: 1.451352
(Iteration 951 / 4900) loss: 1.448472
(Epoch 2 / 10) train acc: 0.487000; val_acc: 0.466000
(Iteration 1001 / 4900) loss: 1.459766
(Iteration 1051 / 4900) loss: 1.487611
(Iteration 1101 / 4900) loss: 1.459427
(Iteration 1151 / 4900) loss: 1.558609
(Iteration 1201 / 4900) loss: 1.380545
(Iteration 1251 / 4900) loss: 1.354350
(Iteration 1301 / 4900) loss: 1.565081
(Iteration 1351 / 4900) loss: 1.123233
(Iteration 1401 / 4900) loss: 1.613201
(Iteration 1451 / 4900) loss: 1.429499
(Epoch 3 / 10) train acc: 0.506000; val_acc: 0.478000
```

```
(Iteration 1501 / 4900) loss: 1.410010
(Iteration 1551 / 4900) loss: 1.298708
(Iteration 1601 / 4900) loss: 1.143147
(Iteration 1651 / 4900) loss: 1.442402
(Iteration 1701 / 4900) loss: 1.323416
(Iteration 1751 / 4900) loss: 1.369062
(Iteration 1801 / 4900) loss: 1.363227
(Iteration 1851 / 4900) loss: 1.242166
(Iteration 1901 / 4900) loss: 1.290352
(Iteration 1951 / 4900) loss: 1.423709
(Epoch 4 / 10) train acc: 0.542000; val_acc: 0.499000
(Iteration 2001 / 4900) loss: 1.264606
(Iteration 2051 / 4900) loss: 1.352938
(Iteration 2101 / 4900) loss: 1.451440
(Iteration 2151 / 4900) loss: 1.492823
(Iteration 2201 / 4900) loss: 1.201178
(Iteration 2251 / 4900) loss: 1.387044
(Iteration 2301 / 4900) loss: 1.331332
(Iteration 2351 / 4900) loss: 1.308441
(Iteration 2401 / 4900) loss: 1.116480
(Epoch 5 / 10) train acc: 0.566000; val_acc: 0.493000
(Iteration 2451 / 4900) loss: 1.122886
(Iteration 2501 / 4900) loss: 1.094335
(Iteration 2551 / 4900) loss: 1.116139
(Iteration 2601 / 4900) loss: 1.055144
(Iteration 2651 / 4900) loss: 1.009194
(Iteration 2701 / 4900) loss: 1.185946
(Iteration 2751 / 4900) loss: 1.180233
(Iteration 2801 / 4900) loss: 1.255893
(Iteration 2851 / 4900) loss: 1.109786
(Iteration 2901 / 4900) loss: 1.152897
(Epoch 6 / 10) train acc: 0.571000; val_acc: 0.522000
(Iteration 2951 / 4900) loss: 1.117053
(Iteration 3001 / 4900) loss: 1.346133
(Iteration 3051 / 4900) loss: 1.142815
(Iteration 3101 / 4900) loss: 1.217458
(Iteration 3151 / 4900) loss: 1.134366
(Iteration 3201 / 4900) loss: 1.081430
(Iteration 3251 / 4900) loss: 1.143034
(Iteration 3301 / 4900) loss: 1.039814
(Iteration 3351 / 4900) loss: 1.162844
(Iteration 3401 / 4900) loss: 1.006388
(Epoch 7 / 10) train acc: 0.611000; val_acc: 0.534000
(Iteration 3451 / 4900) loss: 1.219256
(Iteration 3501 / 4900) loss: 0.907064
(Iteration 3551 / 4900) loss: 1.174738
(Iteration 3601 / 4900) loss: 1.051233
(Iteration 3651 / 4900) loss: 0.958345
```

```
(Iteration 3701 / 4900) loss: 1.052089
(Iteration 3751 / 4900) loss: 1.089412
(Iteration 3801 / 4900) loss: 1.190105
(Iteration 3851 / 4900) loss: 1.135305
(Iteration 3901 / 4900) loss: 0.864292
(Epoch 8 / 10) train acc: 0.626000; val_acc: 0.533000
(Iteration 3951 / 4900) loss: 1.095561
(Iteration 4001 / 4900) loss: 1.068163
(Iteration 4051 / 4900) loss: 0.989233
(Iteration 4101 / 4900) loss: 1.102437
(Iteration 4151 / 4900) loss: 1.015528
(Iteration 4201 / 4900) loss: 0.921740
(Iteration 4251 / 4900) loss: 1.012284
(Iteration 4301 / 4900) loss: 0.931702
(Iteration 4351 / 4900) loss: 0.952177
(Iteration 4401 / 4900) loss: 1.035978
(Epoch 9 / 10) train acc: 0.652000; val_acc: 0.516000
(Iteration 4451 / 4900) loss: 1.080768
(Iteration 4501 / 4900) loss: 1.004819
(Iteration 4551 / 4900) loss: 1.079557
(Iteration 4601 / 4900) loss: 0.880005
(Iteration 4651 / 4900) loss: 0.849844
(Iteration 4701 / 4900) loss: 0.896638
(Iteration 4751 / 4900) loss: 0.973324
(Iteration 4801 / 4900) loss: 0.987399
(Iteration 4851 / 4900) loss: 0.899170
(Epoch 10 / 10) train acc: 0.662000; val_acc: 0.513000
```

[ ]:
```
y_test_pred = np.argmax(model.loss(data['X_test']), axis=1)
y_val_pred = np.argmax(model.loss(data['X_val']), axis=1)
print('Validation set accuracy: {}'.format(np.mean(y_val_pred ==␣
 ↪data['y_val'])))
print('Test set accuracy: {}'.format(np.mean(y_test_pred == data['y_test'])))
```

```
Validation set accuracy: 0.534
Test set accuracy: 0.529
```

```python
import numpy as np

"""
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.
"""

"""
This file implements various first-order update rules that are commonly used for
training neural networks. Each update rule accepts current weights and the
gradient of the loss with respect to those weights and produces the next set of
weights. Each update rule has the same interface:

def update(w, dw, config=None):

Inputs:
  - w: A numpy array giving the current weights.
  - dw: A numpy array of the same shape as w giving the gradient of the
    loss with respect to w.
  - config: A dictionary containing hyperparameter values such as learning rate,
    momentum, etc. If the update rule requires caching values over many
    iterations, then config will also hold these cached values.

Returns:
  - next_w: The next point after the update.
  - config: The config dictionary to be passed to the next iteration of the
    update rule.

NOTE: For most update rules, the default learning rate will probably not perform
well; however the default values of the other hyperparameters should work well
for a variety of different problems.

For efficiency, update rules may perform in-place updates, mutating w and
setting next_w equal to w.
"""


def sgd(w, dw, config=None):
  """
  Performs vanilla stochastic gradient descent.

  config format:
  - learning_rate: Scalar learning rate.
  """
  if config is None: config = {}
  config.setdefault('learning_rate', 1e-2)

  w -= config['learning_rate'] * dw
  return w, config


def sgd_momentum(w, dw, config=None):
  """
  Performs stochastic gradient descent with momentum.

  config format:
  - learning_rate: Scalar learning rate.
  - momentum: Scalar between 0 and 1 giving the momentum value.
    Setting momentum = 0 reduces to sgd.
  - velocity: A numpy array of the same shape as w and dw used to store a moving
    average of the gradients.
  """
```

```python
  if config is None: config = {}
  config.setdefault('learning_rate', 1e-2)
  config.setdefault('momentum', 0.9) # set momentum to 0.9 if it wasn't there
  v = config.get('velocity', np.zeros_like(w))   # gets velocity, else sets it to zero.

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the momentum update formula.  Return the updated weights
  #   as next_w, and the updated velocity as v.
  # ================================================================ #

  v = config['momentum'] * v - config['learning_rate'] * dw
  next_w = w + v

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

  config['velocity'] = v

  return next_w, config

def sgd_nesterov_momentum(w, dw, config=None):
  """
  Performs stochastic gradient descent with Nesterov momentum.

  config format:
  - learning_rate: Scalar learning rate.
  - momentum: Scalar between 0 and 1 giving the momentum value.
    Setting momentum = 0 reduces to sgd.
  - velocity: A numpy array of the same shape as w and dw used to store a moving
    average of the gradients.
  """
  if config is None: config = {}
  config.setdefault('learning_rate', 1e-2)
  config.setdefault('momentum', 0.9) # set momentum to 0.9 if it wasn't there
  v = config.get('velocity', np.zeros_like(w))   # gets velocity, else sets it to zero.

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the momentum update formula.  Return the updated weights
  #   as next_w, and the updated velocity as v.
  # ================================================================ #

  v_temp = v
  v = config['momentum'] * v_temp - config['learning_rate'] * dw
  next_w = w + v + (config['momentum'] * (v - v_temp))

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

  config['velocity'] = v

  return next_w, config

def rmsprop(w, dw, config=None):
  """
  Uses the RMSProp update rule, which uses a moving average of squared gradient
  values to set adaptive per-parameter learning rates.

  config format:
  - learning_rate: Scalar learning rate.
  - decay_rate: Scalar between 0 and 1 giving the decay rate for the squared
    gradient cache.
  - epsilon: Small scalar used for smoothing to avoid dividing by zero.
  - beta: Moving average of second moments of gradients.
```

```python
    """
    if config is None: config = {}
    config.setdefault('learning_rate', 1e-2)
    config.setdefault('decay_rate', 0.99)
    config.setdefault('epsilon', 1e-8)
    config.setdefault('a', np.zeros_like(w))

    next_w = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement RMSProp.  Store the next value of w as next_w.  You need
    #   to also store in config['a'] the moving average of the second
    #   moment gradients, so they can be used for future gradients. Concretely,
    #   config['a'] corresponds to "a" in the lecture notes.
    # ================================================================ #

    config['a'] = config['decay_rate'] * config['a'] + (1 - config['decay_rate']) * dw * dw
    next_w = w - (config['learning_rate'] / (np.sqrt(config['a']) + config['epsilon'])) * dw

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

    return next_w, config


def adam(w, dw, config=None):
    """
    Uses the Adam update rule, which incorporates moving averages of both the
    gradient and its square and a bias correction term.

    config format:
    - learning_rate: Scalar learning rate.
    - beta1: Decay rate for moving average of first moment of gradient.
    - beta2: Decay rate for moving average of second moment of gradient.
    - epsilon: Small scalar used for smoothing to avoid dividing by zero.
    - m: Moving average of gradient.
    - v: Moving average of squared gradient.
    - t: Iteration number.
    """
    if config is None: config = {}
    config.setdefault('learning_rate', 1e-3)
    config.setdefault('beta1', 0.9)
    config.setdefault('beta2', 0.999)
    config.setdefault('epsilon', 1e-8)
    config.setdefault('v', np.zeros_like(w))
    config.setdefault('a', np.zeros_like(w))
    config.setdefault('t', 0)

    next_w = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement Adam.  Store the next value of w as next_w.  You need
    #   to also store in config['a'] the moving average of the second
    #   moment gradients, and in config['v'] the moving average of the
    #   first moments.  Finally, store in config['t'] the increasing time.
    # ================================================================ #

    config['t'] = config['t'] + 1

    config['v'] = (config['beta1'] * config['v']) + (1 - config['beta1']) * dw
    config['a'] = (config['beta2'] * config['a']) + (1 - config['beta2']) * dw * dw

    # Bias correction
    v_corrected = (1 / (1 - (config['beta1'] ** config['t']))) * config['v']
```

```
    a_corrected = (1 / (1 - (config['beta2'] ** config['t']))) * config['a']

    next_w = w - (config['learning_rate'] / (np.sqrt(a_corrected) + config['epsilon'])) *
v_corrected

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

    return next_w, config
```

Batch-Normalization

February 19, 2024

# 1 Batch Normalization

In this notebook, you will implement the batch normalization layers of a neural network to increase its performance. Please review the details of batch normalization from the lecture notes.

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, and their layer structure. 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.fc_net import *
from nndl.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
```

[ ]:
```
# 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 Batchnorm forward pass

Implement the training time batchnorm forward pass, `batchnorm_forward`, in `nndl/layers.py`. After that, test your implementation by running the following cell.

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

# Simulate the forward pass for a two-layer network
N, D1, D2, D3 = 200, 50, 60, 3
X = np.random.randn(N, D1)
W1 = np.random.randn(D1, D2)
W2 = np.random.randn(D2, D3)
a = np.maximum(0, X.dot(W1)).dot(W2)

print('Before batch normalization:')
print('  means: ', a.mean(axis=0))
print('  stds: ', a.std(axis=0))

# Means should be close to zero and stds close to one
print('After batch normalization (gamma=1, beta=0)')
a_norm, _ = batchnorm_forward(a, np.ones(D3), np.zeros(D3), {'mode': 'train'})
print('  mean: ', a_norm.mean(axis=0))
print('  std: ', a_norm.std(axis=0))

# Now means should be close to beta and stds close to gamma
gamma = np.asarray([1.0, 2.0, 3.0])
beta = np.asarray([11.0, 12.0, 13.0])
a_norm, _ = batchnorm_forward(a, gamma, beta, {'mode': 'train'})
print('After batch normalization (nontrivial gamma, beta)')
print('  means: ', a_norm.mean(axis=0))
print('  stds: ', a_norm.std(axis=0))
```

```
Before batch normalization:
  means:  [15.8403627  -3.08777759 15.31647698]
  stds:  [31.08894588 31.20062471 27.83391165]
After batch normalization (gamma=1, beta=0)
  mean:  [-1.01030295e-16  1.42941214e-17  1.22124533e-17]
  std:  [0.99999999 0.99999999 0.99999999]
After batch normalization (nontrivial gamma, beta)
  means:  [11. 12. 13.]
  stds:  [0.99999999 1.99999999 2.99999998]
```

Implement the testing time batchnorm forward pass, `batchnorm_forward`, in `nndl/layers.py`. After that, test your implementation by running the following cell.

[ ]:
```
# Check the test-time forward pass by running the training-time
# forward pass many times to warm up the running averages, and then
# checking the means and variances of activations after a test-time
# forward pass.

N, D1, D2, D3 = 200, 50, 60, 3
W1 = np.random.randn(D1, D2)
W2 = np.random.randn(D2, D3)

bn_param = {'mode': 'train'}
gamma = np.ones(D3)
beta = np.zeros(D3)
for t in np.arange(50):
  X = np.random.randn(N, D1)
  a = np.maximum(0, X.dot(W1)).dot(W2)
  batchnorm_forward(a, gamma, beta, bn_param)
bn_param['mode'] = 'test'
X = np.random.randn(N, D1)
a = np.maximum(0, X.dot(W1)).dot(W2)
a_norm, _ = batchnorm_forward(a, gamma, beta, bn_param)

# Means should be close to zero and stds close to one, but will be
# noisier than training-time forward passes.
print('After batch normalization (test-time):')
print('  means: ', a_norm.mean(axis=0))
print('  stds: ', a_norm.std(axis=0))
```

```
After batch normalization (test-time):
  means:  [ 0.19412904  0.09310373 -0.18460152]
  stds:  [1.09067975 0.97400217 0.99228077]
```

### 1.2 Batchnorm backward pass

Implement the backward pass for the batchnorm layer, `batchnorm_backward` in `nndl/layers.py`. Check your implementation by running the following cell.

[ ]:
```
# Gradient check batchnorm backward pass

N, D = 4, 5
x = 5 * np.random.randn(N, D) + 12
gamma = np.random.randn(D)
beta = np.random.randn(D)
dout = np.random.randn(N, D)

bn_param = {'mode': 'train'}
fx = lambda x: batchnorm_forward(x, gamma, beta, bn_param)[0]
fg = lambda a: batchnorm_forward(x, gamma, beta, bn_param)[0]
fb = lambda b: 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 = batchnorm_forward(x, gamma, beta, bn_param)
dx, dgamma, dbeta = 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:  3.486012774507615e-10
dgamma error:  1.788859956260324e-11
dbeta error:  3.2756409038403876e-12
```

### 1.3 Implement a fully connected neural network with batchnorm layers

Modify the `FullyConnectedNet()` class in `nndl/fc_net.py` to incorporate batchnorm layers. You will need to modify the class in the following areas:

(1) The gammas and betas need to be initialized to 1's and 0's respectively in `__init__`.

(2) The `batchnorm_forward` layer needs to be inserted between each affine and relu layer (except in the output layer) in a forward pass computation in `loss`. You may find it helpful to write an `affine_batchnorm_relu()` layer in `nndl/layer_utils.py` although this is not necessary.

(3) The `batchnorm_backward` layer has to be appropriately inserted when calculating gradients.

After you have done the appropriate modifications, check your implementation by running the following cell.

Note, while the relative error for W3 should be small, as we backprop gradients more, you may find the relative error increases. Our relative error for W1 is on the order of 1e-4.

[ ]:
```
N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))
```

```python
for reg in [0, 3.14]:
  print('Running check with reg = ', reg)
  model = FullyConnectedNet([H1, H2], input_dim=D, num_classes=C,
                            reg=reg, weight_scale=5e-2, dtype=np.float64,
                            use_batchnorm=True)

  loss, grads = model.loss(X, y)
  print('Initial loss: ', loss)

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False,␣
 ↪h=1e-5)
    print('{} relative error: {}'.format(name, rel_error(grad_num,␣
 ↪grads[name])))
  if reg == 0: print('\n')
```

```
Running check with reg =  0
Initial loss:  2.2864890433891327
W1 relative error: 3.0102787062346836e-05
W2 relative error: 4.099051170280919e-06
W3 relative error: 3.706914183196035e-10
b1 relative error: 1.1102230246251565e-08
b2 relative error: 1.1102230246251565e-08
b3 relative error: 1.6920994621533787e-10
beta1 relative error: 2.0087140177382074e-08
beta2 relative error: 6.4286327899011955e-09
gamma1 relative error: 1.9667827876993263e-08
gamma2 relative error: 1.4005529617517133e-08


Running check with reg =  3.14
Initial loss:  7.352744188585346
W1 relative error: 5.772404897672606e-05
W2 relative error: 1.9687788541913548e-06
W3 relative error: 1.0115393991284992e-07
b1 relative error: 3.552713678800501e-07
b2 relative error: 2.7755575615628914e-09
b3 relative error: 2.121091785454902e-10
beta1 relative error: 3.3762290417782724e-09
beta2 relative error: 1.341264252756747e-07
gamma1 relative error: 3.328877625230386e-09
gamma2 relative error: 1.1730886007062147e-07
```

### 1.4 Training a deep fully connected network with batch normalization.

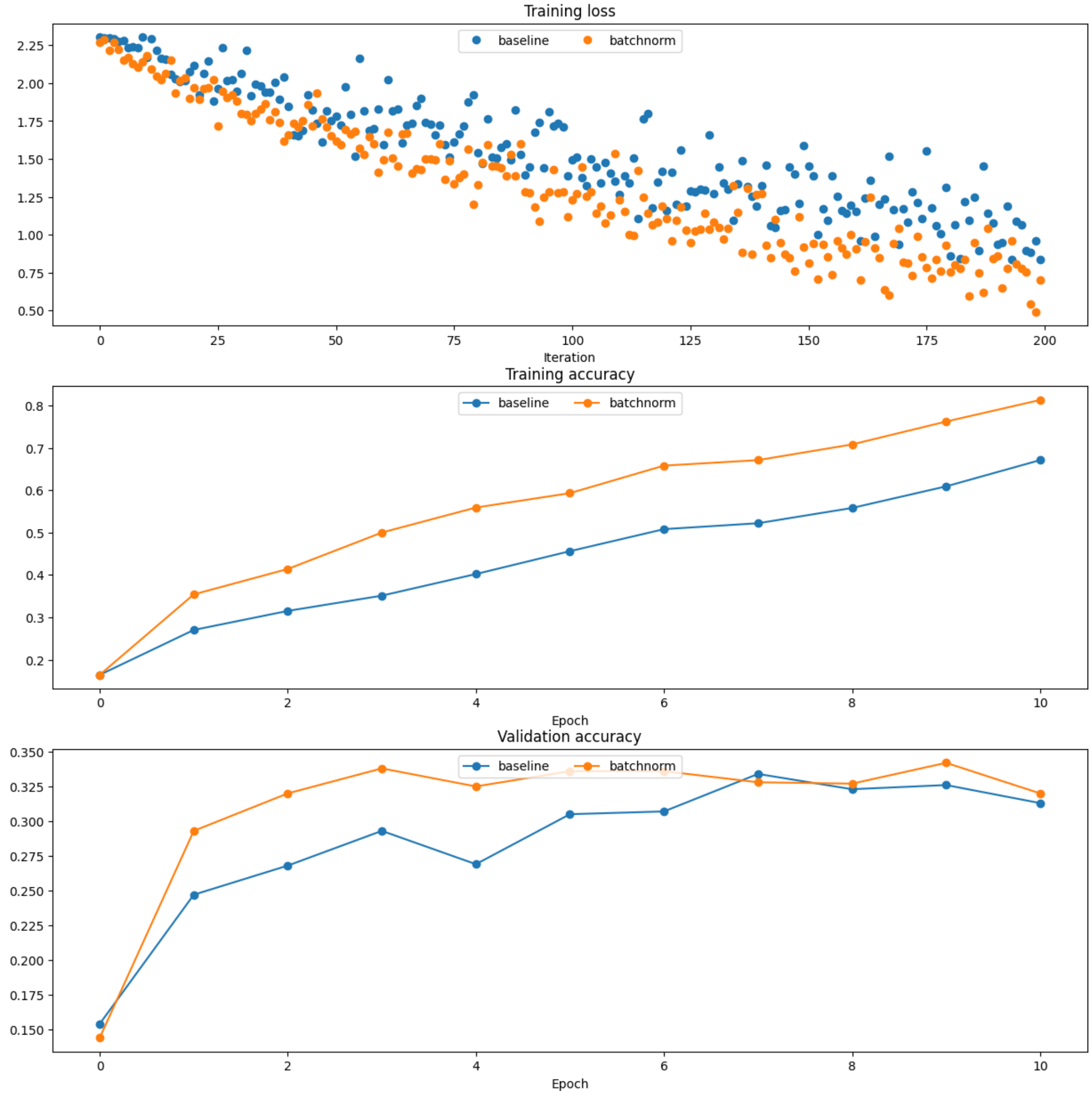
To see if batchnorm helps, let's train a deep neural network with and without batch normalization.

[ ]:
```python
# Try training a very deep net with batchnorm
hidden_dims = [100, 100, 100, 100, 100]

num_train = 1000
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'],
}

weight_scale = 2e-2
bn_model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale,␣
 ↪use_batchnorm=True)
model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale,␣
 ↪use_batchnorm=False)

bn_solver = Solver(bn_model, small_data,
                num_epochs=10, batch_size=50,
                update_rule='adam',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True, print_every=200)
bn_solver.train()

solver = Solver(model, small_data,
                num_epochs=10, batch_size=50,
                update_rule='adam',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True, print_every=200)
solver.train()
```

```
(Iteration 1 / 200) loss: 2.272360
(Epoch 0 / 10) train acc: 0.164000; val_acc: 0.144000
(Epoch 1 / 10) train acc: 0.354000; val_acc: 0.293000
(Epoch 2 / 10) train acc: 0.414000; val_acc: 0.320000
(Epoch 3 / 10) train acc: 0.500000; val_acc: 0.338000
(Epoch 4 / 10) train acc: 0.559000; val_acc: 0.325000
(Epoch 5 / 10) train acc: 0.593000; val_acc: 0.336000
(Epoch 6 / 10) train acc: 0.658000; val_acc: 0.336000
(Epoch 7 / 10) train acc: 0.671000; val_acc: 0.328000
(Epoch 8 / 10) train acc: 0.708000; val_acc: 0.327000
(Epoch 9 / 10) train acc: 0.762000; val_acc: 0.342000
(Epoch 10 / 10) train acc: 0.813000; val_acc: 0.320000
(Iteration 1 / 200) loss: 2.302262
```

```
(Epoch 0 / 10) train acc: 0.164000; val_acc: 0.154000
(Epoch 1 / 10) train acc: 0.270000; val_acc: 0.247000
(Epoch 2 / 10) train acc: 0.315000; val_acc: 0.268000
(Epoch 3 / 10) train acc: 0.351000; val_acc: 0.293000
(Epoch 4 / 10) train acc: 0.402000; val_acc: 0.269000
(Epoch 5 / 10) train acc: 0.456000; val_acc: 0.305000
(Epoch 6 / 10) train acc: 0.508000; val_acc: 0.307000
(Epoch 7 / 10) train acc: 0.522000; val_acc: 0.334000
(Epoch 8 / 10) train acc: 0.558000; val_acc: 0.323000
(Epoch 9 / 10) train acc: 0.609000; val_acc: 0.326000
(Epoch 10 / 10) train acc: 0.671000; val_acc: 0.313000
```

[ ]:
```python
plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 1)
plt.plot(solver.loss_history, 'o', label='baseline')
plt.plot(bn_solver.loss_history, 'o', label='batchnorm')

plt.subplot(3, 1, 2)
plt.plot(solver.train_acc_history, '-o', label='baseline')
plt.plot(bn_solver.train_acc_history, '-o', label='batchnorm')

plt.subplot(3, 1, 3)
plt.plot(solver.val_acc_history, '-o', label='baseline')
plt.plot(bn_solver.val_acc_history, '-o', label='batchnorm')

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```




## 1.5 Batchnorm and initialization

The following cells run an experiment where for a deep network, the initialization is varied. We do training for when batchnorm layers are and are not included.

[ ]:
```
# Try training a very deep net with batchnorm
hidden_dims = [50, 50, 50, 50, 50, 50, 50]

num_train = 1000
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'],
}

bn_solvers = {}
solvers = {}
weight_scales = np.logspace(-4, 0, num=20)
for i, weight_scale in enumerate(weight_scales):
  print('Running weight scale {} / {}'.format(i + 1, len(weight_scales)))
  bn_model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale,␣
 ↪use_batchnorm=True)
  model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale,␣
 ↪use_batchnorm=False)

  bn_solver = Solver(bn_model, small_data,
                  num_epochs=10, batch_size=50,
                  update_rule='adam',
                  optim_config={
                    'learning_rate': 1e-3,
                  },
                  verbose=False, print_every=200)
  bn_solver.train()
  bn_solvers[weight_scale] = bn_solver

  solver = Solver(model, small_data,
                  num_epochs=10, batch_size=50,
                  update_rule='adam',
                  optim_config={
                    'learning_rate': 1e-3,
                  },
                  verbose=False, print_every=200)
  solver.train()
  solvers[weight_scale] = solver
```

```
Running weight scale 1 / 20
Running weight scale 2 / 20
Running weight scale 3 / 20
Running weight scale 4 / 20
Running weight scale 5 / 20
Running weight scale 6 / 20
Running weight scale 7 / 20
Running weight scale 8 / 20
Running weight scale 9 / 20
Running weight scale 10 / 20
Running weight scale 11 / 20
Running weight scale 12 / 20
Running weight scale 13 / 20
Running weight scale 14 / 20
```

```
Running weight scale 15 / 20
Running weight scale 16 / 20
```

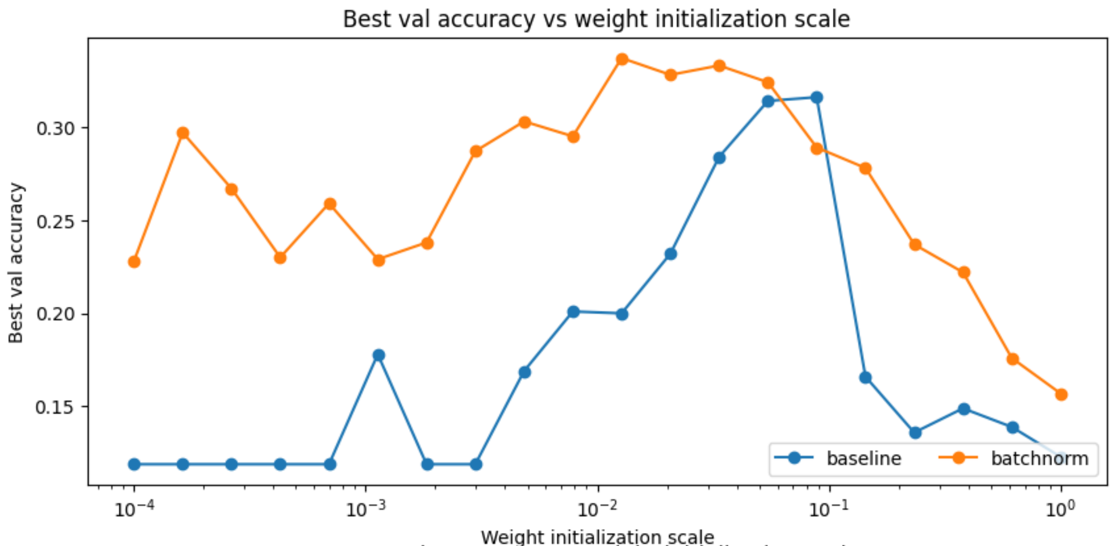
```
/Users/tilboon/Documents/UCLA/Courses/C247/HW4/HW4_code/nndl/layers.py:454:
RuntimeWarning: divide by zero encountered in log
  dx = probs.copy()
```

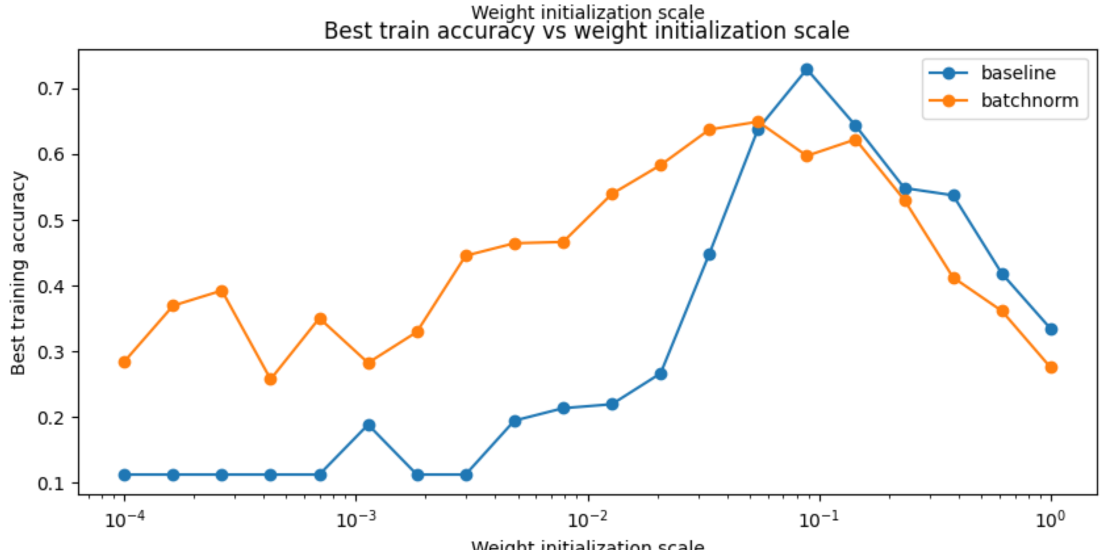
```
Running weight scale 17 / 20
Running weight scale 18 / 20
Running weight scale 19 / 20
Running weight scale 20 / 20
```

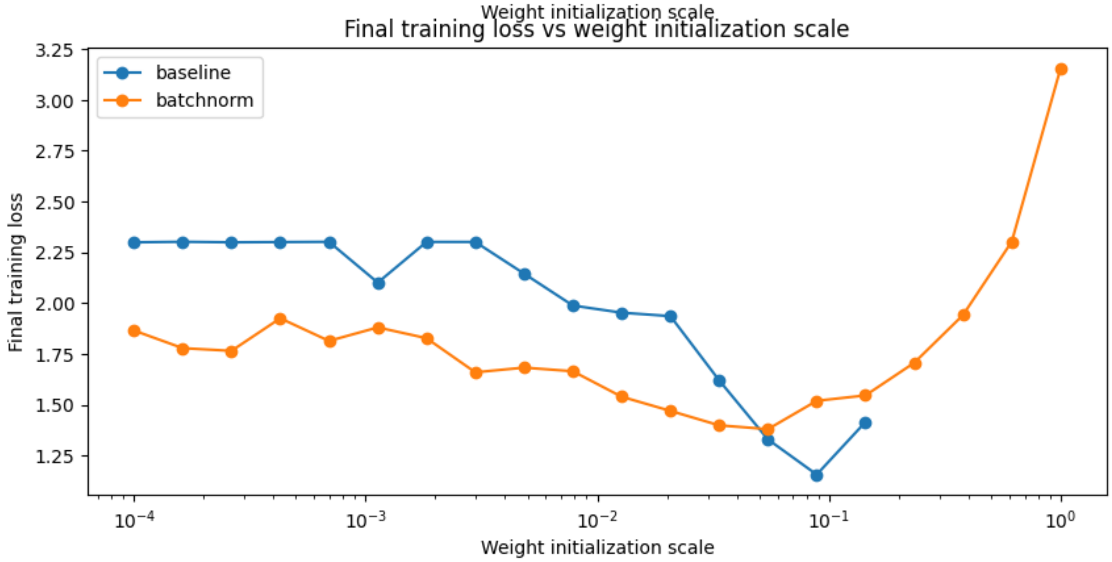
[ ]:

```
# Plot results of weight scale experiment
best_train_accs, bn_best_train_accs = [], []
best_val_accs, bn_best_val_accs = [], []
final_train_loss, bn_final_train_loss = [], []

for ws in weight_scales:
  best_train_accs.append(max(solvers[ws].train_acc_history))
  bn_best_train_accs.append(max(bn_solvers[ws].train_acc_history))

  best_val_accs.append(max(solvers[ws].val_acc_history))
  bn_best_val_accs.append(max(bn_solvers[ws].val_acc_history))

  final_train_loss.append(np.mean(solvers[ws].loss_history[-100:]))
  bn_final_train_loss.append(np.mean(bn_solvers[ws].loss_history[-100:]))

plt.subplot(3, 1, 1)
plt.title('Best val accuracy vs weight initialization scale')
plt.xlabel('Weight initialization scale')
plt.ylabel('Best val accuracy')
plt.semilogx(weight_scales, best_val_accs, '-o', label='baseline')
plt.semilogx(weight_scales, bn_best_val_accs, '-o', label='batchnorm')
plt.legend(ncol=2, loc='lower right')

plt.subplot(3, 1, 2)
plt.title('Best train accuracy vs weight initialization scale')
plt.xlabel('Weight initialization scale')
plt.ylabel('Best training accuracy')
plt.semilogx(weight_scales, best_train_accs, '-o', label='baseline')
plt.semilogx(weight_scales, bn_best_train_accs, '-o', label='batchnorm')
plt.legend()

plt.subplot(3, 1, 3)
plt.title('Final training loss vs weight initialization scale')
plt.xlabel('Weight initialization scale')
plt.ylabel('Final training loss')
plt.semilogx(weight_scales, final_train_loss, '-o', label='baseline')
```

```
plt.semilogx(weight_scales, bn_final_train_loss, '-o', label='batchnorm')
plt.legend()

plt.gcf().set_size_inches(10, 15)
plt.show()
```


Best val accuracy vs weight initialization scale
Best val accuracy
0.30
0.25
0.20
0.15
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
$10^{0}$
Weight initialization scale
baseline
batchnorm


Best train accuracy vs weight initialization scale
Best training accuracy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
$10^{0}$
Weight initialization scale
baseline
batchnorm


Final training loss vs weight initialization scale
Final training loss
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
$10^{0}$
Weight initialization scale
baseline
batchnorm

## 1.6 Question:

In the cell below, summarize the findings of this experiment, and WHY these results make sense.

## 1.7 Answer:

From the experiemtn I see two major findings. The first of which is that the batch-norm model does much better than the baseline with very small/large beginning weight initializations. The baseline has a very low accuracy (and higher loss) in these situations, the batch-norm model has a much better accuracy. Additionally, the batch-norm model is much more smooth and has far less fluctuation. This is due to the fact that batch-norm makes these measurement factors (accuracy, loss) less influenced by the weight initialization. Overall the batch-norm model has a large reduction in variance caused by initial weights, compared to the baseline model.

```python
from .layers import *

"""
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 affine_relu_forward(x, w, b):
  """
  Convenience layer that performs an affine transform followed by a ReLU

  Inputs:
  - x: Input to the affine layer
  - w, b: Weights for the affine layer

  Returns a tuple of:
  - out: Output from the ReLU
  - cache: Object to give to the backward pass
  """
  a, fc_cache = affine_forward(x, w, b)
  out, relu_cache = relu_forward(a)
  cache = (fc_cache, relu_cache)
  return out, cache


def affine_relu_backward(dout, cache):
  """
  Backward pass for the affine-relu convenience layer
  """
  fc_cache, relu_cache = cache
  da = relu_backward(dout, relu_cache)
  dx, dw, db = affine_backward(da, fc_cache)
  return dx, dw, db


def bn_affine_relu_forward(x, w, b, gamma, beta, bn_param):
  """
  Helper function for batch-norm affine relu forward
  """
  out, fc_cache = affine_forward(x, w, b)
  out, bn_cache = batchnorm_forward(out, gamma, beta, bn_param)
  out, relu_cache = relu_forward(out)
  cache = (fc_cache, bn_cache, relu_cache)

  return out, cache


def bn_affine_relu_backward(dout, cache):
  """
  Helper function for batch-norm affine relu backward
  """

  fc_cache, bn_cache, relu_cache = cache
  # da = relu_backward(dout, relu_cache)
  dout = relu_backward(dout, relu_cache)
  dout, dgamma, dbeta = batchnorm_backward(dout, bn_cache)
  dx, dw, db = affine_backward(dout, fc_cache)

  return dx, dw, db, dgamma, dbeta
```

# Dropout

February 19, 2024

## 1 Dropout

In this notebook, you will implement dropout. Then we will ask you to train a network with batchnorm and dropout, and acheive over 55% 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, and their layer structure. 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.fc_net import *
from nndl.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))))
```

[ ]:
```
# 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 Dropout forward pass

Implement the training and test time dropout forward pass, `dropout_forward`, in `nndl/layers.py`. After that, test your implementation by running the following cell.

[ ]:
```
x = np.random.randn(500, 500) + 10

for p in [0.3, 0.6, 0.75]:
  out, _ = dropout_forward(x, {'mode': 'train', 'p': p})
  out_test, _ = dropout_forward(x, {'mode': 'test', 'p': p})

  print('Running tests with p = ', p)
  print('Mean of input: ', x.mean())
  print('Mean of train-time output: ', out.mean())
  print('Mean of test-time output: ', out_test.mean())
  print('Fraction of train-time output set to zero: ', (out == 0).mean())
  print('Fraction of test-time output set to zero: ', (out_test == 0).mean())
```

```
Running tests with p =  0.3
Mean of input:  9.999939317528426
Mean of train-time output:  10.010258362864214
Mean of test-time output:  9.999939317528426
Fraction of train-time output set to zero:  0.6997
Fraction of test-time output set to zero:  0.0
Running tests with p =  0.6
Mean of input:  9.999939317528426
Mean of train-time output:  9.987650768117058
Mean of test-time output:  9.999939317528426
Fraction of train-time output set to zero:  0.400768
Fraction of test-time output set to zero:  0.0
Running tests with p =  0.75
Mean of input:  9.999939317528426
Mean of train-time output:  9.99349662435614
Mean of test-time output:  9.999939317528426
Fraction of train-time output set to zero:  0.250344
Fraction of test-time output set to zero:  0.0
```

### 1.2 Dropout backward pass

Implement the backward pass, `dropout_backward`, in `nndl/layers.py`. After that, test your gradients by running the following cell:

```
[ ]: x = np.random.randn(10, 10) + 10
dout = np.random.randn(*x.shape)

dropout_param = {'mode': 'train', 'p': 0.8, 'seed': 123}
out, cache = dropout_forward(x, dropout_param)
dx = dropout_backward(dout, cache)
dx_num = eval_numerical_gradient_array(lambda xx: dropout_forward(xx, dropout_param)[0], x, dout)

print('dx relative error: ', rel_error(dx, dx_num))
```

```
dx relative error:  5.445608548882915e-11
```

### 1.3 Implement a fully connected neural network with dropout layers

Modify the `FullyConnectedNet()` class in `nndl/fc_net.py` to incorporate dropout. A dropout layer should be incorporated after every ReLU layer. Concretely, there shouldn't be a dropout at the output layer since there is no ReLU at the output layer. You will need to modify the class in the following areas:

(1) In the forward pass, you will need to incorporate a dropout layer after every relu layer.

(2) In the backward pass, you will need to incorporate a dropout backward pass layer.

Check your implementation by running the following code. Our W1 gradient relative error is on the order of 1e-6 (the largest of all the relative errors).

```
[ ]: N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))

for dropout in [0, 0.25, 0.5]:
  print('Running check with dropout = ', dropout)
  model = FullyConnectedNet([H1, H2], input_dim=D, num_classes=C,
                            weight_scale=5e-2, dtype=np.float64,
                            dropout=dropout, seed=123)

  loss, grads = model.loss(X, y)
  print('Initial loss: ', loss)

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False, h=1e-5)
```

```
    print('{} relative error: {}'.format(name, rel_error(grad_num,␣
↪grads[name])))
  print('\n')
```

```
Running check with dropout =  0
Initial loss:  2.303043161170242
W1 relative error: 4.795196815215288e-07
W2 relative error: 1.9717710574314515e-07
W3 relative error: 1.5587099483501822e-07
b1 relative error: 2.033615448560775e-08
b2 relative error: 1.686315567518667e-09
b3 relative error: 1.1144421861081857e-10


Running check with dropout =  0.25
Initial loss:  2.302354247831908
W1 relative error: 1.0017417771677944e-07
W2 relative error: 2.2591349991943916e-09
W3 relative error: 2.5553916704768502e-05
b1 relative error: 9.368619593647787e-10
b2 relative error: 0.2134290559158092
b3 relative error: 1.2466815697171467e-10


Running check with dropout =  0.5
Initial loss:  2.304242617164796
W1 relative error: 1.207883539101674e-07
W2 relative error: 2.4541280438139708e-08
W3 relative error: 8.057428154752603e-07
b1 relative error: 2.276255983500775e-08
b2 relative error: 6.836023983491334e-10
b3 relative error: 1.2833211144563337e-10
```

### 1.4 Dropout as a regularizer

In class, we claimed that dropout acts as a regularizer by effectively bagging. To check this, we will train two small networks, one with dropout and one without dropout.

[ ]:
```
# Train two identical nets, one with dropout and one without

num_train = 500
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'],
```

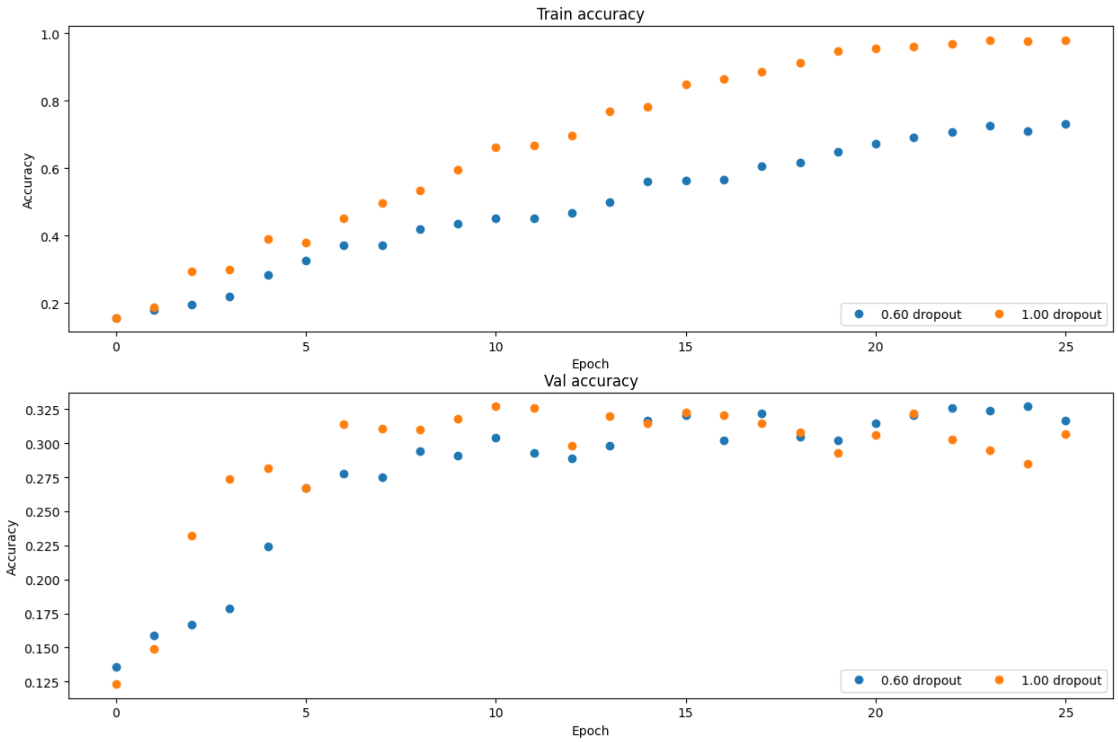
```python
}

solvers = {}
dropout_choices = [0.6, 1.0]
for dropout in dropout_choices:
  model = FullyConnectedNet([100, 100, 100], dropout=dropout)

  solver = Solver(model, small_data,
                  num_epochs=25, batch_size=100,
                  update_rule='adam',
                  optim_config={
                    'learning_rate': 5e-4,
                  },
                  verbose=True, print_every=100)
  solver.train()
  solvers[dropout] = solver
```

```
(Iteration 1 / 125) loss: 2.302846
(Epoch 0 / 25) train acc: 0.156000; val_acc: 0.136000
(Epoch 1 / 25) train acc: 0.178000; val_acc: 0.159000
(Epoch 2 / 25) train acc: 0.194000; val_acc: 0.167000
(Epoch 3 / 25) train acc: 0.220000; val_acc: 0.179000
(Epoch 4 / 25) train acc: 0.282000; val_acc: 0.224000
(Epoch 5 / 25) train acc: 0.326000; val_acc: 0.267000
(Epoch 6 / 25) train acc: 0.372000; val_acc: 0.278000
(Epoch 7 / 25) train acc: 0.370000; val_acc: 0.275000
(Epoch 8 / 25) train acc: 0.420000; val_acc: 0.294000
(Epoch 9 / 25) train acc: 0.434000; val_acc: 0.291000
(Epoch 10 / 25) train acc: 0.452000; val_acc: 0.304000
(Epoch 11 / 25) train acc: 0.450000; val_acc: 0.293000
(Epoch 12 / 25) train acc: 0.466000; val_acc: 0.289000
(Epoch 13 / 25) train acc: 0.500000; val_acc: 0.298000
(Epoch 14 / 25) train acc: 0.560000; val_acc: 0.317000
(Epoch 15 / 25) train acc: 0.562000; val_acc: 0.321000
(Epoch 16 / 25) train acc: 0.566000; val_acc: 0.302000
(Epoch 17 / 25) train acc: 0.606000; val_acc: 0.322000
(Epoch 18 / 25) train acc: 0.616000; val_acc: 0.305000
(Epoch 19 / 25) train acc: 0.648000; val_acc: 0.302000
(Epoch 20 / 25) train acc: 0.672000; val_acc: 0.315000
(Iteration 101 / 125) loss: 1.179990
(Epoch 21 / 25) train acc: 0.692000; val_acc: 0.321000
(Epoch 22 / 25) train acc: 0.708000; val_acc: 0.326000
(Epoch 23 / 25) train acc: 0.726000; val_acc: 0.324000
(Epoch 24 / 25) train acc: 0.710000; val_acc: 0.327000
(Epoch 25 / 25) train acc: 0.732000; val_acc: 0.317000
(Iteration 1 / 125) loss: 2.300483
(Epoch 0 / 25) train acc: 0.156000; val_acc: 0.123000
(Epoch 1 / 25) train acc: 0.186000; val_acc: 0.149000
```

```
(Epoch 2 / 25) train acc: 0.294000; val_acc: 0.232000
(Epoch 3 / 25) train acc: 0.300000; val_acc: 0.274000
(Epoch 4 / 25) train acc: 0.390000; val_acc: 0.282000
(Epoch 5 / 25) train acc: 0.380000; val_acc: 0.267000
(Epoch 6 / 25) train acc: 0.452000; val_acc: 0.314000
(Epoch 7 / 25) train acc: 0.496000; val_acc: 0.311000
(Epoch 8 / 25) train acc: 0.534000; val_acc: 0.310000
(Epoch 9 / 25) train acc: 0.596000; val_acc: 0.318000
(Epoch 10 / 25) train acc: 0.662000; val_acc: 0.327000
(Epoch 11 / 25) train acc: 0.666000; val_acc: 0.326000
(Epoch 12 / 25) train acc: 0.698000; val_acc: 0.298000
(Epoch 13 / 25) train acc: 0.768000; val_acc: 0.320000
(Epoch 14 / 25) train acc: 0.782000; val_acc: 0.315000
(Epoch 15 / 25) train acc: 0.848000; val_acc: 0.323000
(Epoch 16 / 25) train acc: 0.866000; val_acc: 0.321000
(Epoch 17 / 25) train acc: 0.886000; val_acc: 0.315000
(Epoch 18 / 25) train acc: 0.912000; val_acc: 0.308000
(Epoch 19 / 25) train acc: 0.948000; val_acc: 0.293000
(Epoch 20 / 25) train acc: 0.956000; val_acc: 0.306000
(Iteration 101 / 125) loss: 0.239402
(Epoch 21 / 25) train acc: 0.962000; val_acc: 0.322000
(Epoch 22 / 25) train acc: 0.968000; val_acc: 0.303000
(Epoch 23 / 25) train acc: 0.980000; val_acc: 0.295000
(Epoch 24 / 25) train acc: 0.976000; val_acc: 0.285000
(Epoch 25 / 25) train acc: 0.980000; val_acc: 0.307000
```

[ ]:
```python
# Plot train and validation accuracies of the two models

train_accs = []
val_accs = []
for dropout in dropout_choices:
  solver = solvers[dropout]
  train_accs.append(solver.train_acc_history[-1])
  val_accs.append(solver.val_acc_history[-1])

plt.subplot(3, 1, 1)
for dropout in dropout_choices:
  plt.plot(solvers[dropout].train_acc_history, 'o', label='%.2f dropout' %␣
↪dropout)
plt.title('Train accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend(ncol=2, loc='lower right')

plt.subplot(3, 1, 2)
for dropout in dropout_choices:
```

```
  plt.plot(solvers[dropout].val_acc_history, 'o', label='%.2f dropout' %␣
↪dropout)
plt.title('Val accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend(ncol=2, loc='lower right')

plt.gcf().set_size_inches(15, 15)
plt.show()
```




## 1.5 Question

Based off the results of this experiment, is dropout performing regularization? Explain your answer.

## 1.6 Answer:

Based off the results of this experiment, droupout is performing regularization. I tested dropout choices [0.0, 0.6] and [0.6, 1.0] and I have found that even when droput performs worst on the training set, it performs equally, if not better, on the validation set. Dropout specifically targets overfitting of the trainig datset by "dropping units" during training. This was reflected in the results as when dropout is not included, we can see in validation that overfitting occurs.

## 1.7 Final part of the assignment

Get over 55% validation accuracy on CIFAR-10 by using the layers you have implemented. You will be graded according to the following equation:

min(floor((X - 32%)) / 23%, 1) where if you get 55% or higher validation accuracy, you get full points.

[ ]:
```
# ================================================================ #
# YOUR CODE HERE:
#   Implement a FC-net that achieves at least 55% validation accuracy
#   on CIFAR-10.
# ================================================================ #

# Model parameter configuration
hidden_dims = [100, 100, 100, 100]
dropout = 0.95
use_batchnorm = True
weight_scale = 0.01

# Solver parameter configuration
num_epochs = 50
batch_size = 100
update_rule = 'adam'
learning_rate = 4e-4
lr_decay = 0.95
verbose = True
print_every = 100

# Model instantiation
model = FullyConnectedNet(
  hidden_dims=hidden_dims,
  dropout=dropout,
  use_batchnorm=use_batchnorm,
  weight_scale=weight_scale
)

# Solver instantiation
solver = Solver(
  model=model,
  data=data,
  num_epochs=num_epochs,
  batch_size=batch_size,
  update_rule=update_rule,
  optim_config={
    'learning_rate': learning_rate
  },
  lr_decay=lr_decay,
```

```python
  verbose=verbose,
  print_every=print_every
)

solver.train()

y_test_pred = np.argmax(model.loss(data['X_test']), axis=1)
y_val_pred = np.argmax(model.loss(data['X_val']), axis=1)
print('Validation set accuracy: {}'.format(np.mean(y_val_pred ==␣
 ↪data['y_val'])))
print('Test set accuracy: {}'.format(np.mean(y_test_pred == data['y_test'])))

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

```
(Iteration 1 / 24500) loss: 2.309899
(Epoch 0 / 50) train acc: 0.131000; val_acc: 0.132000
(Iteration 101 / 24500) loss: 1.866618
(Iteration 201 / 24500) loss: 1.718992
(Iteration 301 / 24500) loss: 1.596938
(Iteration 401 / 24500) loss: 1.495558
(Epoch 1 / 50) train acc: 0.469000; val_acc: 0.471000
(Iteration 501 / 24500) loss: 1.675077
(Iteration 601 / 24500) loss: 1.558240
(Iteration 701 / 24500) loss: 1.668873
(Iteration 801 / 24500) loss: 1.451140
(Iteration 901 / 24500) loss: 1.332606
(Epoch 2 / 50) train acc: 0.480000; val_acc: 0.508000
(Iteration 1001 / 24500) loss: 1.516851
(Iteration 1101 / 24500) loss: 1.654661
(Iteration 1201 / 24500) loss: 1.410304
(Iteration 1301 / 24500) loss: 1.498030
(Iteration 1401 / 24500) loss: 1.455426
(Epoch 3 / 50) train acc: 0.555000; val_acc: 0.512000
(Iteration 1501 / 24500) loss: 1.662855
(Iteration 1601 / 24500) loss: 1.352332
(Iteration 1701 / 24500) loss: 1.535238
(Iteration 1801 / 24500) loss: 1.396415
(Iteration 1901 / 24500) loss: 1.302943
(Epoch 4 / 50) train acc: 0.543000; val_acc: 0.516000
(Iteration 2001 / 24500) loss: 1.077760
(Iteration 2101 / 24500) loss: 1.264232
(Iteration 2201 / 24500) loss: 1.182357
(Iteration 2301 / 24500) loss: 1.239161
(Iteration 2401 / 24500) loss: 1.239764
(Epoch 5 / 50) train acc: 0.567000; val_acc: 0.522000
(Iteration 2501 / 24500) loss: 1.257635
```

```
(Iteration 2601 / 24500) loss: 1.401854
(Iteration 2701 / 24500) loss: 1.227867
(Iteration 2801 / 24500) loss: 1.074553
(Iteration 2901 / 24500) loss: 1.182340
(Epoch 6 / 50) train acc: 0.566000; val_acc: 0.520000
(Iteration 3001 / 24500) loss: 1.217173
(Iteration 3101 / 24500) loss: 1.139077
(Iteration 3201 / 24500) loss: 0.994905
(Iteration 3301 / 24500) loss: 1.293239
(Iteration 3401 / 24500) loss: 1.210162
(Epoch 7 / 50) train acc: 0.622000; val_acc: 0.547000
(Iteration 3501 / 24500) loss: 1.161972
(Iteration 3601 / 24500) loss: 1.115069
(Iteration 3701 / 24500) loss: 1.295875
(Iteration 3801 / 24500) loss: 1.192754
(Iteration 3901 / 24500) loss: 1.215506
(Epoch 8 / 50) train acc: 0.609000; val_acc: 0.541000
(Iteration 4001 / 24500) loss: 1.055536
(Iteration 4101 / 24500) loss: 1.196554
(Iteration 4201 / 24500) loss: 1.145615
(Iteration 4301 / 24500) loss: 1.187404
(Iteration 4401 / 24500) loss: 1.127986
(Epoch 9 / 50) train acc: 0.617000; val_acc: 0.542000
(Iteration 4501 / 24500) loss: 1.090329
(Iteration 4601 / 24500) loss: 1.026115
(Iteration 4701 / 24500) loss: 1.213346
(Iteration 4801 / 24500) loss: 1.176441
(Epoch 10 / 50) train acc: 0.649000; val_acc: 0.526000
(Iteration 4901 / 24500) loss: 1.116453
(Iteration 5001 / 24500) loss: 1.413444
(Iteration 5101 / 24500) loss: 1.082372
(Iteration 5201 / 24500) loss: 1.202887
(Iteration 5301 / 24500) loss: 1.007759
(Epoch 11 / 50) train acc: 0.632000; val_acc: 0.532000
(Iteration 5401 / 24500) loss: 1.265928
(Iteration 5501 / 24500) loss: 1.194023
(Iteration 5601 / 24500) loss: 0.998210
(Iteration 5701 / 24500) loss: 1.038412
(Iteration 5801 / 24500) loss: 1.277954
(Epoch 12 / 50) train acc: 0.647000; val_acc: 0.545000
(Iteration 5901 / 24500) loss: 0.964349
(Iteration 6001 / 24500) loss: 1.078777
(Iteration 6101 / 24500) loss: 0.898104
(Iteration 6201 / 24500) loss: 1.106749
(Iteration 6301 / 24500) loss: 0.916196
(Epoch 13 / 50) train acc: 0.628000; val_acc: 0.549000
(Iteration 6401 / 24500) loss: 1.140986
(Iteration 6501 / 24500) loss: 0.877311
```

```
(Iteration 6601 / 24500) loss: 1.097259
(Iteration 6701 / 24500) loss: 0.982414
(Iteration 6801 / 24500) loss: 0.865046
(Epoch 14 / 50) train acc: 0.683000; val_acc: 0.545000
(Iteration 6901 / 24500) loss: 0.930988
(Iteration 7001 / 24500) loss: 1.186335
(Iteration 7101 / 24500) loss: 1.137504
(Iteration 7201 / 24500) loss: 1.047523
(Iteration 7301 / 24500) loss: 1.362683
(Epoch 15 / 50) train acc: 0.671000; val_acc: 0.544000
(Iteration 7401 / 24500) loss: 1.019300
(Iteration 7501 / 24500) loss: 0.924499
(Iteration 7601 / 24500) loss: 1.064929
(Iteration 7701 / 24500) loss: 0.955549
(Iteration 7801 / 24500) loss: 1.062574
(Epoch 16 / 50) train acc: 0.693000; val_acc: 0.544000
(Iteration 7901 / 24500) loss: 1.109003
(Iteration 8001 / 24500) loss: 0.956318
(Iteration 8101 / 24500) loss: 0.885780
(Iteration 8201 / 24500) loss: 1.022016
(Iteration 8301 / 24500) loss: 0.843696
(Epoch 17 / 50) train acc: 0.697000; val_acc: 0.544000
(Iteration 8401 / 24500) loss: 1.129848
(Iteration 8501 / 24500) loss: 1.010513
(Iteration 8601 / 24500) loss: 1.029730
(Iteration 8701 / 24500) loss: 1.015755
(Iteration 8801 / 24500) loss: 0.956425
(Epoch 18 / 50) train acc: 0.715000; val_acc: 0.540000
(Iteration 8901 / 24500) loss: 1.048353
(Iteration 9001 / 24500) loss: 0.974311
(Iteration 9101 / 24500) loss: 1.006743
(Iteration 9201 / 24500) loss: 1.237507
(Iteration 9301 / 24500) loss: 0.978970
(Epoch 19 / 50) train acc: 0.707000; val_acc: 0.551000
(Iteration 9401 / 24500) loss: 0.967101
(Iteration 9501 / 24500) loss: 1.078612
(Iteration 9601 / 24500) loss: 0.845386
(Iteration 9701 / 24500) loss: 1.179538
(Epoch 20 / 50) train acc: 0.714000; val_acc: 0.547000
(Iteration 9801 / 24500) loss: 0.912087
(Iteration 9901 / 24500) loss: 0.950588
(Iteration 10001 / 24500) loss: 0.809592
(Iteration 10101 / 24500) loss: 0.799876
(Iteration 10201 / 24500) loss: 0.891645
(Epoch 21 / 50) train acc: 0.704000; val_acc: 0.542000
(Iteration 10301 / 24500) loss: 1.144767
(Iteration 10401 / 24500) loss: 1.036633
(Iteration 10501 / 24500) loss: 1.134391
```

```
(Iteration 10601 / 24500) loss: 1.016582
(Iteration 10701 / 24500) loss: 0.838502
(Epoch 22 / 50) train acc: 0.713000; val_acc: 0.547000
(Iteration 10801 / 24500) loss: 0.940446
(Iteration 10901 / 24500) loss: 0.957821
(Iteration 11001 / 24500) loss: 0.863714
(Iteration 11101 / 24500) loss: 0.812743
(Iteration 11201 / 24500) loss: 0.954878
(Epoch 23 / 50) train acc: 0.697000; val_acc: 0.534000
(Iteration 11301 / 24500) loss: 1.065244
(Iteration 11401 / 24500) loss: 1.063900
(Iteration 11501 / 24500) loss: 0.879651
(Iteration 11601 / 24500) loss: 0.838008
(Iteration 11701 / 24500) loss: 0.915915
(Epoch 24 / 50) train acc: 0.730000; val_acc: 0.556000
(Iteration 11801 / 24500) loss: 0.633101
(Iteration 11901 / 24500) loss: 1.135219
(Iteration 12001 / 24500) loss: 0.895735
(Iteration 12101 / 24500) loss: 0.901940
(Iteration 12201 / 24500) loss: 0.996008
(Epoch 25 / 50) train acc: 0.720000; val_acc: 0.548000
(Iteration 12301 / 24500) loss: 0.982373
(Iteration 12401 / 24500) loss: 1.048043
(Iteration 12501 / 24500) loss: 0.784439
(Iteration 12601 / 24500) loss: 0.971330
(Iteration 12701 / 24500) loss: 0.827522
(Epoch 26 / 50) train acc: 0.689000; val_acc: 0.543000
(Iteration 12801 / 24500) loss: 1.007480
(Iteration 12901 / 24500) loss: 0.949627
(Iteration 13001 / 24500) loss: 1.012613
(Iteration 13101 / 24500) loss: 0.893711
(Iteration 13201 / 24500) loss: 0.651754
(Epoch 27 / 50) train acc: 0.762000; val_acc: 0.557000
(Iteration 13301 / 24500) loss: 0.783713
(Iteration 13401 / 24500) loss: 0.819004
(Iteration 13501 / 24500) loss: 0.889076
(Iteration 13601 / 24500) loss: 0.898212
(Iteration 13701 / 24500) loss: 0.754875
(Epoch 28 / 50) train acc: 0.723000; val_acc: 0.559000
(Iteration 13801 / 24500) loss: 0.711990
(Iteration 13901 / 24500) loss: 0.841989
(Iteration 14001 / 24500) loss: 0.747193
(Iteration 14101 / 24500) loss: 0.904623
(Iteration 14201 / 24500) loss: 0.868194
(Epoch 29 / 50) train acc: 0.745000; val_acc: 0.538000
(Iteration 14301 / 24500) loss: 0.810182
(Iteration 14401 / 24500) loss: 1.041365
(Iteration 14501 / 24500) loss: 0.592201
```

```
(Iteration 14601 / 24500) loss: 0.956923
(Epoch 30 / 50) train acc: 0.757000; val_acc: 0.541000
(Iteration 14701 / 24500) loss: 0.813912
(Iteration 14801 / 24500) loss: 0.700483
(Iteration 14901 / 24500) loss: 0.928101
(Iteration 15001 / 24500) loss: 0.837813
(Iteration 15101 / 24500) loss: 0.973395
(Epoch 31 / 50) train acc: 0.741000; val_acc: 0.551000
(Iteration 15201 / 24500) loss: 0.945654
(Iteration 15301 / 24500) loss: 0.876836
(Iteration 15401 / 24500) loss: 0.781459
(Iteration 15501 / 24500) loss: 0.761614
(Iteration 15601 / 24500) loss: 0.814122
(Epoch 32 / 50) train acc: 0.782000; val_acc: 0.538000
(Iteration 15701 / 24500) loss: 0.748537
(Iteration 15801 / 24500) loss: 0.908091
(Iteration 15901 / 24500) loss: 1.117483
(Iteration 16001 / 24500) loss: 0.934632
(Iteration 16101 / 24500) loss: 0.983172
(Epoch 33 / 50) train acc: 0.755000; val_acc: 0.539000
(Iteration 16201 / 24500) loss: 0.882517
(Iteration 16301 / 24500) loss: 0.759011
(Iteration 16401 / 24500) loss: 0.737251
(Iteration 16501 / 24500) loss: 0.667217
(Iteration 16601 / 24500) loss: 0.702816
(Epoch 34 / 50) train acc: 0.743000; val_acc: 0.545000
(Iteration 16701 / 24500) loss: 0.951171
(Iteration 16801 / 24500) loss: 0.766670
(Iteration 16901 / 24500) loss: 0.929767
(Iteration 17001 / 24500) loss: 0.861767
(Iteration 17101 / 24500) loss: 0.831672
(Epoch 35 / 50) train acc: 0.735000; val_acc: 0.549000
(Iteration 17201 / 24500) loss: 0.805188
(Iteration 17301 / 24500) loss: 0.754041
(Iteration 17401 / 24500) loss: 0.814125
(Iteration 17501 / 24500) loss: 0.858323
(Iteration 17601 / 24500) loss: 0.711753
(Epoch 36 / 50) train acc: 0.751000; val_acc: 0.541000
(Iteration 17701 / 24500) loss: 0.981485
(Iteration 17801 / 24500) loss: 0.887595
(Iteration 17901 / 24500) loss: 0.865561
(Iteration 18001 / 24500) loss: 1.052693
(Iteration 18101 / 24500) loss: 0.821450
(Epoch 37 / 50) train acc: 0.756000; val_acc: 0.542000
(Iteration 18201 / 24500) loss: 0.656655
(Iteration 18301 / 24500) loss: 0.654024
(Iteration 18401 / 24500) loss: 0.810589
(Iteration 18501 / 24500) loss: 0.721276
```

```
(Iteration 18601 / 24500) loss: 0.855766
(Epoch 38 / 50) train acc: 0.767000; val_acc: 0.539000
(Iteration 18701 / 24500) loss: 0.988117
(Iteration 18801 / 24500) loss: 0.848304
(Iteration 18901 / 24500) loss: 0.943380
(Iteration 19001 / 24500) loss: 0.748473
(Iteration 19101 / 24500) loss: 0.748982
(Epoch 39 / 50) train acc: 0.772000; val_acc: 0.537000
(Iteration 19201 / 24500) loss: 0.711835
(Iteration 19301 / 24500) loss: 1.041719
(Iteration 19401 / 24500) loss: 0.903313
(Iteration 19501 / 24500) loss: 0.674720
(Epoch 40 / 50) train acc: 0.770000; val_acc: 0.541000
(Iteration 19601 / 24500) loss: 0.706645
(Iteration 19701 / 24500) loss: 0.771446
(Iteration 19801 / 24500) loss: 0.842413
(Iteration 19901 / 24500) loss: 0.758576
(Iteration 20001 / 24500) loss: 0.943846
(Epoch 41 / 50) train acc: 0.764000; val_acc: 0.551000
(Iteration 20101 / 24500) loss: 0.736160
(Iteration 20201 / 24500) loss: 0.892561
(Iteration 20301 / 24500) loss: 0.566785
(Iteration 20401 / 24500) loss: 0.671496
(Iteration 20501 / 24500) loss: 0.821424
(Epoch 42 / 50) train acc: 0.781000; val_acc: 0.533000
(Iteration 20601 / 24500) loss: 0.794014
(Iteration 20701 / 24500) loss: 0.866569
(Iteration 20801 / 24500) loss: 0.780973
(Iteration 20901 / 24500) loss: 0.862082
(Iteration 21001 / 24500) loss: 0.736835
(Epoch 43 / 50) train acc: 0.796000; val_acc: 0.534000
(Iteration 21101 / 24500) loss: 0.689307
(Iteration 21201 / 24500) loss: 0.814803
(Iteration 21301 / 24500) loss: 0.705534
(Iteration 21401 / 24500) loss: 0.672100
(Iteration 21501 / 24500) loss: 0.762810
(Epoch 44 / 50) train acc: 0.800000; val_acc: 0.552000
(Iteration 21601 / 24500) loss: 0.862723
(Iteration 21701 / 24500) loss: 0.869883
(Iteration 21801 / 24500) loss: 0.818486
(Iteration 21901 / 24500) loss: 0.707606
(Iteration 22001 / 24500) loss: 0.878560
(Epoch 45 / 50) train acc: 0.773000; val_acc: 0.542000
(Iteration 22101 / 24500) loss: 0.749141
(Iteration 22201 / 24500) loss: 0.590047
(Iteration 22301 / 24500) loss: 0.728439
(Iteration 22401 / 24500) loss: 0.546963
(Iteration 22501 / 24500) loss: 0.818315
```

```
(Epoch 46 / 50) train acc: 0.773000; val_acc: 0.538000
(Iteration 22601 / 24500) loss: 0.733699
(Iteration 22701 / 24500) loss: 0.632203
(Iteration 22801 / 24500) loss: 0.925055
(Iteration 22901 / 24500) loss: 0.793780
(Iteration 23001 / 24500) loss: 0.742589
(Epoch 47 / 50) train acc: 0.792000; val_acc: 0.537000
(Iteration 23101 / 24500) loss: 0.856480
(Iteration 23201 / 24500) loss: 0.624789
(Iteration 23301 / 24500) loss: 0.841813
(Iteration 23401 / 24500) loss: 0.612569
(Iteration 23501 / 24500) loss: 0.838897
(Epoch 48 / 50) train acc: 0.809000; val_acc: 0.541000
(Iteration 23601 / 24500) loss: 0.640904
(Iteration 23701 / 24500) loss: 0.787501
(Iteration 23801 / 24500) loss: 0.693472
(Iteration 23901 / 24500) loss: 0.781707
(Iteration 24001 / 24500) loss: 0.726345
(Epoch 49 / 50) train acc: 0.773000; val_acc: 0.545000
(Iteration 24101 / 24500) loss: 0.828772
(Iteration 24201 / 24500) loss: 0.648704
(Iteration 24301 / 24500) loss: 0.635414
(Iteration 24401 / 24500) loss: 0.721156
(Epoch 50 / 50) train acc: 0.780000; val_acc: 0.536000
Validation set accuracy: 0.566
Test set accuracy: 0.55
```

```python
import numpy as np
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 affine_forward(x, w, b):
  """
  Computes the forward pass for an affine (fully-connected) layer.

  The input x has shape (N, d_1, ..., d_k) and contains a minibatch of N
  examples, where each example x[i] has shape (d_1, ..., d_k). We will
  reshape each input into a vector of dimension D = d_1 * ... * d_k, and
  then transform it to an output vector of dimension M.

  Inputs:
  - x: A numpy array containing input data, of shape (N, d_1, ..., d_k)
  - w: A numpy array of weights, of shape (D, M)
  - b: A numpy array of biases, of shape (M,)

  Returns a tuple of:
  - out: output, of shape (N, M)
  - cache: (x, w, b)
  """

  # ================================================================ #
  # YOUR CODE HERE:
  #   Calculate the output of the forward pass.  Notice the dimensions
  #   of w are D x M, which is the transpose of what we did in earlier
  #   assignments.
  # ================================================================ #

  reshaped_input = np.reshape(x, (x.shape[0], -1))
  out = np.dot(reshaped_input, w) + b

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

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


def affine_backward(dout, cache):
  """
  Computes the backward pass for an affine layer.

  Inputs:
  - dout: Upstream derivative, of shape (N, M)
  - cache: Tuple of:
    - x: A numpy array containing input data, of shape (N, d_1, ..., d_k)
    - w: A numpy array of weights, of shape (D, M)
    - b: A numpy array of biases, of shape (M,)

  Returns a tuple of:
  - dx: Gradient with respect to x, of shape (N, d1, ..., d_k)
  - dw: Gradient with respect to w, of shape (D, M)
  - db: Gradient with respect to b, of shape (M,)
  """
  x, w, b = cache
```

```
  dx, dw, db = None, None, None

  # ================================================================== #
  # YOUR CODE HERE:
  #   Calculate the gradients for the backward pass.
  # Notice:
  #   dout is N x M
  #   dx should be N x d1 x ... x dk; it relates to dout through multiplication with w, which
is D x M
  #   dw should be D x M; it relates to dout through multiplication with x, which is N x D
after reshaping
  #   db should be M; it is just the sum over dout examples
  # ================================================================== #

  dx = np.dot(dout, w.T).reshape(x.shape)
  dw = np.dot(x.reshape(x.shape[0], -1).T, dout)
  db = np.sum(dout, axis=0)

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

  return dx, dw, db

def relu_forward(x):
  """
  Computes the forward pass for a layer of rectified linear units (ReLUs).

  Input:
  - x: Inputs, of any shape

  Returns a tuple of:
  - out: Output, of the same shape as x
  - cache: x
  """
  # ================================================================== #
  # YOUR CODE HERE:
  #   Implement the ReLU forward pass.
  # ================================================================== #

  out = np.maximum(x, 0)

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

  cache = x
  return out, cache


def relu_backward(dout, cache):
  """
  Computes the backward pass for a layer of rectified linear units (ReLUs).

  Input:
  - dout: Upstream derivatives, of any shape
  - cache: Input x, of same shape as dout

  Returns:
  - dx: Gradient with respect to x
  """
  x = cache

  # ================================================================== #
  # YOUR CODE HERE:
  #   Implement the ReLU backward pass
  # ================================================================== #
```

```python
  # ReLU directs linearly to those > 0
  dx = dout * (x >= 0)

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

  return dx

def batchnorm_forward(x, gamma, beta, bn_param):
  """
  Forward pass for batch normalization.

  During training the sample mean and (uncorrected) sample variance are
  computed from minibatch statistics and used to normalize the incoming data.
  During training we also keep an exponentially decaying running mean of the mean
  and variance of each feature, and these averages are used to normalize data
  at test-time.

  At each timestep we update the running averages for mean and variance using
  an exponential decay based on the momentum parameter:

  running_mean = momentum * running_mean + (1 - momentum) * sample_mean
  running_var = momentum * running_var + (1 - momentum) * sample_var

  Note that the batch normalization paper suggests a different test-time
  behavior: they compute sample mean and variance for each feature using a
  large number of training images rather than using a running average. For
  this implementation we have chosen to use running averages instead since
  they do not require an additional estimation step; the torch7 implementation
  of batch normalization also uses running averages.

  Input:
  - x: Data of shape (N, D)
  - gamma: Scale parameter of shape (D,)
  - beta: Shift paremeter of shape (D,)
  - bn_param: Dictionary with the following keys:
    - mode: 'train' or 'test'; required
    - eps: Constant for numeric stability
    - momentum: Constant for running mean / variance.
    - 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: of shape (N, D)
  - cache: A tuple of values needed in the backward pass
  """
  mode = bn_param['mode']
  eps = bn_param.get('eps', 1e-5)
  momentum = bn_param.get('momentum', 0.9)

  N, D = x.shape
  running_mean = bn_param.get('running_mean', np.zeros(D, dtype=x.dtype))
  running_var = bn_param.get('running_var', np.zeros(D, dtype=x.dtype))

  out, cache = None, None
  if mode == 'train':

    # ================================================================ #
    # YOUR CODE HERE:
    #   A few steps here:
    #     (1) Calculate the running mean and variance of the minibatch.
    #     (2) Normalize the activations with the sample mean and variance.
    #     (3) Scale and shift the normalized activations.  Store this
    #         as the variable 'out'
    #     (4) Store any variables you may need for the backward pass in
```

```python
    #         the 'cache' variable.
    # ================================================================ #

    # compute the mean and covariance from minibatch examples x
    mu = x.mean(axis=0)
    var = x.var(axis=0)

    # keep an exponentially decaying running mean and variance
    # these averages are used to normalize data at test-time
    running_mean = momentum * running_mean + (1 - momentum) * mu
    running_var = momentum * running_var + (1 - momentum) * var

    # normalize the activations x with the current mean and variance
    # subtract the mean and divide by the variance
    xc = x - mu
    xn = (xc)/(np.sqrt(var + eps))

    # scale and shift the normalized activations
    out = (gamma * xn) + beta

    # store variables you need for the backward pass in the 'cache' variable
    cache = [eps, mu, var, gamma, beta, x, xc, xn]

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

  elif mode == 'test':

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the testing time normalized activation.  Normalize using
    #   the running mean and variance, and then scale and shift appropriately.
    #   Store the output as 'out'.
    # ================================================================ #

    # calculate the testing time normalized activation

    # normalize using the running mean and variance from training statistics
    xn = (x - running_mean)/(np.sqrt(running_var + eps))

    # scale and shift appropriately
    out = (gamma * xn) + beta

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

  else:
    raise ValueError('Invalid forward batchnorm mode "%s"' % mode)

  # Store the updated running means back into bn_param
  bn_param['running_mean'] = running_mean
  bn_param['running_var'] = running_var

  return out, cache

def batchnorm_backward(dout, cache):
  """
  Backward pass for batch normalization.

  For this implementation, you should write out a computation graph for
  batch normalization on paper and propagate gradients backward through
  intermediate nodes.

  Inputs:
  - dout: Upstream derivatives, of shape (N, D)
```

```
    - cache: Variable of intermediates from batchnorm_forward.

    Returns a tuple of:
    - dx: Gradient with respect to inputs x, of shape (N, D)
    - dgamma: Gradient with respect to scale parameter gamma, of shape (D,)
    - dbeta: Gradient with respect to shift parameter beta, of shape (D,)
    """
    dx, dgamma, dbeta = None, None, None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the batchnorm backward pass, calculating dx, dgamma, and dbeta.
    # ================================================================ #


    # first pull out relevant variables
    eps, mu, var, gamma, beta, x, xc, xn = cache
    N, D = dout.shape

    # the easy ones: dbeta, dgamma
    dbeta = np.sum(dout, axis=0)
    dgamma = np.sum(dout * xn, axis=0)

    # the hard one: dx
    dxn = dout * gamma

    # intermediate variables
    sqrtvar = np.sqrt(var + eps)
    ivar = 1 / sqrtvar

    # intermediate steps for dvar
    dvar = np.sum(dxn * xc, axis=0) * (-0.5) * np.power(sqrtvar, -3)

    # intermediate steps for dmu
    # dmu = -np.sum(dxn * ivar, axis=0) + dvar * np.sum(-2. * xc, axis=0) / N
    dmu = -ivar * np.sum(dxn, axis=0) - (2.0 * dvar / N) * np.sum(xc, axis=0)

    # final step for dx
    # dx = (dxn * ivar) + (dvar * 2 * xc / N) + (dmu / N)
    dx = (ivar * dxn) + (dmu * np.ones_like(dout) / N) + (2.0 * xc / N) * dvar

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

    return dx, dgamma, dbeta

def dropout_forward(x, dropout_param):
    """
    Performs the forward pass for (inverted) dropout.

    Inputs:
    - x: Input data, of any shape
    - dropout_param: A dictionary with the following keys:
      - p: Dropout parameter. We keep each neuron output with probability p.
      - mode: 'test' or 'train'. If the mode is train, then perform dropout;
        if the mode is test, then just return the input.
      - seed: Seed for the random number generator. Passing seed makes this
        function deterministic, which is needed for gradient checking but not in
        real networks.

    Outputs:
    - out: Array of the same shape as x.
    - cache: A tuple (dropout_param, mask). In training mode, mask is the dropout
      mask that was used to multiply the input; in test mode, mask is None.
    """
    p, mode = dropout_param['p'], dropout_param['mode']
```

```python
  if 'seed' in dropout_param:
    np.random.seed(dropout_param['seed'])

  mask = None
  out = None

  if mode == 'train':
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the inverted dropout forward pass during training time.
    #   Store the masked and scaled activations in out, and store the
    #   dropout mask as the variable mask.
    # ================================================================ #

    mask = (np.random.rand(*x.shape) < p) / p
    out = x * mask

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

  elif mode == 'test':

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the inverted dropout forward pass during test time.
    # ================================================================ #

    out = x

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

  cache = (dropout_param, mask)
  out = out.astype(x.dtype, copy=False)

  return out, cache

def dropout_backward(dout, cache):
  """
  Perform the backward pass for (inverted) dropout.

  Inputs:
  - dout: Upstream derivatives, of any shape
  - cache: (dropout_param, mask) from dropout_forward.
  """
  dropout_param, mask = cache
  mode = dropout_param['mode']

  dx = None
  if mode == 'train':
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the inverted dropout backward pass during training time.
    # ================================================================ #

    dx = dout * mask

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #
  elif mode == 'test':
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the inverted dropout backward pass during test time.
    # ================================================================ #
```

```python
    dx = dout

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #
  return dx

def svm_loss(x, y):
  """
  Computes the loss and gradient using for multiclass SVM classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

  Returns a tuple of:
  - loss: Scalar giving the loss
  - dx: Gradient of the loss with respect to x
  """
  N = x.shape[0]
  correct_class_scores = x[np.arange(N), y]
  margins = np.maximum(0, x - correct_class_scores[:, np.newaxis] + 1.0)
  margins[np.arange(N), y] = 0
  loss = np.sum(margins) / N
  num_pos = np.sum(margins > 0, axis=1)
  dx = np.zeros_like(x)
  dx[margins > 0] = 1
  dx[np.arange(N), y] -= num_pos
  dx /= N
  return loss, dx


def softmax_loss(x, y):
  """
  Computes the loss and gradient for softmax classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

  Returns a tuple of:
  - loss: Scalar giving the loss
  - dx: Gradient of the loss with respect to x
  """

  probs = np.exp(x - np.max(x, axis=1, keepdims=True))
  probs /= np.sum(probs, axis=1, keepdims=True)
  N = x.shape[0]
  loss = -np.sum(np.log(probs[np.arange(N), y])) / N
  dx = probs.copy()
  dx[np.arange(N), y] -= 1
  dx /= N
  return loss, dx
```

```python
import numpy as np
import pdb

from .layers import *
from .layer_utils import *

"""
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 TwoLayerNet(object):
  """
  A two-layer fully-connected neural network with ReLU nonlinearity and
  softmax loss that uses a modular layer design. We assume an input dimension
  of D, a hidden dimension of H, and perform classification over C classes.

  The architecure should be affine - relu - affine - softmax.

  Note that this class does not implement gradient descent; instead, it
  will interact with a separate Solver object that is responsible for running
  optimization.

  The learnable parameters of the model are stored in the dictionary
  self.params that maps parameter names to numpy arrays.
  """

  def __init__(self, input_dim=3*32*32, hidden_dims=100, num_classes=10,
               dropout=0, weight_scale=1e-3, reg=0.0):
    """
    Initialize a new network.

    Inputs:
    - input_dim: An integer giving the size of the input
    - hidden_dims: An integer giving the size of the hidden layer
    - num_classes: An integer giving the number of classes to classify
    - dropout: Scalar between 0 and 1 giving dropout strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - reg: Scalar giving L2 regularization strength.
    """
    self.params = {}
    self.reg = reg

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize W1, W2, b1, and b2.  Store these as self.params['W1'],
    #   self.params['W2'], self.params['b1'] and self.params['b2']. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    #   The dimensions of W1 should be (input_dim, hidden_dim) and the
    #   dimensions of W2 should be (hidden_dims, num_classes)
    # ================================================================ #


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

  def loss(self, X, y=None):
    """
    Compute loss and gradient for a minibatch of data.
```

```python
    Inputs:
    - X: Array of input data of shape (N, d_1, ..., d_k)
    - y: Array of labels, of shape (N,). y[i] gives the label for X[i].

    Returns:
    If y is None, then run a test-time forward pass of the model and return:
    - scores: Array of shape (N, C) giving classification scores, where
      scores[i, c] is the classification score for X[i] and class c.

    If y is not None, then run a training-time forward and backward pass and
    return a tuple of:
    - loss: Scalar value giving the loss
    - grads: Dictionary with the same keys as self.params, mapping parameter
      names to gradients of the loss with respect to those parameters.
    """
    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the forward pass of the two-layer neural network. Store
    #   the class scores as the variable 'scores'.  Be sure to use the layers
    #   you prior implemented.
    # ================================================================ #

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

    # If y is None then we are in test mode so just return scores
    if y is None:
      return scores

    loss, grads = 0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backward pass of the two-layer neural net.  Store
    #   the loss as the variable 'loss' and store the gradients in the
    #   'grads' dictionary.  For the grads dictionary, grads['W1'] holds
    #   the gradient for W1, grads['b1'] holds the gradient for b1, etc.
    #   i.e., grads[k] holds the gradient for self.params[k].
    #
    #   Add L2 regularization, where there is an added cost 0.5*self.reg*W^2
    #   for each W.  Be sure to include the 0.5 multiplying factor to
    #   match our implementation.
    #
    #   And be sure to use the layers you prior implemented.
    # ================================================================ #

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

    return loss, grads


class FullyConnectedNet(object):
  """
  A fully-connected neural network with an arbitrary number of hidden layers,
  ReLU nonlinearities, and a softmax loss function. This will also implement
  dropout and batch normalization as options. For a network with L layers,
  the architecture will be

  {affine - [batch norm] - relu - [dropout]} x (L - 1) - affine - softmax

  where batch normalization and dropout are optional, and the {...} block is
  repeated L - 1 times.
```

```
  Similar to the TwoLayerNet above, learnable parameters are stored in the
  self.params dictionary and will be learned using the Solver class.
  """

  def __init__(self, hidden_dims, input_dim=3*32*32, num_classes=10,
               dropout=0, use_batchnorm=False, reg=0.0,
               weight_scale=1e-2, dtype=np.float32, seed=None):
    """
    Initialize a new FullyConnectedNet.

    Inputs:
    - hidden_dims: A list of integers giving the size of each hidden layer.
    - input_dim: An integer giving the size of the input.
    - num_classes: An integer giving the number of classes to classify.
    - dropout: Scalar between 0 and 1 giving dropout strength. If dropout=0 then
      the network should not use dropout at all.
    - use_batchnorm: Whether or not the network should use batch normalization.
    - reg: Scalar giving L2 regularization strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - dtype: A numpy datatype object; all computations will be performed using
      this datatype. float32 is faster but less accurate, so you should use
      float64 for numeric gradient checking.
    - seed: If not None, then pass this random seed to the dropout layers. This
      will make the dropout layers deteriminstic so we can gradient check the
      model.
    """
    self.use_batchnorm = use_batchnorm
    self.use_dropout = dropout > 0
    self.reg = reg
    self.num_layers = 1 + len(hidden_dims)
    self.dtype = dtype
    self.params = {}

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize all parameters of the network in the self.params dictionary.
    #   The weights and biases of layer 1 are W1 and b1; and in general the
    #   weights and biases of layer i are Wi and bi. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    # ================================================================ #

    layer_dims = [input_dim] + hidden_dims + [num_classes]

    for i in range(self.num_layers):
      self.params["W{}".format(i + 1)] = weight_scale * np.random.randn(layer_dims[i],
layer_dims[i + 1])
      self.params["b{}".format(i + 1)] = np.zeros(layer_dims[i + 1])

    # Batchnorm
    if self.use_batchnorm:
      for i in range(self.num_layers - 1):
        self.params["gamma{}".format(i + 1)] = np.ones(layer_dims[i + 1])
        self.params["beta{}".format(i + 1)] = np.zeros(layer_dims[i + 1])

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

    # When using dropout we need to pass a dropout_param dictionary to each
    # dropout layer so that the layer knows the dropout probability and the mode
    # (train / test). You can pass the same dropout_param to each dropout layer.
    self.dropout_param = {}
    if self.use_dropout:
      self.dropout_param = {'mode': 'train', 'p': dropout}
```

```python
    if seed is not None:
      self.dropout_param['seed'] = seed

  # With batch normalization we need to keep track of running means and
  # variances, so we need to pass a special bn_param object to each batch
  # normalization layer. You should pass self.bn_params[0] to the forward pass
  # of the first batch normalization layer, self.bn_params[1] to the forward
  # pass of the second batch normalization layer, etc.
  self.bn_params = []
  if self.use_batchnorm:
    self.bn_params = [{'mode': 'train'} for i in np.arange(self.num_layers - 1)]

  # Cast all parameters to the correct datatype
  for k, v in self.params.items():
    self.params[k] = v.astype(dtype)


def loss(self, X, y=None):
  """
  Compute loss and gradient for the fully-connected net.

  Input / output: Same as TwoLayerNet above.
  """
  X = X.astype(self.dtype)
  mode = 'test' if y is None else 'train'

  # Set train/test mode for batchnorm params and dropout param since they
  # behave differently during training and testing.
  if self.dropout_param is not None:
    self.dropout_param['mode'] = mode
  if self.use_batchnorm:
    for bn_param in self.bn_params:
      bn_param[mode] = mode

  scores = None

  # ================================================================ #
  # YOUR CODE HERE:
  #   Implement the forward pass of the FC net and store the output
  #   scores as the variable "scores".
  # ================================================================ #

  caches = {}

  for i in range(1, self.num_layers):
    W = self.params[f"W{i}"]
    b = self.params[f"b{i}"]

    if self.use_batchnorm:
      gamma = self.params[f"gamma{i}"]
      beta = self.params[f"beta{i}"]
      X, cache = bn_affine_relu_forward(X, W, b, gamma, beta, self.bn_params[i - 1])
    else:
      X, cache = affine_relu_forward(X, W, b)

    caches[i] = cache

    if self.use_dropout:
      X, cache = dropout_forward(X, self.dropout_param)
      caches[f"dropout{i}"] = cache

  W, b = self.params[f"W{self.num_layers}"], self.params[f"b{self.num_layers}"]
  scores, cache = affine_forward(X, W, b)
  caches[self.num_layers] = cache

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

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

    # If test mode return early
    if mode == 'test':
      return scores

    loss, grads = 0.0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backwards pass of the FC net and store the gradients
    #   in the grads dict, so that grads[k] is the gradient of self.params[k]
    #   Be sure your L2 regularization includes a 0.5 factor.
    # ================================================================ #

    # Compute the softmax loss and its gradient
    loss, dout = softmax_loss(scores, y)

    # Add regularization to the loss more efficiently
    weights_squared_sum = sum(np.sum(self.params['W' + str(i + 1)] ** 2) for i in
range(self.num_layers))
    loss += 0.5 * self.reg * weights_squared_sum

    # Backward pass for the last layer
    dout, grads['W' + str(self.num_layers)], grads['b' + str(self.num_layers)] =
affine_backward(dout, caches[self.num_layers])
    grads['W' + str(self.num_layers)] += self.reg * self.params['W' + str(self.num_layers)]

    # Iterate over layers in reverse for backprop, skipping the last since it's already handled
    for i in range(self.num_layers - 1, 0, -1):
      if self.use_dropout:
        dout = dropout_backward(dout, caches[f'dropout{i}']) # dropout

      if self.use_batchnorm:
        dout, dw, db, dgamma, dbeta = bn_affine_relu_backward(dout, caches[i])
        grads[f"gamma{i}"] = dgamma
        grads[f"beta{i}"] = dbeta
      else:
        dout, dw, db = affine_relu_backward(dout, caches[i])

      W = self.params[f"W{i}"]
      grads[f"W{i}"] = dw + self.reg * W
      grads[f"b{i}"] = db

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #
    return loss, grads
```