Fully-Connected Neural Nets

In the previous homework you implemented a fully-connected two-layer neural network on CIFAR-10. The implementation was simple but not very modular since the loss and gradient were computed in a single monolithic function. This is manageable for a simple two-layer network, but would become impractical as we move to bigger models. Ideally we want to build networks using a more modular design so that we can implement different layer types in isolation and then snap them together into models with different architectures.

In this exercise we will implement fully-connected networks using a more modular approach. For each layer we will implement a `forward` and a `backward` function. The `forward` function will receive inputs, weights, and other parameters and will return both an output and a `cache` object storing data needed for the backward pass, like this:

```
def layer_forward(x, w):
  """ Receive inputs x and weights w """
  # Do some computations ...
  z = # ... some intermediate value
  # Do some more computations ...
  out = # the output

  cache = (x, w, z, out) # Values we need to compute gradients

  return out, cache
```

The backward pass will receive upstream derivatives and the `cache` object, and will return gradients with respect to the inputs and weights, like this:

```
def layer_backward(dout, cache):
  """
  Receive dout (derivative of loss with respect to outputs) and cache,
  and compute derivative with respect to inputs.
  """
  # Unpack cache values
  x, w, z, out = cache

  # Use values in cache to compute derivatives
  dx = # Derivative of loss with respect to x
  dw = # Derivative of loss with respect to w

  return dx, dw
```

After implementing a bunch of layers this way, we will be able to easily combine them to build classifiers with different architectures.

In addition to implementing fully-connected networks of arbitrary depth, we will also explore different update rules for optimization, and introduce Batch/Layer Normalization as a tool to more efficiently optimize deep networks.

In [ ]:
```
# This mounts your Google Drive to the Colab VM.
from google.colab import drive
drive.mount('/content/drive')

# TODO: Enter the foldername in your Drive where you have saved the unzipped
# assignment folder, e.g. 'cs6353/assignments/assignment3/'
FOLDERNAME = 'CS6353/Assignments/assignment3/assignment3/'
assert FOLDERNAME is not None, "[!] Enter the foldername."

# Now that we've mounted your Drive, this ensures that
# the Python interpreter of the Colab VM can load
# python files from within it.
import sys
sys.path.append('/content/drive/My Drive/{}'.format(FOLDERNAME))

# This downloads the CIFAR-10 dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs6353/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME

# Install requirements from colab_requirements.txt
# TODO: Please change your path below to the colab_requirements.txt file
! python -m pip install -r /content/drive/My\ Drive/$FOLDERNAME/requirements.txt
```

```
Mounted at /content/drive
/content/drive/My Drive/CS6353/Assignments/assignment3/assignment3/cs6353/datasets
--2024-10-18 16:18:31--  https://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz
Resolving www.cs.toronto.edu (www.cs.toronto.edu)... 128.100.3.30
Connecting to www.cs.toronto.edu (www.cs.toronto.edu)|128.100.3.30|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 170498071 (163M) [application/x-gzip]
Saving to: ‘cifar-10-python.tar.gz’

cifar-10-python.tar 100%[===================>] 162.60M  36.4MB/s    in 4.7s

2024-10-18 16:18:35 (34.5 MB/s) - ‘cifar-10-python.tar.gz’ saved [170498071/17049807
1]

cifar-10-batches-py/
cifar-10-batches-py/data_batch_4
cifar-10-batches-py/readme.html
cifar-10-batches-py/test_batch
cifar-10-batches-py/data_batch_3
cifar-10-batches-py/batches.meta
cifar-10-batches-py/data_batch_2
cifar-10-batches-py/data_batch_5
cifar-10-batches-py/data_batch_1
/content/drive/My Drive/CS6353/Assignments/assignment3/assignment3
Collecting attrs==19.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignme
nt3/assignment3//requirements.txt (line 1))
  Downloading attrs-19.1.0-py2.py3-none-any.whl.metadata (10 kB)
Collecting backcall==0.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assign
ment3/assignment3//requirements.txt (line 2))
  Downloading backcall-0.1.0.zip (11 kB)
  Preparing metadata (setup.py) ... done
Collecting bleach==3.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignme
nt3/assignment3//requirements.txt (line 3))
  Downloading bleach-3.1.0-py2.py3-none-any.whl.metadata (19 kB)
Collecting certifi==2019.6.16 (from -r /content/drive/My Drive/CS6353/Assignments/ass
ignment3/assignment3//requirements.txt (line 4))
  Downloading certifi-2019.6.16-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting cycler==0.10.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 5))
  Downloading cycler-0.10.0-py2.py3-none-any.whl.metadata (722 bytes)
Collecting decorator==4.4.0 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 6))
  Downloading decorator-4.4.0-py2.py3-none-any.whl.metadata (3.7 kB)
Collecting defusedxml==0.6.0 (from -r /content/drive/My Drive/CS6353/Assignments/assi
gnment3/assignment3//requirements.txt (line 7))
  Downloading defusedxml-0.6.0-py2.py3-none-any.whl.metadata (31 kB)
Collecting entrypoints==0.3 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 8))
  Downloading entrypoints-0.3-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting future==0.17.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 9))
  Downloading future-0.17.1.tar.gz (829 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 829.1/829.1 kB 33.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting imageio==2.5.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 10))
  Downloading imageio-2.5.0-py3-none-any.whl.metadata (2.8 kB)
Collecting ipykernel==5.1.2 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 11))
  Downloading ipykernel-5.1.2-py3-none-any.whl.metadata (919 bytes)
```

```
Collecting ipython==7.8.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 12))
  Downloading ipython-7.8.0-py3-none-any.whl.metadata (4.3 kB)
Requirement already satisfied: ipython-genutils==0.2.0 in /usr/local/lib/python3.10/dist-packages (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 13)) (0.2.0)
Collecting ipywidgets==7.5.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 14))
  Downloading ipywidgets-7.5.1-py2.py3-none-any.whl.metadata (1.8 kB)
Collecting jedi==0.15.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 15))
  Downloading jedi-0.15.1-py2.py3-none-any.whl.metadata (15 kB)
Collecting Jinja2==2.10.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 16))
  Downloading Jinja2-2.10.1-py2.py3-none-any.whl.metadata (2.2 kB)
Collecting jsonschema==3.0.2 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 17))
  Downloading jsonschema-3.0.2-py2.py3-none-any.whl.metadata (7.4 kB)
Collecting jupyter==1.0.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 18))
  Downloading jupyter-1.0.0-py2.py3-none-any.whl.metadata (995 bytes)
Collecting jupyter-client==5.3.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 19))
  Downloading jupyter_client-5.3.1-py2.py3-none-any.whl.metadata (3.6 kB)
Collecting jupyter-console==6.0.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 20))
  Downloading jupyter_console-6.0.0-py2.py3-none-any.whl.metadata (955 bytes)
Collecting jupyter-core==4.5.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 21))
  Downloading jupyter_core-4.5.0-py2.py3-none-any.whl.metadata (884 bytes)
Collecting kiwisolver==1.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 22))
  Downloading kiwisolver-1.1.0.tar.gz (30 kB)
  Preparing metadata (setup.py) ... done
Collecting MarkupSafe==1.1.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 23))
  Downloading MarkupSafe-1.1.1.tar.gz (19 kB)
  Preparing metadata (setup.py) ... done
Collecting matplotlib==3.1.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 24))
  Downloading matplotlib-3.1.1.tar.gz (37.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 37.8/37.8 MB 27.7 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: mistune==0.8.4 in /usr/local/lib/python3.10/dist-packages (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 25)) (0.8.4)
ERROR: Could not find a version that satisfies the requirement mkl-fft==1.0.6 (from versions: 1.3.6, 1.3.8)
ERROR: No matching distribution found for mkl-fft==1.0.6
```

In [ ]:

```
# As usual, a bit of setup
from __future__ import print_function
import time
import numpy as np
import matplotlib.pyplot as plt
from cs6353.classifiers.fc_net import *
from cs6353.data_utils import get_CIFAR10_data
from cs6353.gradient_check import eval_numerical_gradient, eval_numerical_gradient_arr
from cs6353.solver import Solver
```

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

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

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

In [ ]:
```
# Load the (preprocessed) CIFAR10 data.
data = get_CIFAR10_data()
for k, v in list(data.items()):
  print(('%s: ' % k, v.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,))
```

Affine layer: foward

Open the file `cs6353/layers.py` and implement the `affine_forward` function.

Once you are done you can test your implementaion by running the following:

In [ ]:
```
# Test the affine_forward function

num_inputs = 2
input_shape = (4, 5, 6)
output_dim = 3

input_size = num_inputs * np.prod(input_shape)
weight_size = output_dim * np.prod(input_shape)

x = np.linspace(-0.1, 0.5, num=input_size).reshape(num_inputs, *input_shape)
w = np.linspace(-0.2, 0.3, num=weight_size).reshape(np.prod(input_shape), output_dim)
b = np.linspace(-0.3, 0.1, num=output_dim)

out, _ = affine_forward(x, w, b)
correct_out = np.array([[ 1.49834967,  1.70660132,  1.91485297],
                        [ 3.25553199,  3.5141327,   3.77273342]])

# Compare your output with ours. The error should be around e-9 or less.
print('Testing affine_forward function:')
print('difference: ', rel_error(out, correct_out))
```

```
Testing affine_forward function:
difference:  9.769849468192957e-10
```

Affine layer: backward

Now implement the `affine_backward` function and test your implementation using numeric gradient checking.

In [ ]:
```
# Test the affine_backward function
np.random.seed(231)
x = np.random.randn(10, 2, 3)
w = np.random.randn(6, 5)
b = np.random.randn(5)
dout = np.random.randn(10, 5)

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

_, cache = affine_forward(x, w, b)
dx, dw, db = affine_backward(dout, cache)

# The error should be around e-10 or less
print('Testing affine_backward function:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
Testing affine_backward function:
dx error:  5.399100368651805e-11
dw error:  9.904211865398145e-11
db error:  2.4122867568119087e-11
```

ReLU activation: forward

Implement the forward pass for the ReLU activation function in the `relu_forward` function and test your implementation using the following:

In [ ]:
```
# Test the relu_forward function

x = np.linspace(-0.5, 0.5, num=12).reshape(3, 4)

out, _ = relu_forward(x)
correct_out = np.array([[ 0.,          0.,          0.,          0.,        ],
                        [ 0.,          0.,          0.04545455,  0.13636364,],
                        [ 0.22727273,  0.31818182,  0.40909091,  0.5,       ]])

# Compare your output with ours. The error should be on the order of e-8
print('Testing relu_forward function:')
print('difference: ', rel_error(out, correct_out))
```

```
Testing relu_forward function:
difference:  4.999999798022158e-08
```

ReLU activation: backward

Now implement the backward pass for the ReLU activation function in the `relu_backward` function and test your implementation using numeric gradient checking:

In [ ]:
```
np.random.seed(231)
x = np.random.randn(10, 10)
dout = np.random.randn(*x.shape)

dx_num = eval_numerical_gradient_array(lambda x: relu_forward(x)[0], x, dout)

_, cache = relu_forward(x)
dx = relu_backward(dout, cache)

# The error should be on the order of e-12
print('Testing relu_backward function:')
print('dx error: ', rel_error(dx_num, dx))
```

```
Testing relu_backward function:
dx error:  3.2756349136310288e-12
```

Inline Question 1:

We've only asked you to implement ReLU, but there are a number of different activation functions that one could use in neural networks, each with its pros and cons. In particular, an issue commonly seen with activation functions is getting zero (or close to zero) gradient flow during backpropagation. Which of the following activation functions have this problem? If you consider these functions in the one dimensional case, what types of input would lead to this behaviour?

1. Sigmoid
2. ReLU
3. Leaky ReLU

Answer:

1. Sigmoid: The sigmoid function is prone to the vanishing gradient problem because its gradient diminishes near zero when the input values are very large or very small. For instance, input values like [-10000, 100000] would cause this issue as the function enters saturation, where the gradients become tiny.

2. ReLU: ReLU, compared to sigmoid, is less prone to vanishing gradients because it responds linearly to positive inputs. Its gradient is either 1 for positive inputs or 0 for negative inputs. However, when all inputs are negative, the output becomes zero, leading to no gradient flow, which is when we call it the "dying ReLU" problem. An example of this would be if the inputs were all negative, like [-12, -13, -14].

3. Leaky ReLU: To address the problem of dying RELU, Leaky ReLU introduces a small negative slope for negative input values, i.e., for $x < 0$, it outputs $0.0001x$, and for $x \geq 0$, it returns x. This modification helps mitigate the vanishing gradient problem. However, since the

function isn't continuous at x = 0, the gradient is undefined at that point. If this isn't handled properly in the implementation, an example input like [0, 0, 0] could result in zero gradients.

"Sandwich" layers

There are some common patterns of layers that are frequently used in neural nets. For example, affine layers are frequently followed by a ReLU nonlinearity. To make these common patterns easy, we define several convenience layers in the file `cs6353/layer_utils.py`.

For now take a look at the `affine_relu_forward` and `affine_relu_backward` functions, and run the following to numerically gradient check the backward pass:

In [ ]:
```
from cs6353.layer_utils import affine_relu_forward, affine_relu_backward
np.random.seed(231)
x = np.random.randn(2, 3, 4)
w = np.random.randn(12, 10)
b = np.random.randn(10)
dout = np.random.randn(2, 10)

out, cache = affine_relu_forward(x, w, b)
dx, dw, db = affine_relu_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: affine_relu_forward(x, w, b)[0], x, d
dw_num = eval_numerical_gradient_array(lambda w: affine_relu_forward(x, w, b)[0], w, d
db_num = eval_numerical_gradient_array(lambda b: affine_relu_forward(x, w, b)[0], b, d

# Relative error should be around e-10 or less
print('Testing affine_relu_forward and affine_relu_backward:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
Testing affine_relu_forward and affine_relu_backward:
dx error:  2.299579177309368e-11
dw error:  8.162011105764925e-11
db error:  7.826724021458994e-12
```

Loss layers: Softmax and SVM

You implemented these loss functions in the last assignment, so we'll give them to you for free here. You should still make sure you understand how they work by looking at the implementations in `cs6353/layers.py`.

You can make sure that the implementations are correct by running the following:

In [ ]:
```
np.random.seed(231)
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)
```

```
dx_num = eval_numerical_gradient(lambda x: svm_loss(x, y)[0], x, verbose=False)
loss, dx = svm_loss(x, y)

# Test svm_loss function. Loss should be around 9 and dx error should be around the or
print('Testing svm_loss:')
print('loss: ', loss)
print('dx error: ', rel_error(dx_num, dx))

dx_num = eval_numerical_gradient(lambda x: softmax_loss(x, y)[0], x, verbose=False)
loss, dx = softmax_loss(x, y)

# Test softmax_loss function. Loss should be close to 2.3 and dx error should be aroun
print('\nTesting softmax_loss:')
print('loss: ', loss)
print('dx error: ', rel_error(dx_num, dx))
```

```
Testing svm_loss:
loss:  8.999602749096233
dx error:  1.4021566006651672e-09

Testing softmax_loss:
loss:  2.302545844500738
dx error:  9.384673161989355e-09
```

Two-layer network

In the previous assignment you implemented a two-layer neural network in a single monolithic class. Now that you have implemented modular versions of the necessary layers, you will reimplement the two layer network using these modular implementations.

Open the file `cs6353/classifiers/fc_net.py` and complete the implementation of the `TwoLayerNet` class. This class will serve as a model for the other networks you will implement in this assignment, so read through it to make sure you understand the API. You can run the cell below to test your implementation.

In [ ]:

```
np.random.seed(231)
N, D, H, C = 3, 5, 50, 7
X = np.random.randn(N, D)
y = np.random.randint(C, size=N)

std = 1e-3
model = TwoLayerNet(input_dim=D, hidden_dim=H, num_classes=C, weight_scale=std)

print('Testing initialization ... ')
W1_std = abs(model.params['W1'].std() - std)
b1 = model.params['b1']
W2_std = abs(model.params['W2'].std() - std)
b2 = model.params['b2']
assert W1_std < std / 10, 'First layer weights do not seem right'
assert np.all(b1 == 0), 'First layer biases do not seem right'
assert W2_std < std / 10, 'Second layer weights do not seem right'
assert np.all(b2 == 0), 'Second layer biases do not seem right'

print('Testing test-time forward pass ... ')
model.params['W1'] = np.linspace(-0.7, 0.3, num=D*H).reshape(D, H)
```

```
model.params['b1'] = np.linspace(-0.1, 0.9, num=H)
model.params['W2'] = np.linspace(-0.3, 0.4, num=H*C).reshape(H, C)
model.params['b2'] = np.linspace(-0.9, 0.1, num=C)
X = np.linspace(-5.5, 4.5, num=N*D).reshape(D, N).T
scores = model.loss(X)
correct_scores = np.asarray(
  [[11.53165108,  12.2917344,   13.05181771,  13.81190102,  14.57198434, 15.33206765,
   [12.05769098,  12.74614105,  13.43459113,  14.1230412,   14.81149128, 15.49994135,
   [12.58373087,  13.20054771,  13.81736455,  14.43418138,  15.05099822, 15.66781506,
scores_diff = np.abs(scores - correct_scores).sum()
assert scores_diff < 1e-6, 'Problem with test-time forward pass'

print('Testing training loss (no regularization)')
y = np.asarray([0, 5, 1])
loss, grads = model.loss(X, y)
correct_loss = 3.4702243556
assert abs(loss - correct_loss) < 1e-10, 'Problem with training-time loss'

model.reg = 1.0
loss, grads = model.loss(X, y)
correct_loss = 26.5948426952
assert abs(loss - correct_loss) < 1e-10, 'Problem with regularization loss'

# Errors should be around e-7 or less
for reg in [0.0, 0.7]:
  print('Running numeric gradient check with reg = ', reg)
  model.reg = reg
  loss, grads = model.loss(X, y)

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False)
    print('%s relative error: %.2e' % (name, rel_error(grad_num, grads[name])))
```

```
Testing initialization ...
Testing test-time forward pass ...
Testing training loss (no regularization)
Running numeric gradient check with reg =  0.0
W1 relative error: 1.83e-08
W2 relative error: 3.12e-10
b1 relative error: 9.83e-09
b2 relative error: 4.33e-10
Running numeric gradient check with reg =  0.7
W1 relative error: 2.53e-07
W2 relative error: 2.85e-08
b1 relative error: 1.56e-08
b2 relative error: 7.76e-10
```

Solver

In the previous assignment, the logic for training models was coupled to the models themselves. Following a more modular design, for this assignment we have split the logic for training models into a separate class.

Open the file `cs6353/solver.py` and read through it to familiarize yourself with the API. After doing so, use a `Solver` instance to train a `TwoLayerNet` that achieves at least `50%` accuracy

on the validation set.

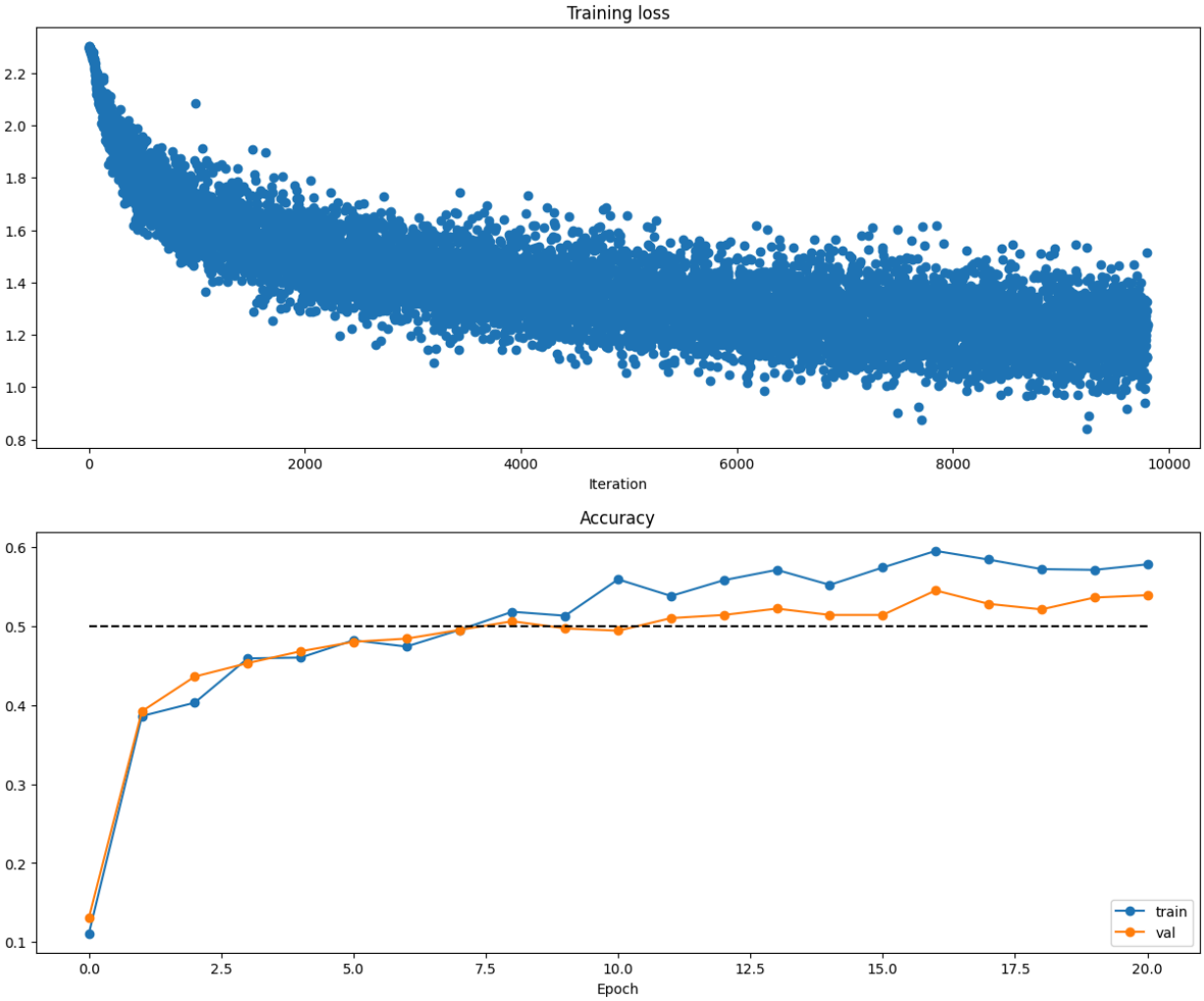
In [ ]:
```
model = TwoLayerNet()
solver = None

################################################################################
# TODO: Use a Solver instance to train a TwoLayerNet that achieves at least  #
# 50% accuracy on the validation set.                                          #
################################################################################

# lr_list = [0.0001, 0.0002, 0.0003, 0.0004]
# reg_list = [0.00001, 0.000015, 0.00002, 0.000025]
# layer_list = [25, 50, 100]
# epoch_list = [5, 10, 15, 20]

lr_list = [0.0001, 0.0002, 0.0003]
reg_list = [0.00001, 0.000015, 0.00002, 0.000025]
layer_list = [50, 100]
epoch_list = [10, 15, 20]

# lr_list = [0.0003]
# reg_list = [0.000025]
# layer_list = [100]
# epoch_list = [20]

best_val_acc = -1
for lr in lr_list:
  for regularization in reg_list:
    for layer in layer_list:
      for epoch in epoch_list:
        print('lr: {}, reg: {}, layer: {}, epoch: {}'.format(lr, regularization, layer
        model = TwoLayerNet(hidden_dim=layer, reg=regularization)
        candidate_solver = Solver(model, data, optim_config={'learning_rate':lr},
                                  lr_decay = 0.95, num_epochs=epoch, verbose=False)
        candidate_solver.train()
        if candidate_solver.best_val_acc > best_val_acc:
          best_val_acc = candidate_solver.best_val_acc
          solver = candidate_solver
          print('lr: {}, reg: {}, layer: {}, epoch: {}, Val acc: {}'.format(lr, regula

print('Best val acc: {}'.format(best_val_acc))
################################################################################
#                              END OF YOUR CODE                                #
################################################################################
```

```
lr: 0.0001, reg: 1e-05, layer: 50, epoch: 10
lr: 0.0001, reg: 1e-05, layer: 50, epoch: 10, Val acc: 0.461
lr: 0.0001, reg: 1e-05, layer: 50, epoch: 15
lr: 0.0001, reg: 1e-05, layer: 50, epoch: 15, Val acc: 0.477
lr: 0.0001, reg: 1e-05, layer: 50, epoch: 20
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 10
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 15
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 15, Val acc: 0.487
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 20, Val acc: 0.492
lr: 0.0001, reg: 1.5e-05, layer: 50, epoch: 10
lr: 0.0001, reg: 1.5e-05, layer: 50, epoch: 15
lr: 0.0001, reg: 1.5e-05, layer: 50, epoch: 20
lr: 0.0001, reg: 1.5e-05, layer: 100, epoch: 10
lr: 0.0001, reg: 1.5e-05, layer: 100, epoch: 15
lr: 0.0001, reg: 1.5e-05, layer: 100, epoch: 20
lr: 0.0001, reg: 1.5e-05, layer: 100, epoch: 20, Val acc: 0.496
lr: 0.0001, reg: 2e-05, layer: 50, epoch: 10
lr: 0.0001, reg: 2e-05, layer: 50, epoch: 15
lr: 0.0001, reg: 2e-05, layer: 50, epoch: 20
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 10
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 15
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 20, Val acc: 0.499
lr: 0.0001, reg: 2.5e-05, layer: 50, epoch: 10
lr: 0.0001, reg: 2.5e-05, layer: 50, epoch: 15
lr: 0.0001, reg: 2.5e-05, layer: 50, epoch: 20
lr: 0.0001, reg: 2.5e-05, layer: 100, epoch: 10
lr: 0.0001, reg: 2.5e-05, layer: 100, epoch: 15
lr: 0.0001, reg: 2.5e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 1e-05, layer: 50, epoch: 10
lr: 0.0002, reg: 1e-05, layer: 50, epoch: 10, Val acc: 0.505
lr: 0.0002, reg: 1e-05, layer: 50, epoch: 15
lr: 0.0002, reg: 1e-05, layer: 50, epoch: 20
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 10
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 15
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 20, Val acc: 0.527
lr: 0.0002, reg: 1.5e-05, layer: 50, epoch: 10
lr: 0.0002, reg: 1.5e-05, layer: 50, epoch: 15
lr: 0.0002, reg: 1.5e-05, layer: 50, epoch: 20
lr: 0.0002, reg: 1.5e-05, layer: 100, epoch: 10
lr: 0.0002, reg: 1.5e-05, layer: 100, epoch: 15
lr: 0.0002, reg: 1.5e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 2e-05, layer: 50, epoch: 10
lr: 0.0002, reg: 2e-05, layer: 50, epoch: 15
lr: 0.0002, reg: 2e-05, layer: 50, epoch: 20
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 10
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 15
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 20, Val acc: 0.545
lr: 0.0002, reg: 2.5e-05, layer: 50, epoch: 10
lr: 0.0002, reg: 2.5e-05, layer: 50, epoch: 15
lr: 0.0002, reg: 2.5e-05, layer: 50, epoch: 20
lr: 0.0002, reg: 2.5e-05, layer: 100, epoch: 10
lr: 0.0002, reg: 2.5e-05, layer: 100, epoch: 15
lr: 0.0002, reg: 2.5e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 1e-05, layer: 50, epoch: 10
lr: 0.0003, reg: 1e-05, layer: 50, epoch: 15
lr: 0.0003, reg: 1e-05, layer: 50, epoch: 20
```

```
lr: 0.0003, reg: 1e-05, layer: 100, epoch: 10
lr: 0.0003, reg: 1e-05, layer: 100, epoch: 15
lr: 0.0003, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 1.5e-05, layer: 50, epoch: 10
lr: 0.0003, reg: 1.5e-05, layer: 50, epoch: 15
lr: 0.0003, reg: 1.5e-05, layer: 50, epoch: 20
lr: 0.0003, reg: 1.5e-05, layer: 100, epoch: 10
lr: 0.0003, reg: 1.5e-05, layer: 100, epoch: 15
lr: 0.0003, reg: 1.5e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 2e-05, layer: 50, epoch: 10
lr: 0.0003, reg: 2e-05, layer: 50, epoch: 15
lr: 0.0003, reg: 2e-05, layer: 50, epoch: 20
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 10
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 15
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 2.5e-05, layer: 50, epoch: 10
lr: 0.0003, reg: 2.5e-05, layer: 50, epoch: 15
lr: 0.0003, reg: 2.5e-05, layer: 50, epoch: 20
lr: 0.0003, reg: 2.5e-05, layer: 100, epoch: 10
lr: 0.0003, reg: 2.5e-05, layer: 100, epoch: 15
lr: 0.0003, reg: 2.5e-05, layer: 100, epoch: 20
Best val acc: 0.545
```

In [ ]:

```
# Run this cell to visualize training loss and train / val accuracy

plt.subplot(2, 1, 1)
plt.title('Training loss')
plt.plot(solver.loss_history, 'o')
plt.xlabel('Iteration')

plt.subplot(2, 1, 2)
plt.title('Accuracy')
plt.plot(solver.train_acc_history, '-o', label='train')
plt.plot(solver.val_acc_history, '-o', label='val')
plt.plot([0.5] * len(solver.val_acc_history), 'k--')
plt.xlabel('Epoch')
plt.legend(loc='lower right')
plt.gcf().set_size_inches(15, 12)
plt.show()
```



Multilayer network

Next you will implement a fully-connected network with an arbitrary number of hidden layers.

Read through the `FullyConnectedNet` class in the file `cs6353/classifiers/fc_net.py`.

Implement the initialization, the forward pass, and the backward pass. For the moment don't worry about implementing batch/layer normalization; we will add those features soon.

Initial loss and gradient check

As a sanity check, run the following to check the initial loss and to gradient check the network both with and without regularization. Do the initial losses seem reasonable?

For gradient checking, you should expect to see errors around 1e-7 or less.

In [ ]:
```
np.random.seed(231)
N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))
```

```
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)

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

  # Most of the errors should be on the order of e-7 or smaller.
  # NOTE: It is fine however to see an error for W2 on the order of e-5
  # for the check when reg = 0.0
  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('%s relative error: %.2e' % (name, rel_error(grad_num, grads[name])))
```

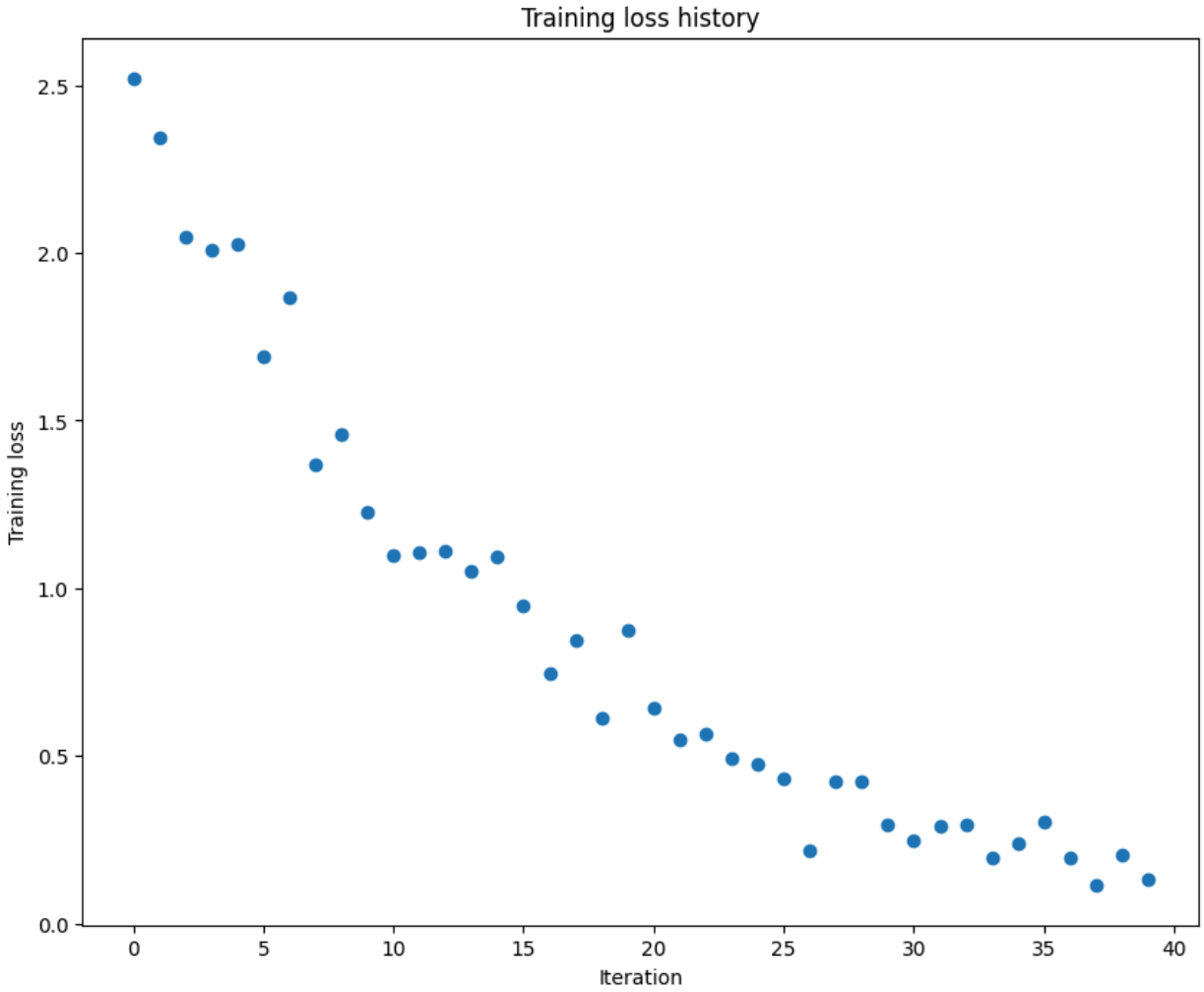
```
Running check with reg =  0
Initial loss:  2.3004790897684924
W1 relative error: 1.48e-07
W2 relative error: 2.21e-05
W3 relative error: 3.53e-07
b1 relative error: 5.38e-09
b2 relative error: 2.09e-09
b3 relative error: 5.80e-11
Running check with reg =  3.14
Initial loss:  7.052114776533016
W1 relative error: 3.90e-09
W2 relative error: 6.87e-08
W3 relative error: 2.13e-08
b1 relative error: 1.48e-08
b2 relative error: 1.72e-09
b3 relative error: 1.57e-10
```

As another sanity check, make sure you can overfit a small dataset of 50 images. First we will try a three-layer network with 100 units in each hidden layer. In the following cell, tweak the learning rate and initialization scale to overfit and achieve 100% training accuracy within 20 epochs.

In [ ]:
```
# TODO: Use a three-layer Net to overfit 50 training examples by
# tweaking just the learning rate and initialization scale.

num_train = 50
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 = 1.5e-2
learning_rate = 3e-3
model = FullyConnectedNet([100, 100],
              weight_scale=weight_scale, dtype=np.float64)
solver = Solver(model, small_data,
                print_every=10, num_epochs=20, batch_size=25,
                update_rule='sgd',
                optim_config={
                  'learning_rate': learning_rate,
                }
```

```
        )
solver.train()

plt.plot(solver.loss_history, 'o')
plt.title('Training loss history')
plt.xlabel('Iteration')
plt.ylabel('Training loss')
plt.show()
```

```
(Iteration 1 / 40) loss: 2.521152
(Epoch 0 / 20) train acc: 0.180000; val_acc: 0.121000
(Epoch 1 / 20) train acc: 0.240000; val_acc: 0.151000
(Epoch 2 / 20) train acc: 0.440000; val_acc: 0.153000
(Epoch 3 / 20) train acc: 0.460000; val_acc: 0.173000
(Epoch 4 / 20) train acc: 0.520000; val_acc: 0.168000
(Epoch 5 / 20) train acc: 0.700000; val_acc: 0.170000
(Iteration 11 / 40) loss: 1.097557
(Epoch 6 / 20) train acc: 0.840000; val_acc: 0.187000
(Epoch 7 / 20) train acc: 0.800000; val_acc: 0.173000
(Epoch 8 / 20) train acc: 0.880000; val_acc: 0.202000
(Epoch 9 / 20) train acc: 0.920000; val_acc: 0.192000
(Epoch 10 / 20) train acc: 0.920000; val_acc: 0.180000
(Iteration 21 / 40) loss: 0.641836
(Epoch 11 / 20) train acc: 0.980000; val_acc: 0.156000
(Epoch 12 / 20) train acc: 0.980000; val_acc: 0.176000
(Epoch 13 / 20) train acc: 0.980000; val_acc: 0.169000
(Epoch 14 / 20) train acc: 0.980000; val_acc: 0.188000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.196000
(Iteration 31 / 40) loss: 0.249410
(Epoch 16 / 20) train acc: 0.980000; val_acc: 0.193000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.189000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.192000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.193000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.187000
```

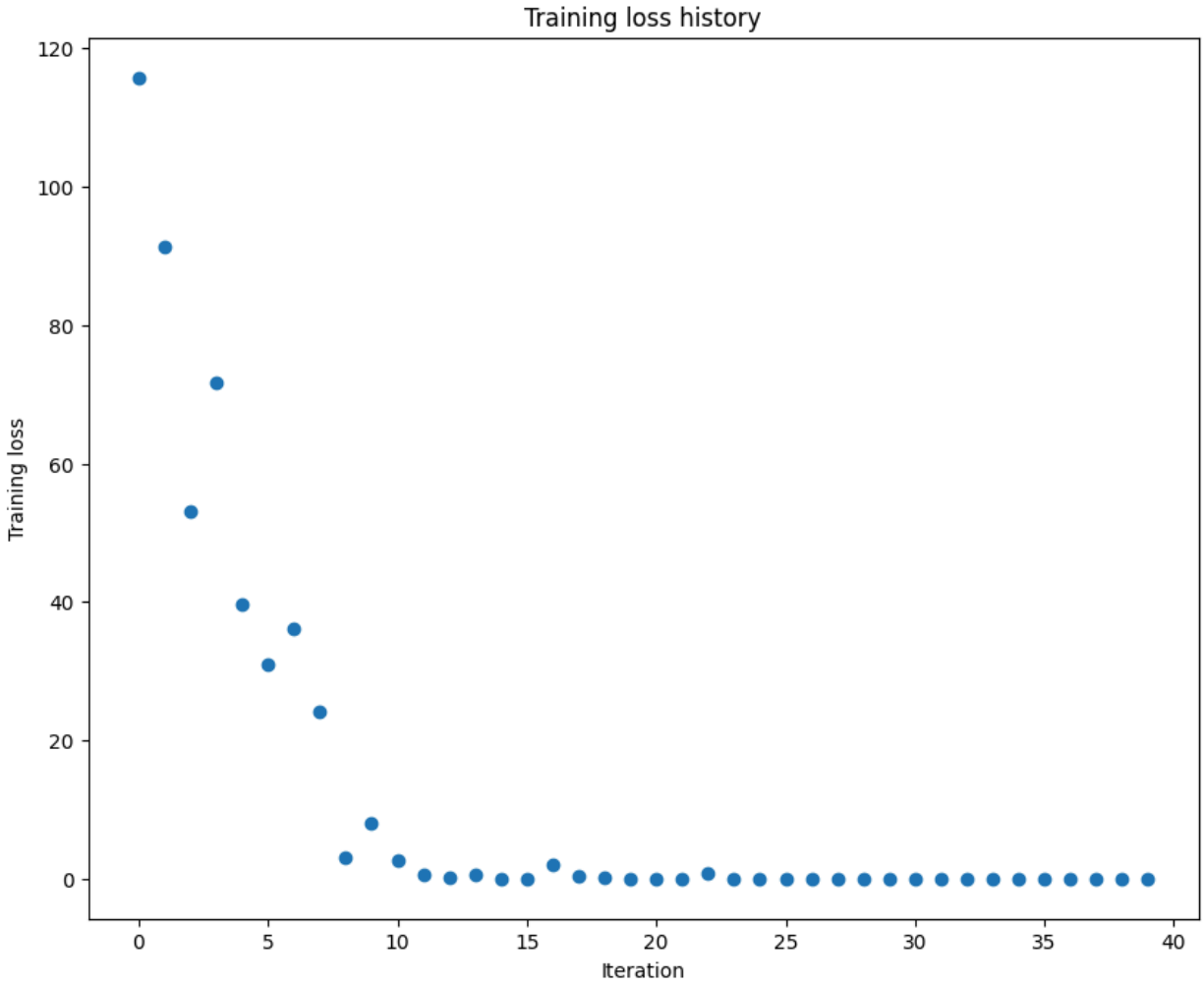


Now try to use a five-layer network with 100 units on each layer to overfit 50 training examples. Again you will have to adjust the learning rate and weight initialization, but you should be able to achieve 100% training accuracy within 20 epochs.

In [35]:
```
# TODO: Use a five-layer Net to overfit 50 training examples by
# tweaking just the learning rate and initialization scale.

num_train = 50
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'],
}

# learning_rate = 2e-3
learning_rate = 1e-3
weight_scale = 1e-1
model = FullyConnectedNet([100, 100, 100, 100],
                weight_scale=weight_scale, dtype=np.float64)
solver = Solver(model, small_data,
                print_every=10, num_epochs=20, batch_size=25,
                update_rule='sgd',
                optim_config={
                  'learning_rate': learning_rate,
                }
```

```
        )
solver.train()

plt.plot(solver.loss_history, 'o')
plt.title('Training loss history')
plt.xlabel('Iteration')
plt.ylabel('Training loss')
plt.show()
```

```
(Iteration 1 / 40) loss: 115.774329
(Epoch 0 / 20) train acc: 0.140000; val_acc: 0.112000
(Epoch 1 / 20) train acc: 0.160000; val_acc: 0.126000
(Epoch 2 / 20) train acc: 0.320000; val_acc: 0.085000
(Epoch 3 / 20) train acc: 0.360000; val_acc: 0.143000
(Epoch 4 / 20) train acc: 0.680000; val_acc: 0.132000
(Epoch 5 / 20) train acc: 0.760000; val_acc: 0.133000
(Iteration 11 / 40) loss: 2.715416
(Epoch 6 / 20) train acc: 0.880000; val_acc: 0.132000
(Epoch 7 / 20) train acc: 0.940000; val_acc: 0.145000
(Epoch 8 / 20) train acc: 0.940000; val_acc: 0.144000
(Epoch 9 / 20) train acc: 0.960000; val_acc: 0.146000
(Epoch 10 / 20) train acc: 0.980000; val_acc: 0.148000
(Iteration 21 / 40) loss: 0.000212
(Epoch 11 / 20) train acc: 0.980000; val_acc: 0.148000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.148000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.147000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.147000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.147000
(Iteration 31 / 40) loss: 0.000291
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.147000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.147000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.148000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.148000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.148000
```



Inline Question 2:

Did you notice anything about the comparative difficulty of training the three-layer net vs training the five layer net? In particular, based on your experience, which network seemed more sensitive to the initialization scale? Why do you think that is the case?

Answer:

Training a five-layer network is significantly more sensitive to weight initialization than a three-layer network due to its greater depth. In deep networks, small weight scales cause vanishing gradients, while large scales cause exploding gradients. As depth increases, the chance of encountering these issues rises, making it harder to find the right weight scale. The five-layer net was more sensetive to the weight initialization than the learning rate. Increasing the weight intialization causes exploding gradients quickly. While shallower networks also face sensitivity, their weight scale is easier to identify.

Update rules

So far we have used vanilla stochastic gradient descent (SGD) as our update rule. More sophisticated update rules can make it easier to train deep networks. We will implement a few of the most commonly used update rules and compare them to vanilla SGD.

SGD+Momentum

Stochastic gradient descent with momentum is a widely used update rule that tends to make deep networks converge faster than vanilla stochastic gradient descent.

Open the file `cs6353/optim.py` and read the documentation at the top of the file to make sure you understand the API. Implement the SGD+momentum update rule in the function `sgd_momentum` and run the following to check your implementation. You should see errors less than e-8.

In [36]:
```
from cs6353.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    ]])

# Should see relative errors around e-8 or less
print('next_w error: ', rel_error(next_w, expected_next_w))
print('velocity error: ', rel_error(expected_velocity, config['velocity']))
```

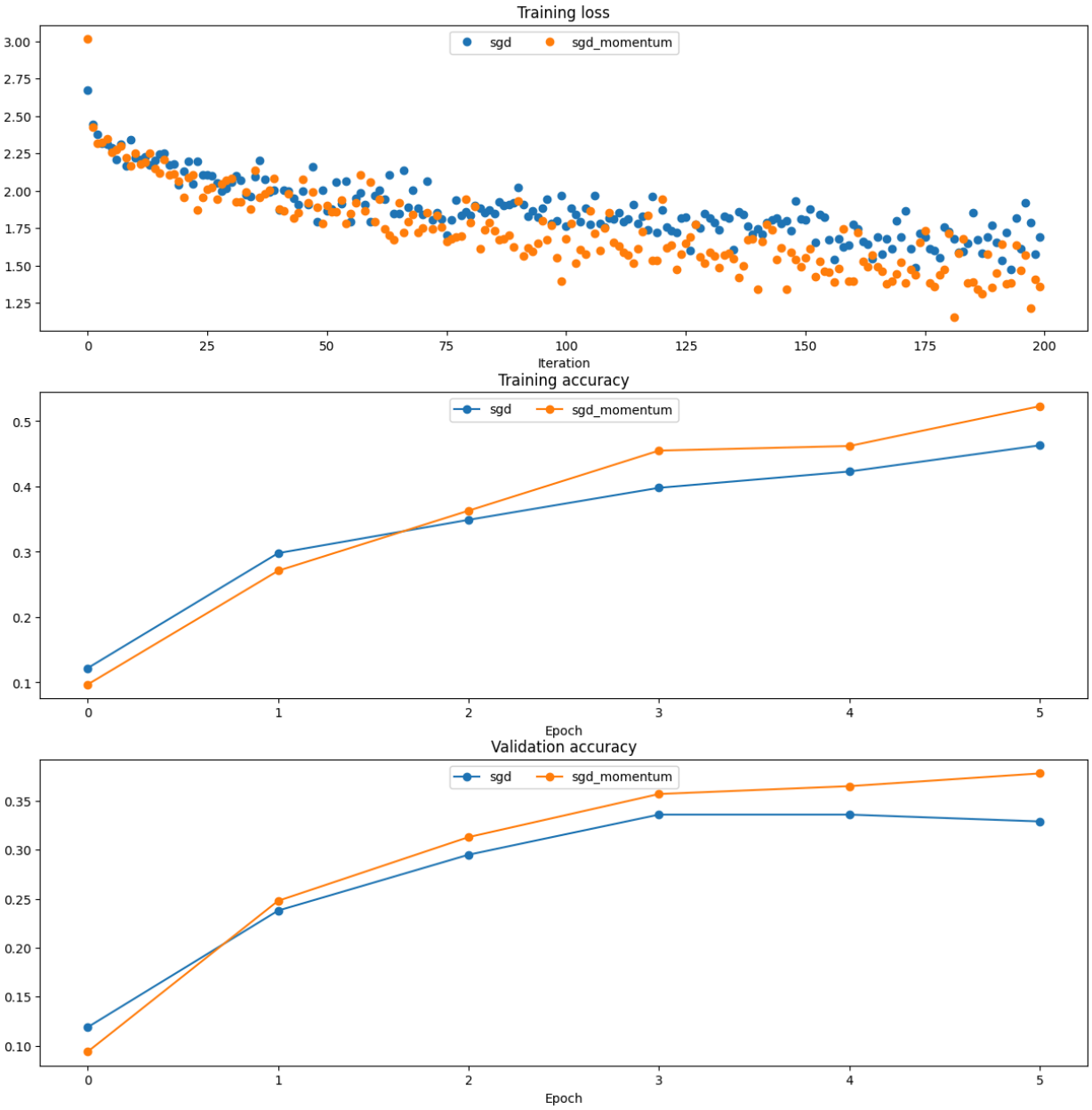
```
next_w error:  8.882347033505819e-09
velocity error:  4.269287743278663e-09
```

Once you have done so, run the following to train a six-layer network with both SGD and SGD+momentum. You should see the SGD+momentum update rule converge faster.

In [37]:
```
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']:
  print('running with ', 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=True)
  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 list(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()
```

```
running with  sgd
(Iteration 1 / 200) loss: 2.675409
(Epoch 0 / 5) train acc: 0.122000; val_acc: 0.119000
(Iteration 11 / 200) loss: 2.221112
(Iteration 21 / 200) loss: 2.130692
(Iteration 31 / 200) loss: 2.058671
(Epoch 1 / 5) train acc: 0.298000; val_acc: 0.238000
(Iteration 41 / 200) loss: 1.869888
(Iteration 51 / 200) loss: 1.866691
(Iteration 61 / 200) loss: 1.967924
(Iteration 71 / 200) loss: 1.840155
(Epoch 2 / 5) train acc: 0.349000; val_acc: 0.295000
(Iteration 81 / 200) loss: 1.832665
(Iteration 91 / 200) loss: 2.024011
(Iteration 101 / 200) loss: 1.764401
(Iteration 111 / 200) loss: 1.813284
(Epoch 3 / 5) train acc: 0.398000; val_acc: 0.336000
(Iteration 121 / 200) loss: 1.869099
(Iteration 131 / 200) loss: 1.818794
(Iteration 141 / 200) loss: 1.745466
(Iteration 151 / 200) loss: 1.803553
(Epoch 4 / 5) train acc: 0.423000; val_acc: 0.336000
(Iteration 161 / 200) loss: 1.773484
(Iteration 171 / 200) loss: 1.688030
(Iteration 181 / 200) loss: 1.725408
(Iteration 191 / 200) loss: 1.651351
(Epoch 5 / 5) train acc: 0.463000; val_acc: 0.329000

running with  sgd_momentum
(Iteration 1 / 200) loss: 3.015264
(Epoch 0 / 5) train acc: 0.097000; val_acc: 0.094000
(Iteration 11 / 200) loss: 2.253214
(Iteration 21 / 200) loss: 1.954631
(Iteration 31 / 200) loss: 2.079628
(Epoch 1 / 5) train acc: 0.271000; val_acc: 0.248000
(Iteration 41 / 200) loss: 1.876052
(Iteration 51 / 200) loss: 1.903648
(Iteration 61 / 200) loss: 1.790516
(Iteration 71 / 200) loss: 1.751532
(Epoch 2 / 5) train acc: 0.363000; val_acc: 0.313000
(Iteration 81 / 200) loss: 1.787554
(Iteration 91 / 200) loss: 1.932285
(Iteration 101 / 200) loss: 1.680905
(Iteration 111 / 200) loss: 1.655849
(Epoch 3 / 5) train acc: 0.455000; val_acc: 0.357000
(Iteration 121 / 200) loss: 1.944164
(Iteration 131 / 200) loss: 1.587576
(Iteration 141 / 200) loss: 1.340575
(Iteration 151 / 200) loss: 1.554041
(Epoch 4 / 5) train acc: 0.462000; val_acc: 0.365000
(Iteration 161 / 200) loss: 1.393865
(Iteration 171 / 200) loss: 1.520558
(Iteration 181 / 200) loss: 1.716010
(Iteration 191 / 200) loss: 1.446959
(Epoch 5 / 5) train acc: 0.523000; val_acc: 0.378000
```



Train a good model!

Train the best fully-connected model that you can on CIFAR-10, storing your best model in the `best_model` variable. We require you to get at least 50% accuracy on the validation set using a fully-connected net.

If you are careful it should be possible to get accuracies above 55%, but we don't require it for this part and won't assign extra credit for doing so. Later in the assignment we will ask you to train the best convolutional network that you can on CIFAR-10, and we would prefer that you spend your effort working on convolutional nets rather than fully-connected nets.

You might find it useful to complete the `BatchNormalization.ipynb` notebook before completing this part, since those techniques can help you train powerful models.

In [42]:
```
best_model = None
################################################################################
# TODO: Train the best FullyConnectedNet that you can on CIFAR-10. You might   #
# find batch/layer normalization useful. Store your best model in  #
# the best_model variable.                                                     #
################################################################################
# lr_list = [0.0001, 0.0002, 0.0003, 0.0004]
# reg_list = [0.00001, 0.000015, 0.00002, 0.000025]
# layer_list = [25, 50, 100]
# epoch_list = [5, 10, 15, 20]

lr_list = [0.0001, 0.0002, 0.0003]
reg_list = [0.00001, 0.00002]
layer_list = [100]
epoch_list = [20, 25]

# lr_list = [0.0003]
# reg_list = [0.00002]
# layer_list = [100]
# epoch_list = [20]

best_val_acc = -1
for lr in lr_list:
  for regularization in reg_list:
    for layer in layer_list:
      for epoch in epoch_list:
        print('lr: {}, reg: {}, layer: {}, epoch: {}'.format(lr, regularization, layer
        model = TwoLayerNet(hidden_dim=layer, reg=regularization)
        candidate_solver = Solver(model, data, update_rule='sgd', optim_config={'learr
                                  lr_decay = 0.95, num_epochs=epoch, batch_size=100, v
        candidate_solver.train()
        if candidate_solver.best_val_acc > best_val_acc:
          best_val_acc = candidate_solver.best_val_acc
          best_model = model
          print('lr: {}, reg: {}, layer: {}, epoch: {}, Val acc: {}'.format(lr, regula

print('Best val acc: {}'.format(best_val_acc))
################################################################################
#                              END OF YOUR CODE                                #
################################################################################
```

```
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 20, Val acc: 0.482
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 25
lr: 0.0001, reg: 1e-05, layer: 100, epoch: 25, Val acc: 0.5
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 25
lr: 0.0001, reg: 2e-05, layer: 100, epoch: 25, Val acc: 0.504
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 20, Val acc: 0.525
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 25
lr: 0.0002, reg: 1e-05, layer: 100, epoch: 25, Val acc: 0.533
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0002, reg: 2e-05, layer: 100, epoch: 25
lr: 0.0003, reg: 1e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 1e-05, layer: 100, epoch: 25
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 20
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 25
lr: 0.0003, reg: 2e-05, layer: 100, epoch: 25, Val acc: 0.537
Best val acc: 0.537
```

Test your model!

Run your best model on the validation and test sets. You should achieve above 50% accuracy on the validation set.

In [43]:
```
y_test_pred = np.argmax(best_model.loss(data['X_test']), axis=1)
y_val_pred = np.argmax(best_model.loss(data['X_val']), axis=1)
print('Validation set accuracy: ', (y_val_pred == data['y_val']).mean())
print('Test set accuracy: ', (y_test_pred == data['y_test']).mean())
```

```
Validation set accuracy:  0.537
Test set accuracy:  0.527
```

In [112…

```
# This mounts your Google Drive to the Colab VM.
from google.colab import drive
drive.mount('/content/drive')

# TODO: Enter the foldername in your Drive where you have saved the unzipped
# assignment folder, e.g. 'cs6353/assignments/assignment3/'
FOLDERNAME = 'CS6353/Assignments/assignment3/assignment3/'
assert FOLDERNAME is not None, "[!] Enter the foldername."

# Now that we've mounted your Drive, this ensures that
# the Python interpreter of the Colab VM can load
# python files from within it.
import sys
sys.path.append('/content/drive/My Drive/{}'.format(FOLDERNAME))

# This downloads the CIFAR-10 dataset to your Drive
# if it doesn't already exist.

%cd /content/drive/My\ Drive/$FOLDERNAME/cs6353/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME

# Install requirements from colab_requirements.txt
# TODO: Please change your path below to the colab_requirements.txt file
! python -m pip install -r /content/drive/My\ Drive/$FOLDERNAME/requirements.txt
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.m
ount("/content/drive", force_remount=True).
/content/drive/My Drive/CS6353/Assignments/assignment3/assignment3/cs6353/datasets
--2024-10-19 01:38:07--  https://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz
Resolving www.cs.toronto.edu (www.cs.toronto.edu)... 128.100.3.30
Connecting to www.cs.toronto.edu (www.cs.toronto.edu)|128.100.3.30|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 170498071 (163M) [application/x-gzip]
Saving to: ‘cifar-10-python.tar.gz’

cifar-10-python.tar 100%[===================>] 162.60M  43.1MB/s    in 4.2s

2024-10-19 01:38:11 (38.8 MB/s) - ‘cifar-10-python.tar.gz’ saved [170498071/17049807
1]

cifar-10-batches-py/
cifar-10-batches-py/data_batch_4
cifar-10-batches-py/readme.html
cifar-10-batches-py/test_batch
cifar-10-batches-py/data_batch_3
cifar-10-batches-py/batches.meta
cifar-10-batches-py/data_batch_2
cifar-10-batches-py/data_batch_5
cifar-10-batches-py/data_batch_1
/content/drive/My Drive/CS6353/Assignments/assignment3/assignment3
Collecting attrs==19.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignme
nt3/assignment3//requirements.txt (line 1))
  Using cached attrs-19.1.0-py2.py3-none-any.whl.metadata (10 kB)
Collecting backcall==0.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assign
ment3/assignment3//requirements.txt (line 2))
  Using cached backcall-0.1.0.zip (11 kB)
  Preparing metadata (setup.py) ... done
Collecting bleach==3.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignme
nt3/assignment3//requirements.txt (line 3))
  Using cached bleach-3.1.0-py2.py3-none-any.whl.metadata (19 kB)
Collecting certifi==2019.6.16 (from -r /content/drive/My Drive/CS6353/Assignments/ass
ignment3/assignment3//requirements.txt (line 4))
  Using cached certifi-2019.6.16-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting cycler==0.10.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 5))
  Using cached cycler-0.10.0-py2.py3-none-any.whl.metadata (722 bytes)
Collecting decorator==4.4.0 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 6))
  Using cached decorator-4.4.0-py2.py3-none-any.whl.metadata (3.7 kB)
Collecting defusedxml==0.6.0 (from -r /content/drive/My Drive/CS6353/Assignments/assi
gnment3/assignment3//requirements.txt (line 7))
  Using cached defusedxml-0.6.0-py2.py3-none-any.whl.metadata (31 kB)
Collecting entrypoints==0.3 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 8))
  Using cached entrypoints-0.3-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting future==0.17.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 9))
  Using cached future-0.17.1.tar.gz (829 kB)
  Preparing metadata (setup.py) ... done
Collecting imageio==2.5.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignm
ent3/assignment3//requirements.txt (line 10))
  Using cached imageio-2.5.0-py3-none-any.whl.metadata (2.8 kB)
Collecting ipykernel==5.1.2 (from -r /content/drive/My Drive/CS6353/Assignments/assig
nment3/assignment3//requirements.txt (line 11))
  Using cached ipykernel-5.1.2-py3-none-any.whl.metadata (919 bytes)
```

```
Collecting ipython==7.8.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 12))
  Using cached ipython-7.8.0-py3-none-any.whl.metadata (4.3 kB)
Requirement already satisfied: ipython-genutils==0.2.0 in /usr/local/lib/python3.10/dist-packages (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 13)) (0.2.0)
Collecting ipywidgets==7.5.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 14))
  Using cached ipywidgets-7.5.1-py2.py3-none-any.whl.metadata (1.8 kB)
Collecting jedi==0.15.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 15))
  Using cached jedi-0.15.1-py2.py3-none-any.whl.metadata (15 kB)
Collecting Jinja2==2.10.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 16))
  Using cached Jinja2-2.10.1-py2.py3-none-any.whl.metadata (2.2 kB)
Collecting jsonschema==3.0.2 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 17))
  Using cached jsonschema-3.0.2-py2.py3-none-any.whl.metadata (7.4 kB)
Collecting jupyter==1.0.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 18))
  Using cached jupyter-1.0.0-py2.py3-none-any.whl.metadata (995 bytes)
Collecting jupyter-client==5.3.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 19))
  Using cached jupyter_client-5.3.1-py2.py3-none-any.whl.metadata (3.6 kB)
Collecting jupyter-console==6.0.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 20))
  Using cached jupyter_console-6.0.0-py2.py3-none-any.whl.metadata (955 bytes)
Collecting jupyter-core==4.5.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 21))
  Using cached jupyter_core-4.5.0-py2.py3-none-any.whl.metadata (884 bytes)
Collecting kiwisolver==1.1.0 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 22))
  Using cached kiwisolver-1.1.0.tar.gz (30 kB)
  Preparing metadata (setup.py) ... done
Collecting MarkupSafe==1.1.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 23))
  Using cached MarkupSafe-1.1.1.tar.gz (19 kB)
  Preparing metadata (setup.py) ... done
Collecting matplotlib==3.1.1 (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 24))
  Using cached matplotlib-3.1.1.tar.gz (37.8 MB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: mistune==0.8.4 in /usr/local/lib/python3.10/dist-packages (from -r /content/drive/My Drive/CS6353/Assignments/assignment3/assignment3//requirements.txt (line 25)) (0.8.4)
ERROR: Could not find a version that satisfies the requirement mkl-fft==1.0.6 (from versions: 1.3.6, 1.3.8)
ERROR: No matching distribution found for mkl-fft==1.0.6
```

Batch Normalization

One way to make deep networks easier to train is to use more sophisticated optimization procedures such as SGD+momentum, and RMSProp. Another strategy is to change the architecture of the network to make it easier to train. One idea along these lines is batch normalization which was proposed by [3] in 2015.

The idea is relatively straightforward. Machine learning methods tend to work better when their input data consists of uncorrelated features with zero mean and unit variance. When training a neural network, we can preprocess the data before feeding it to the network to explicitly decorrelate its features; this will ensure that the first layer of the network sees data that follows a nice distribution. However, even if we preprocess the input data, the activations at deeper layers of the network will likely no longer be decorrelated and will no longer have zero mean or unit variance since they are output from earlier layers in the network. Even worse, during the training process the distribution of features at each layer of the network will shift as the weights of each layer are updated.

The authors of [1] hypothesize that the shifting distribution of features inside deep neural networks may make training deep networks more difficult. To overcome this problem, [3] proposes to insert batch normalization layers into the network. At training time, a batch normalization layer uses a minibatch of data to estimate the mean and standard deviation of each feature. These estimated means and standard deviations are then used to center and normalize the features of the minibatch. A running average of these means and standard deviations is kept during training, and at test time these running averages are used to center and normalize features.

It is possible that this normalization strategy could reduce the representational power of the network, since it may sometimes be optimal for certain layers to have features that are not zero-mean or unit variance. To this end, the batch normalization layer includes learnable shift and scale parameters for each feature dimension.

[1] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ICML 2015.

In [113…

```
# As usual, a bit of setup
import time
import numpy as np
import matplotlib.pyplot as plt
from cs6353.classifiers.fc_net import *
from cs6353.data_utils import get_CIFAR10_data
from cs6353.gradient_check import eval_numerical_gradient, eval_numerical_gradient_arr
from cs6353.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'

%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))))

def print_mean_std(x,axis=0):
    print('  means: ', x.mean(axis=axis))
```

```
    print('  stds:  ', x.std(axis=axis))
    print()
```

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

In [114…

```
# Load the (preprocessed) CIFAR10 data.
data = get_CIFAR10_data()
for k, v in data.items():
  print('%s: ' % k, v.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,)
```

Batch Normalization: Forward

In the file `cs6353/layers.py`, implement the batch normalization forward pass in the function `batchnorm_forward`. Once you have done so, run the following to test your implementation.

Referencing the paper linked to above would be helpful!

In [115…

```
# 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
np.random.seed(231)
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_mean_std(a,axis=0)

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

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

```
Before batch normalization:
  means:  [ -2.3814598  -13.18038246   1.91780462]
  stds:   [27.18502186 34.21455511 37.68611762]

After batch normalization (gamma=1, beta=0)
  means:  [5.99520433e-17 7.16093851e-17 8.32667268e-19]
  stds:   [0.99999999 1.         1.        ]

After batch normalization (gamma= [1. 2. 3.] , beta= [11. 12. 13.] )
  means:  [11. 12. 13.]
  stds:   [0.99999999 1.99999999 2.99999999]
```

In [116…

```
# 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.

np.random.seed(231)
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 range(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_mean_std(a_norm,axis=0)
```

```
After batch normalization (test-time):
  means:  [-0.03927354 -0.04349152 -0.10452688]
  stds:   [1.01531428 1.01238373 0.97819988]
```

Batch normalization: Backward Pass

Now implement the backward pass for batch normalization in the function `batchnorm_backward`.

To derive the backward pass you should write out the computation graph for batch normalization and backprop through each of the intermediate nodes. Some intermediates may have multiple outgoing branches; make sure to sum gradients across these branches in the backward pass.

Once you have finished, run the following to numerically check your backward pass.

In [117…

```
# Gradient check batchnorm backward pass
np.random.seed(231)
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, a, beta, bn_param)[0]
fb = lambda b: batchnorm_forward(x, gamma, b, bn_param)[0]

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

_, cache = batchnorm_forward(x, gamma, beta, bn_param)
dx, dgamma, dbeta = batchnorm_backward(dout, cache)

#You should expect to see relative errors between 1e-13 and 1e-8
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:  1.6674604875341426e-09
dgamma error:  7.417225040694815e-13
dbeta error:  2.379446949959628e-12
```

Batch Normalization: Alternative Backward

In class we talked about two different implementations for the sigmoid backward pass. One strategy is to write out a computation graph composed of simple operations and backprop through all intermediate values. Another strategy is to work out the derivatives on paper. For example, you can derive a very simple formula for the sigmoid function's backward pass by simplifying gradients on paper.

Surprisingly, it turns out that you can do a similar simplification for the batch normalization backward pass too.

Given a set of inputs $X = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix}$, we first calculate the mean $\mu = \frac{1}{N}\sum_{k=1}^{N} x_k$ and variance $v = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$.
With μ and v calculated, we can calculate the standard deviation $\sigma = \sqrt{v + \epsilon}$ and normalized data Y with $y_i = \frac{x_i - \mu}{\sigma}$.

The meat of our problem is to get $\frac{\partial L}{\partial X}$ from the upstream gradient $\frac{\partial L}{\partial Y}$. It might be challenging to directly reason about the gradients over X and Y - try reasoning about it in terms of x_i and

y_i first.

You will need to come up with the derivations for $\frac{\partial L}{\partial x_i}$, by relying on the Chain Rule to first calculate the intermediate $\frac{\partial \mu}{\partial x_i}$, $\frac{\partial v}{\partial x_i}$, $\frac{\partial \sigma}{\partial x_i}$, then assemble these pieces to calculate $\frac{\partial y_i}{\partial x_i}$. You should make sure each of the intermediary steps are all as simple as possible.

After doing so, implement the simplified batch normalization backward pass in the function `batchnorm_backward_alt` and compare the two implementations by running the following. Your two implementations should compute nearly identical results, but the alternative implementation should be a bit faster.

In [118…

```
np.random.seed(231)
N, D = 100, 500
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'}
out, cache = batchnorm_forward(x, gamma, beta, bn_param)

t1 = time.time()
dx1, dgamma1, dbeta1 = batchnorm_backward(dout, cache)
t2 = time.time()
dx2, dgamma2, dbeta2 = batchnorm_backward_alt(dout, cache)
t3 = time.time()

print('dx difference: ', rel_error(dx1, dx2))
print('dgamma difference: ', rel_error(dgamma1, dgamma2))
print('dbeta difference: ', rel_error(dbeta1, dbeta2))
print('speedup: %.2fx' % ((t2 - t1) / (t3 - t2)))
```

```
dx difference:  9.890497291190823e-13
dgamma difference:  0.0
dbeta difference:  0.0
speedup: 0.21x
```

Fully Connected Networks with Batch Normalization

Now that you have a working implementation for batch normalization, go back to your `FullyConnectedNet` in the file `cs6353/classifiers/fc_net.py`. Modify your implementation to add batch normalization.

Concretely, when the `normalization` flag is set to `"batchnorm"` in the constructor, you should insert a batch normalization layer before each ReLU nonlinearity. The outputs from the last layer of the network should not be normalized. Once you are done, run the following to gradient-check your implementation.

HINT: You might find it useful to define an additional helper layer similar to those in the file `cs6353/layer_utils.py`.

In [119…

```
np.random.seed(231)
N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))

# You should expect losses between 1e-4~1e-10 for W,
# losses between 1e-08~1e-10 for b,
# and losses between 1e-08~1e-09 for beta and gammas.
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,
                            normalization='batchnorm')
  # model = FullyConnectedNet([H1, H2], input_dim=D, num_classes=C,
  #                           reg=reg, weight_scale=5e-2, dtype=np.float64,
  #                           normalization=None)

  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('%s relative error: %.2e' % (name, rel_error(grad_num, grads[name])))
  if reg == 0: print()
```

```
Running check with reg =  0
Initial loss:  2.2611955101340957
W1 relative error: 1.10e-04
W2 relative error: 3.11e-06
W3 relative error: 4.05e-10
b1 relative error: 2.66e-07
b2 relative error: 2.72e-07
b3 relative error: 1.01e-10
beta1 relative error: 7.33e-09
beta2 relative error: 1.89e-09
gamma1 relative error: 6.96e-09
gamma2 relative error: 2.41e-09

Running check with reg =  3.14
Initial loss:  6.996533220108303
W1 relative error: 1.98e-06
W2 relative error: 2.28e-06
W3 relative error: 1.11e-08
b1 relative error: 1.38e-08
b2 relative error: 7.99e-07
b3 relative error: 1.42e-10
beta1 relative error: 6.65e-09
beta2 relative error: 3.48e-09
gamma1 relative error: 6.27e-09
gamma2 relative error: 5.28e-09
```

Batch Normalization for Deep Networks

Run the following to train a six-layer network on a subset of 1000 training examples both with and without batch normalization.

In [120…

```
np.random.seed(231)
# 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
weight_scale = 5e-2
bn_model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normalization='ba
model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normalization=None)

bn_solver = Solver(bn_model, small_data,
                num_epochs=10, batch_size=50,
                update_rule='sgd_momentum',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True,print_every=20)
bn_solver.train()

solver = Solver(model, small_data,
                num_epochs=10, batch_size=50,
                update_rule='sgd_momentum',
                optim_config={
                  'learning_rate': 1e-3,
                },
                verbose=True, print_every=20)
solver.train()
```

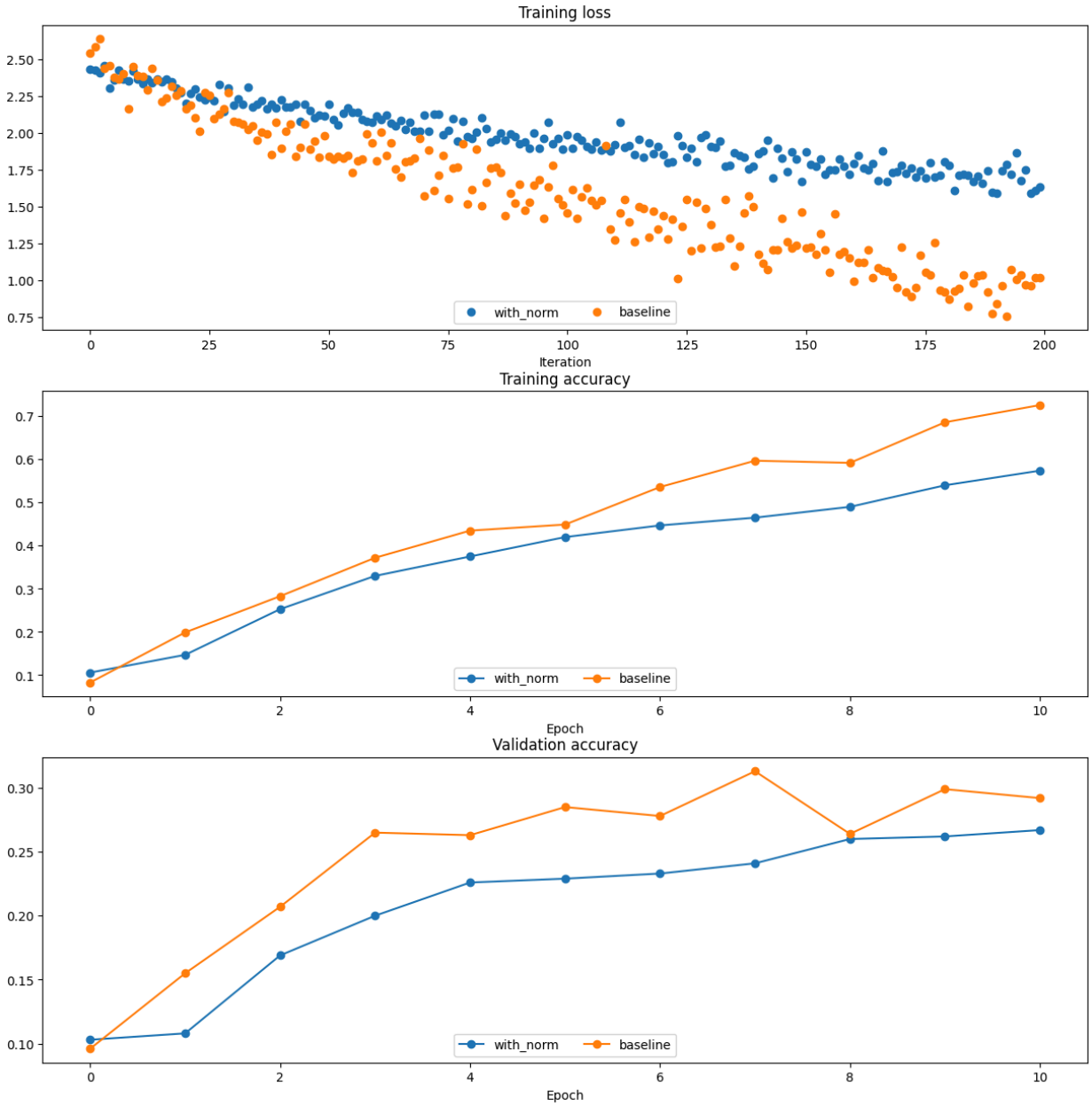
```
(Iteration 1 / 200) loss: 2.427352
(Epoch 0 / 10) train acc: 0.106000; val_acc: 0.103000
(Epoch 1 / 10) train acc: 0.147000; val_acc: 0.108000
(Iteration 21 / 200) loss: 2.200218
(Epoch 2 / 10) train acc: 0.253000; val_acc: 0.169000
(Iteration 41 / 200) loss: 2.223512
(Epoch 3 / 10) train acc: 0.330000; val_acc: 0.200000
(Iteration 61 / 200) loss: 2.112095
(Epoch 4 / 10) train acc: 0.375000; val_acc: 0.226000
(Iteration 81 / 200) loss: 1.958597
(Epoch 5 / 10) train acc: 0.420000; val_acc: 0.229000
(Iteration 101 / 200) loss: 1.985634
(Epoch 6 / 10) train acc: 0.447000; val_acc: 0.233000
(Iteration 121 / 200) loss: 1.849872
(Epoch 7 / 10) train acc: 0.465000; val_acc: 0.241000
(Iteration 141 / 200) loss: 1.859622
(Epoch 8 / 10) train acc: 0.490000; val_acc: 0.260000
(Iteration 161 / 200) loss: 1.790350
(Epoch 9 / 10) train acc: 0.540000; val_acc: 0.262000
(Iteration 181 / 200) loss: 1.778871
(Epoch 10 / 10) train acc: 0.574000; val_acc: 0.267000
(Iteration 1 / 200) loss: 2.539770
(Epoch 0 / 10) train acc: 0.083000; val_acc: 0.096000
(Epoch 1 / 10) train acc: 0.199000; val_acc: 0.155000
(Iteration 21 / 200) loss: 2.162250
(Epoch 2 / 10) train acc: 0.283000; val_acc: 0.207000
(Iteration 41 / 200) loss: 1.894397
(Epoch 3 / 10) train acc: 0.372000; val_acc: 0.265000
(Iteration 61 / 200) loss: 1.807557
(Epoch 4 / 10) train acc: 0.435000; val_acc: 0.263000
(Iteration 81 / 200) loss: 1.614385
(Epoch 5 / 10) train acc: 0.449000; val_acc: 0.285000
(Iteration 101 / 200) loss: 1.456777
(Epoch 6 / 10) train acc: 0.536000; val_acc: 0.278000
(Iteration 121 / 200) loss: 1.438367
(Epoch 7 / 10) train acc: 0.597000; val_acc: 0.313000
(Iteration 141 / 200) loss: 1.177516
(Epoch 8 / 10) train acc: 0.592000; val_acc: 0.264000
(Iteration 161 / 200) loss: 0.994768
(Epoch 9 / 10) train acc: 0.686000; val_acc: 0.299000
(Iteration 181 / 200) loss: 0.871267
(Epoch 10 / 10) train acc: 0.726000; val_acc: 0.292000
```

Run the following to visualize the results from two networks trained above. You should find that using batch normalization helps the network to converge much faster.

In [121…

```
def plot_training_history(title, label, baseline, bn_solvers, plot_fn, bl_marker='.',
    """utility function for plotting training history"""
    plt.title(title)
    plt.xlabel(label)
    bn_plots = [plot_fn(bn_solver) for bn_solver in bn_solvers]
    bl_plot = plot_fn(baseline)
    num_bn = len(bn_plots)
    for i in range(num_bn):
        label='with_norm'
        if labels is not None:
            label += str(labels[i])
        plt.plot(bn_plots[i], bn_marker, label=label)
    label='baseline'
```

```
    if labels is not None:
        label += str(labels[0])
    plt.plot(bl_plot, bl_marker, label=label)
    plt.legend(loc='lower center', ncol=num_bn+1)


plt.subplot(3, 1, 1)
plot_training_history('Training loss','Iteration', solver, [bn_solver], \
                      lambda x: x.loss_history, bl_marker='o', bn_marker='o')
plt.subplot(3, 1, 2)
plot_training_history('Training accuracy','Epoch', solver, [bn_solver], \
                      lambda x: x.train_acc_history, bl_marker='-o', bn_marker='-o')
plt.subplot(3, 1, 3)
plot_training_history('Validation accuracy','Epoch', solver, [bn_solver], \
                      lambda x: x.val_acc_history, bl_marker='-o', bn_marker='-o')

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



Batch Normalization and Initialization

We will now run a small experiment to study the interaction of batch normalization and weight initialization.

The first cell will train 8-layer networks both with and without batch normalization using different scales for weight initialization. The second layer will plot training accuracy, validation set accuracy, and training loss as a function of the weight initialization scale.

In [122…

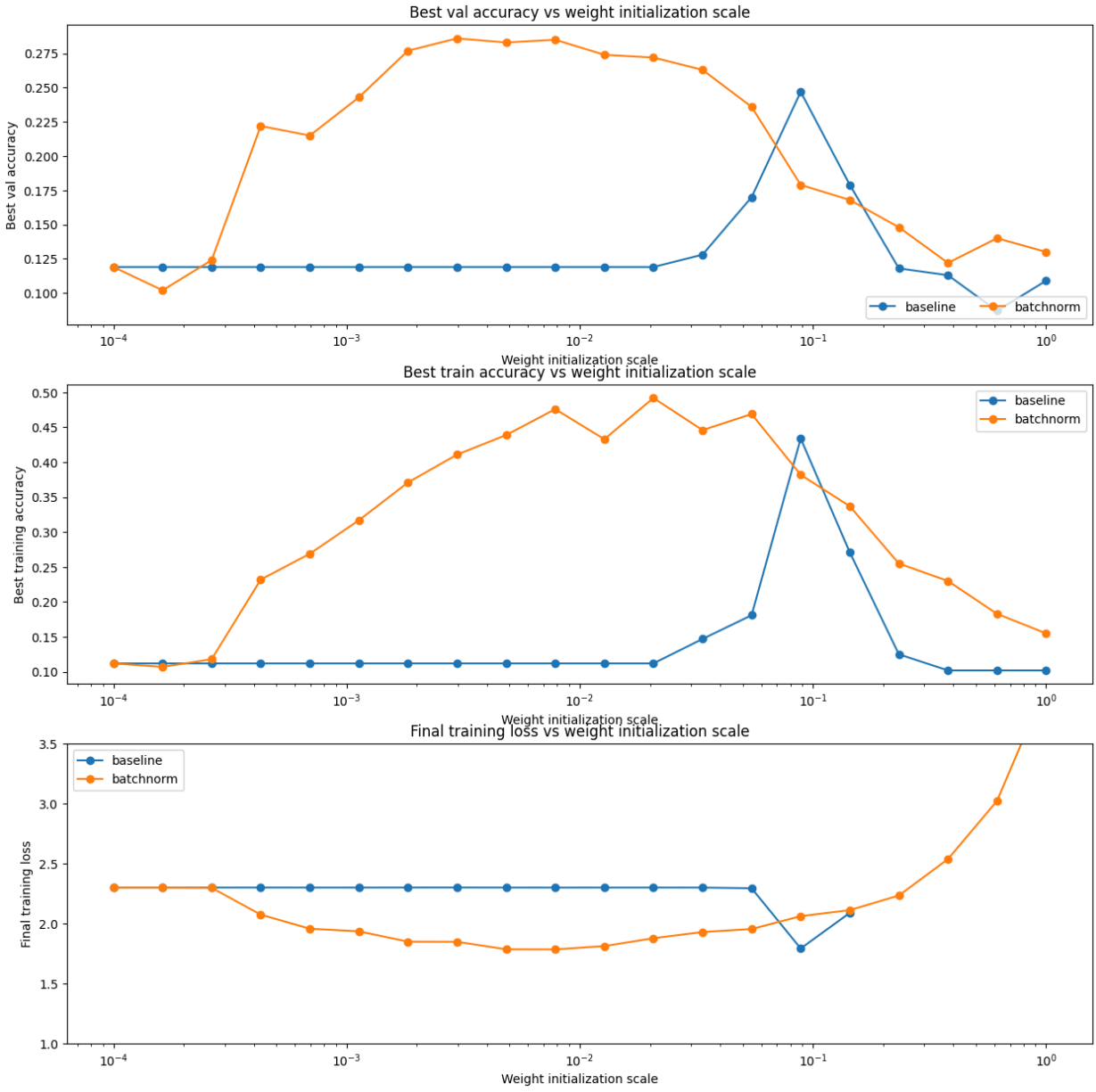
```
np.random.seed(231)
# Try training a very deep net with batchnorm
hidden_dims = [50, 50, 50, 50, 50, 50, 50]
num_train = 1000
small_data = {
  'X_train': data['X_train'][:num_train],
  'y_train': data['y_train'][:num_train],
  'X_val': data['X_val'],
  'y_val': data['y_val'],
}

bn_solvers_ws = {}
solvers_ws = {}
weight_scales = np.logspace(-4, 0, num=20)
for i, weight_scale in enumerate(weight_scales):
  print('Running weight scale %d / %d' % (i + 1, len(weight_scales)))
  bn_model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normalization='
  model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normalization=None

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

  solver = Solver(model, small_data,
                  num_epochs=10, batch_size=50,
                  update_rule='sgd_momentum',
                  optim_config={
                    'learning_rate': 1e-3,
                  },
                  verbose=False, print_every=200)
  solver.train()
  solvers_ws[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
Running weight scale 17 / 20
/content/drive/MyDrive/CS6353/Assignments/assignment3/assignment3/cs6353/layers.py:472: RuntimeWarning: invalid value encountered in subtract
Running weight scale 18 / 20
Running weight scale 19 / 20
Running weight scale 20 / 20
```

In [123…

```
# 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[ws].train_acc_history))
  bn_best_train_accs.append(max(bn_solvers_ws[ws].train_acc_history))

  best_val_accs.append(max(solvers_ws[ws].val_acc_history))
  bn_best_val_accs.append(max(bn_solvers_ws[ws].val_acc_history))

  final_train_loss.append(np.mean(solvers_ws[ws].loss_history[-100:]))
  bn_final_train_loss.append(np.mean(bn_solvers_ws[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.gca().set_ylim(1.0, 3.5)

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



Inline Question 1:

Describe the results of this experiment. How does the scale of weight initialization affect models with/without batch normalization differently, and why?

Answer:

In the three plots, batch normalization (BN) shows clear advantages over the baseline model at different weight initialization scales:

1. Best Validation Accuracy Plot: The baseline model's validation accuracy remains flat and low across all scales, indicating poor generalization, while the BN model achieves higher accuracy consistently, showing resilience to initialization scale.

2. Best Training Accuracy Plot: The baseline model only improves at very high weight scales, but performance quickly deteriorates due to exploding gradients. In contrast, BN maintains high accuracy across a broad range of scales, indicating more stable learning.

3. Final Training Loss Plot: The baseline model shows a high loss across most scales, only improving briefly before diverging. The BN model, however, has consistently lower training loss, reflecting better convergence and mitigating the exploding gradient problem seen at large scales in the baseline.

Thus, BN reduces both vanishing and exploding gradients, leading to better performance overall.

Batch normalization and batch size

We will now run a small experiment to study the interaction of batch normalization and batch size.

The first cell will train 6-layer networks both with and without batch normalization using different batch sizes. The second layer will plot training accuracy and validation set accuracy over time.

In [124…

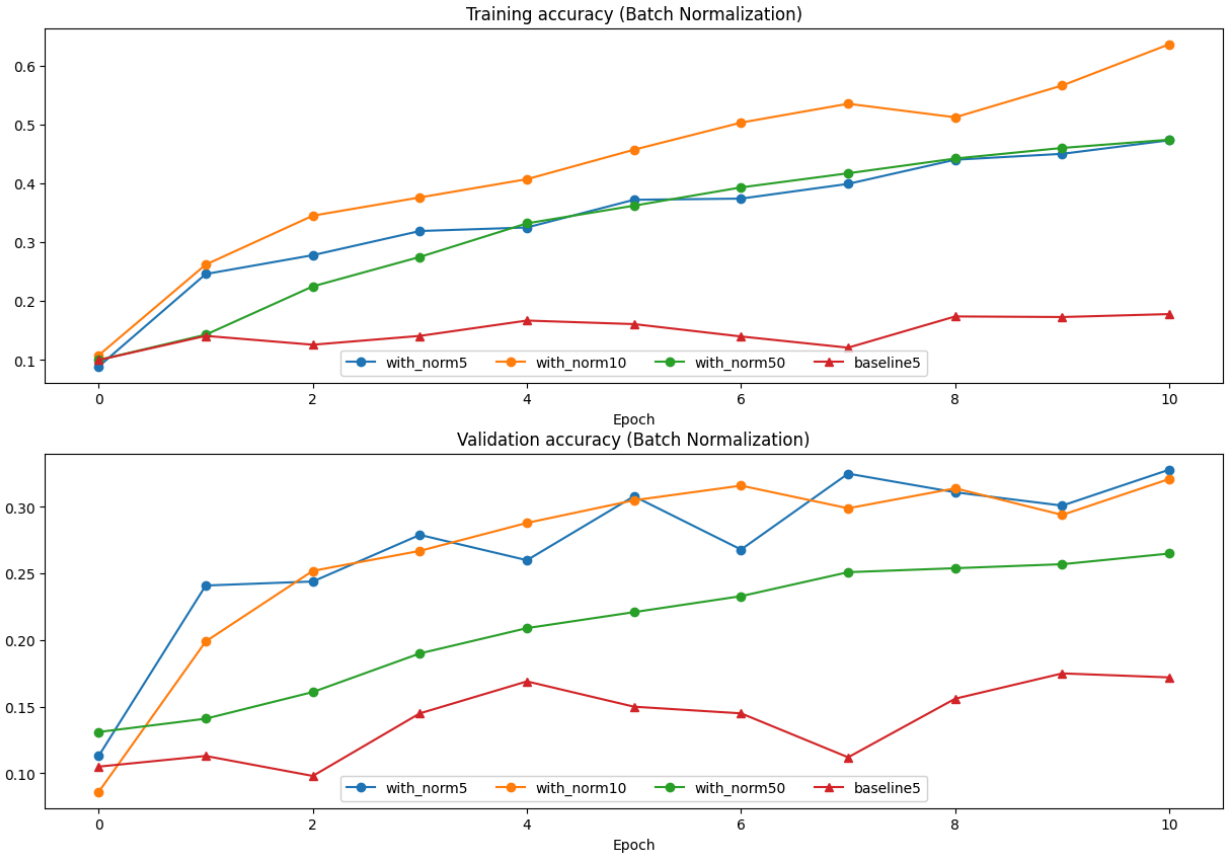
```
def run_batchsize_experiments(normalization_mode):
    np.random.seed(231)
    # Try training a very deep net with batchnorm
    hidden_dims = [100, 100, 100, 100, 100]
    num_train = 1000
    small_data = {
      'X_train': data['X_train'][:num_train],
      'y_train': data['y_train'][:num_train],
      'X_val': data['X_val'],
      'y_val': data['y_val'],
    }
    n_epochs=10
    weight_scale = 2e-2
    batch_sizes = [5,10,50]
    lr = 10**(-3.5)
    solver_bsize = batch_sizes[0]

    print('No normalization: batch size = ',solver_bsize)
    model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normalization=No
    solver = Solver(model, small_data,
                    num_epochs=n_epochs, batch_size=solver_bsize,
                    update_rule='sgd_momentum',
                    optim_config={
                      'learning_rate': lr,
                    },
                    verbose=False)
```

```
    solver.train()

    bn_solvers = []
    for i in range(len(batch_sizes)):
        b_size=batch_sizes[i]
        print('Normalization: batch size = ',b_size)
        bn_model = FullyConnectedNet(hidden_dims, weight_scale=weight_scale, normaliza
        bn_solver = Solver(bn_model, small_data,
                        num_epochs=n_epochs, batch_size=b_size,
                        update_rule='sgd_momentum',
                        optim_config={
                          'learning_rate': lr,
                        },
                        verbose=False)
        bn_solver.train()
        bn_solvers.append(bn_solver)

    return bn_solvers, solver, batch_sizes

batch_sizes = [5,10,50]
bn_solvers_bsize, solver_bsize, batch_sizes = run_batchsize_experiments('batchnorm')
```

```
No normalization: batch size =  5
Normalization: batch size =  5
Normalization: batch size =  10
Normalization: batch size =  50
```

In [125…

```
plt.subplot(2, 1, 1)
plot_training_history('Training accuracy (Batch Normalization)','Epoch', solver_bsize,
                      lambda x: x.train_acc_history, bl_marker='-^', bn_marker='-o', l
plt.subplot(2, 1, 2)
plot_training_history('Validation accuracy (Batch Normalization)','Epoch', solver_bsiz
                      lambda x: x.val_acc_history, bl_marker='-^', bn_marker='-o', lab

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



Inline Question 2:

Describe the results of this experiment. What does this imply about the relationship between batch normalization and batch size? Why is this relationship observed?

Answer:

Training Accuracy:

With_norm10 consistently achieves the highest training accuracy throughout the epochs, showing faster and better learning compared to other batch sizes.

With_norm5 performs moderately well, although not as strong as with_norm10, but better than with_norm50.

With_norm50 shows slower improvement in training accuracy, with a flatter curve, implying less effective learning initially.

Baseline5 , which does not use batch normalization, has the worst performance. It almost stagnates after the first epoch, showing minimal improvement across the epochs.

Validation Accuracy:

With_norm5 and with_norm10 show competitive performance on the validation set, with both achieving similar peaks but experiencing some fluctuations.

With_norm50 demonstrates stable, steady improvement without much fluctuation, although it reaches a lower peak compared to with_norm5 and with_norm10.

Baseline5 again performs the worst, showing very minimal improvement in validation accuracy.

Conclusion:

Batch normalization is more effective with smaller batch sizes (e.g., 5 and 10), leading to faster learning but some fluctuation in accuracy. Larger batch sizes (e.g., 50) provide more stable learning with smoother accuracy curves but slower convergence. Without batch normalization, performance is poor. The relationship arises because smaller batches yield noisier batch statistics, speeding up learning but causing instability, while larger batches offer more reliable estimates but slower adaptation.

Layer Normalization

Batch normalization has proved to be effective in making networks easier to train, but the dependency on batch size makes it less useful in complex networks which have a cap on the input batch size due to hardware limitations.

Several alternatives to batch normalization have been proposed to mitigate this problem; one such technique is Layer Normalization [2]. Instead of normalizing over the batch, we normalize over the features. In other words, when using Layer Normalization, each feature vector corresponding to a single datapoint is normalized based on the sum of all terms within that feature vector.

[2] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer Normalization." stat 1050 (2016): 21.

Inline Question 3:

Which of these data preprocessing steps is analogous to batch normalization, and which is analogous to layer normalization?

1. Scaling each image in the dataset, so that the RGB channels for each row of pixels within an image sums up to 1.
2. Scaling each image in the dataset, so that the RGB channels for all pixels within an image sums up to 1.
3. Subtracting the mean image of the dataset from each image in the dataset.
4. Setting all RGB values to either 0 or 1 depending on a given threshold.

Answer:

1. The second approach in the question is analogous to layer normalization. In layer normalization, we scale across the features (or dimensions) of each data point. In this case, that means scaling within each image independently, ensuring that the RGB channels across all pixels in an image sum to 1. There's no shifting of the data, just scaling.

2. The third approach in the question is analogous to batch normalization. In batch normalization, we shift and scale based on the statistics computed over a batch of examples. Similarly, subtracting the mean image (calculated from all images) adjusts each image based on the overall dataset.

Layer Normalization: Implementation

Now you'll implement layer normalization. This step should be relatively straightforward, as conceptually the implementation is almost identical to that of batch normalization. One significant difference though is that for layer normalization, we do not keep track of the moving moments, and the testing phase is identical to the training phase, where the mean and variance are directly calculated per datapoint.

Here's what you need to do:

- In `cs6353/layers.py`, implement the forward pass for layer normalization in the function `layernorm_backward`.

Run the cell below to check your results.

- In `cs6353/layers.py`, implement the backward pass for layer normalization in the function `layernorm_backward`.

Run the second cell below to check your results.

- Modify `cs6353/classifiers/fc_net.py` to add layer normalization to the `FullyConnectedNet`. When the `normalization` flag is set to `"layernorm"` in the constructor, you should insert a layer normalization layer before each ReLU nonlinearity.

Run the third cell below to run the batch size experiment on layer normalization.

In [126…

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

# Simulate the forward pass for a two-layer network
np.random.seed(231)
N, D1, D2, D3 =4, 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 layer normalization:')
print_mean_std(a,axis=1)
```

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

gamma = np.asarray([3.0,3.0,3.0])
beta = np.asarray([5.0,5.0,5.0])
# Now means should be close to beta and stds close to gamma
print('After layer normalization (gamma=', gamma, ', beta=', beta, ')')
a_norm, _ = layernorm_forward(a, gamma, beta, {'mode': 'train'})
print_mean_std(a_norm,axis=1)
```

```
Before layer normalization:
  means:  [-59.06673243 -47.60782686 -43.31137368 -26.40991744]
  stds:   [10.07429373 28.39478981 35.28360729  4.01831507]

After layer normalization (gamma=1, beta=0)
  means:  [ 4.81096644e-16 -7.40148683e-17  2.22044605e-16 -5.92118946e-16]
  stds:   [0.99999995 0.99999999 1.         0.99999969]

After layer normalization (gamma= [3. 3. 3.] , beta= [5. 5. 5.] )
  means:  [5. 5. 5. 5.]
  stds:   [2.99999985 2.99999998 2.99999999 2.99999907]
```

In [127…

```
# Gradient check batchnorm backward pass
np.random.seed(231)
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)

ln_param = {}
fx = lambda x: layernorm_forward(x, gamma, beta, ln_param)[0]
fg = lambda a: layernorm_forward(x, a, beta, ln_param)[0]
fb = lambda b: layernorm_forward(x, gamma, b, ln_param)[0]

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

_, cache = layernorm_forward(x, gamma, beta, ln_param)
dx, dgamma, dbeta = layernorm_backward(dout, cache)

#You should expect to see relative errors between 1e-12 and 1e-8
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:  2.1072766107135477e-09
dgamma error:  1.980045566295477e-12
dbeta error:  2.5842537629899423e-12
```

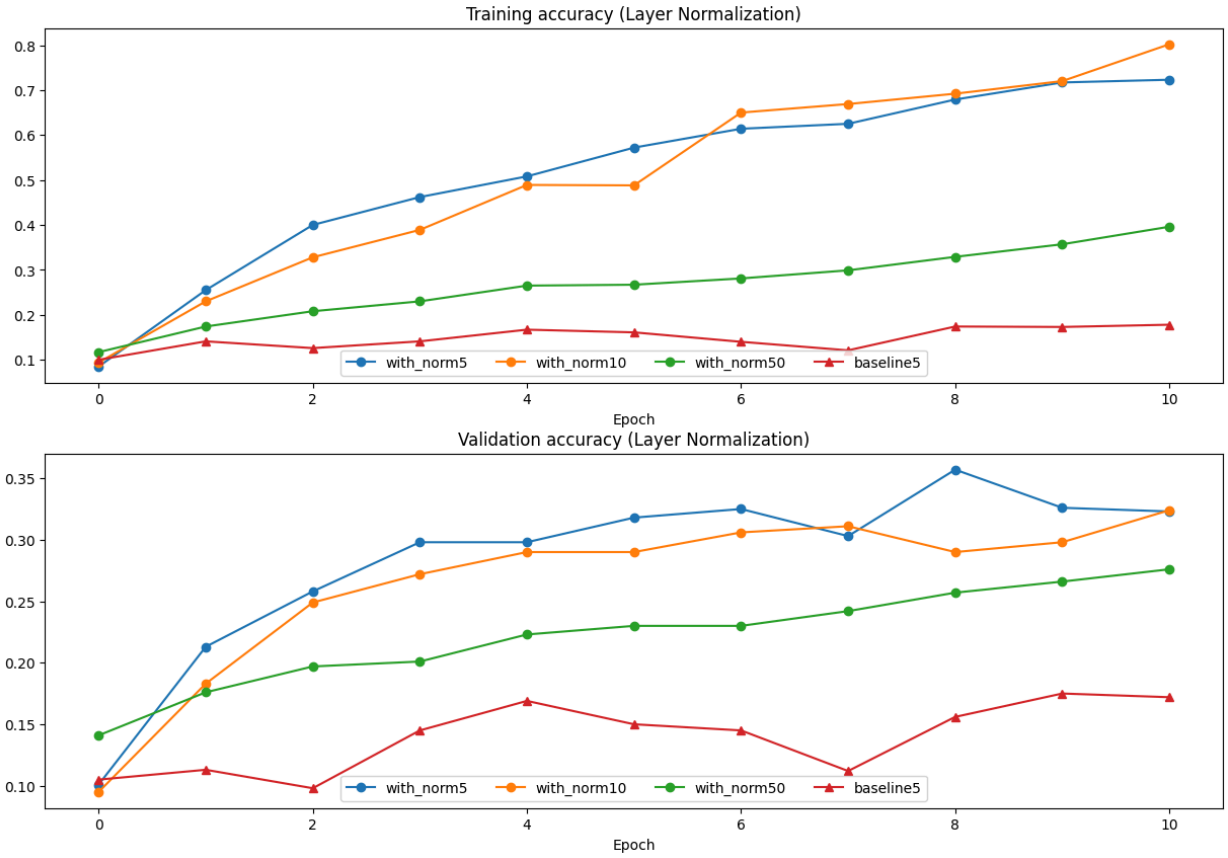
Layer Normalization and batch size

We will now run the previous batch size experiment with layer normalization instead of batch normalization. Compared to the previous experiment, you should see a markedly smaller influence of batch size on the training history!

In [128…

```
ln_solvers_bsize, solver_bsize, batch_sizes = run_batchsize_experiments('layernorm')

plt.subplot(2, 1, 1)
plot_training_history('Training accuracy (Layer Normalization)','Epoch', solver_bsize,
                      lambda x: x.train_acc_history, bl_marker='-^', bn_marker='-o', l
plt.subplot(2, 1, 2)
plot_training_history('Validation accuracy (Layer Normalization)','Epoch', solver_bsiz
                      lambda x: x.val_acc_history, bl_marker='-^', bn_marker='-o', lab

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

```
No normalization: batch size =  5
Normalization: batch size =  5
Normalization: batch size =  10
Normalization: batch size =  50
```



Inline Question 4:

When is layer normalization likely to not work well, and why?

1. Using it in a very deep network
2. Having a very small dimension of features
3. Having a high regularization term

Answer:

1. Using it in a very deep network

Will work well. Layer normalization generally works well in deep networks. For example, even in networks with several layers, layer normalization has been shown to improve performance and speed up training. So, it's not likely to struggle in deep architectures.

1. Having a very small feature dimension

May not work well. A small number of features can reduce the effectiveness of layer normalization. This is similar to the challenges batch normalization faces with small batch sizes. In layer normalization, the statistics are calculated based on the hidden units, which are influenced by the dimensionality of the data. When there are fewer features, the computed statistics can become noisy, impacting performance.

1. Using a high regularization term

May not work well. High regularization can negatively impact layer normalization. When regularization is too strong, the model tends to underfit, meaning it learns overly simple patterns, which can hinder the benefits of layer normalization. This leads to poorer overall performance.