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.

Affine layer: foward

Open the file `cs231n/layers.py` and implement the `affine_forward` function.

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

In [3]:

```
# 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:  9.769847728806635e-10
```

Affine layer: backward

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

In [4]:

```
# 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 [5]:

```
# 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 [6]:

```
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, 2

1. When the input value is negative, it will get zero gradient flow.
2. When the input value is really small or large, it will get close to zero gradient flow

"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 `cs231n/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 [7]:

```
from cs231n.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, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_relu_forward(x, w, b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_relu_forward(x, w, b)[0], b, dout)

# 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:  6.750562121603446e-11
dw error:  8.162015570444288e-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 `cs231n/layers.py`.

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

In [8]:

```
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 order of e-9
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 around e-8
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 `cs231n/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 [9]:

```
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,  16.09215096]
,
   [12.05769098,  12.74614105,  13.43459113,  14.1230412,   14.81149128, 15.49994135,  16.18839143]
,
   [12.58373087,  13.20054771,  13.81736455,  14.43418138,  15.05099822, 15.66781506,  16.2846319 ]
])
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)
```

```
Testing training loss (no regularization)
Running numeric gradient check with reg =  0.0
W1 relative error: 1.22e-08
W2 relative error: 3.48e-10
b1 relative error: 6.55e-09
b2 relative error: 4.33e-10
Running numeric gradient check with reg =  0.7
W1 relative error: 8.18e-07
W2 relative error: 2.85e-08
b1 relative error: 1.09e-09
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 `cs231n/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 [10]:

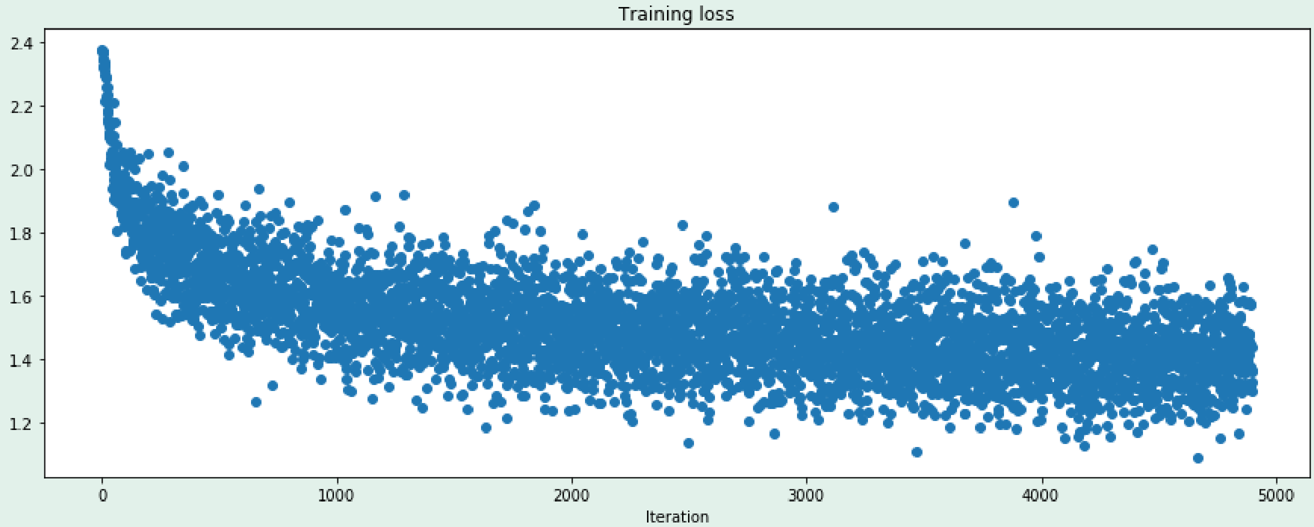
```
model = TwoLayerNet()
solver = None

##############################################################################
# TODO: Use a Solver instance to train a TwoLayerNet that achieves at least  #
# 50% accuracy on the validation set.                                        #
##############################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

model = TwoLayerNet(reg = 0.5)

solver = Solver(model, data,
                    update_rule='sgd',
                    optim_config={
                      'learning_rate': 1e-3,
                    },
                    lr_decay=0.95,
                    num_epochs=10, batch_size=100,
                    print_every=100)

solver.train()

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
##############################################################################
#                              END OF YOUR CODE                              #
##############################################################################
```

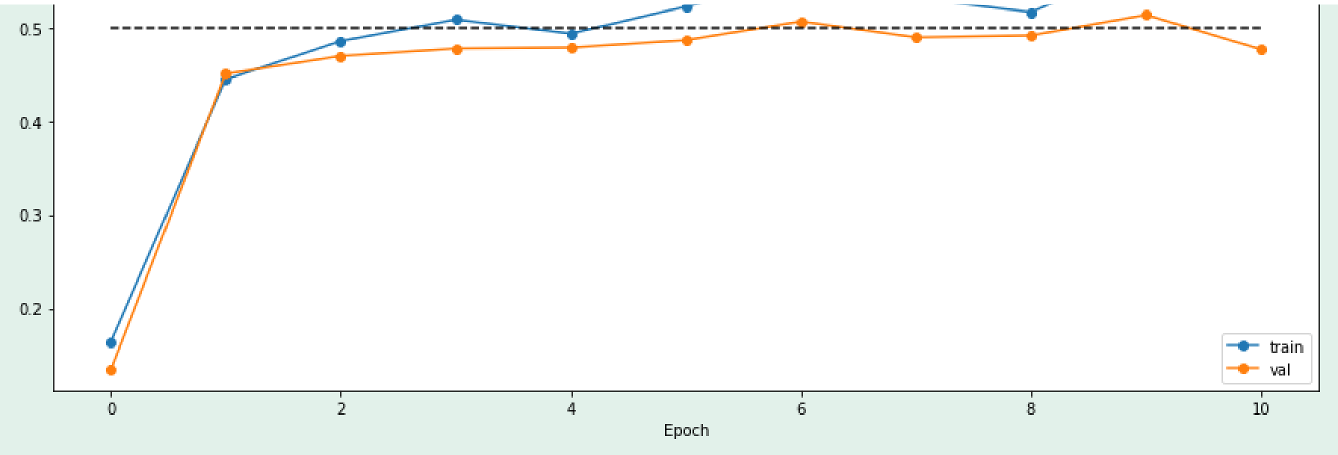
```
(Iteration 1 / 4900) loss: 2.378087
(Epoch 0 / 10) train acc: 0.164000; val_acc: 0.134000
(Iteration 101 / 4900) loss: 1.905230
(Iteration 201 / 4900) loss: 2.049261
(Iteration 301 / 4900) loss: 1.713070
(Iteration 401 / 4900) loss: 1.582693
(Epoch 1 / 10) train acc: 0.445000; val_acc: 0.451000
(Iteration 501 / 4900) loss: 1.673854
(Iteration 601 / 4900) loss: 1.542241
(Iteration 701 / 4900) loss: 1.661488
(Iteration 801 / 4900) loss: 1.698760
(Iteration 901 / 4900) loss: 1.523104
(Epoch 2 / 10) train acc: 0.486000; val_acc: 0.470000
(Iteration 1001 / 4900) loss: 1.575808
(Iteration 1101 / 4900) loss: 1.559150
(Iteration 1201 / 4900) loss: 1.512748
(Iteration 1301 / 4900) loss: 1.403581
(Iteration 1401 / 4900) loss: 1.582180
(Epoch 3 / 10) train acc: 0.509000; val_acc: 0.478000
(Iteration 1501 / 4900) loss: 1.520756
(Iteration 1601 / 4900) loss: 1.502950
(Iteration 1701 / 4900) loss: 1.504738
(Iteration 1801 / 4900) loss: 1.639809
(Iteration 1901 / 4900) loss: 1.588290
(Epoch 4 / 10) train acc: 0.494000; val_acc: 0.479000
(Iteration 2001 / 4900) loss: 1.617043
```

```
(Iteration 2101 / 4900) loss: 1.571903
(Iteration 2201 / 4900) loss: 1.580801
(Iteration 2301 / 4900) loss: 1.319308
(Iteration 2401 / 4900) loss: 1.405106
(Epoch 5 / 10) train acc: 0.523000; val_acc: 0.487000
(Iteration 2501 / 4900) loss: 1.462077
(Iteration 2601 / 4900) loss: 1.478338
(Iteration 2701 / 4900) loss: 1.452321
(Iteration 2801 / 4900) loss: 1.501195
(Iteration 2901 / 4900) loss: 1.444775
(Epoch 6 / 10) train acc: 0.553000; val_acc: 0.507000
(Iteration 3001 / 4900) loss: 1.377007
(Iteration 3101 / 4900) loss: 1.252317
(Iteration 3201 / 4900) loss: 1.703810
(Iteration 3301 / 4900) loss: 1.449870
(Iteration 3401 / 4900) loss: 1.579887
(Epoch 7 / 10) train acc: 0.532000; val_acc: 0.490000
(Iteration 3501 / 4900) loss: 1.427413
(Iteration 3601 / 4900) loss: 1.284990
(Iteration 3701 / 4900) loss: 1.474521
(Iteration 3801 / 4900) loss: 1.398789
(Iteration 3901 / 4900) loss: 1.239221
(Epoch 8 / 10) train acc: 0.517000; val_acc: 0.492000
(Iteration 4001 / 4900) loss: 1.355188
(Iteration 4101 / 4900) loss: 1.393246
(Iteration 4201 / 4900) loss: 1.277518
(Iteration 4301 / 4900) loss: 1.229724
(Iteration 4401 / 4900) loss: 1.707493
(Epoch 9 / 10) train acc: 0.570000; val_acc: 0.514000
(Iteration 4501 / 4900) loss: 1.374712
(Iteration 4601 / 4900) loss: 1.592834
(Iteration 4701 / 4900) loss: 1.579716
(Iteration 4801 / 4900) loss: 1.336092
(Epoch 10 / 10) train acc: 0.566000; val_acc: 0.477000
```

In [11]:

```python
# 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 `cs231n/classifiers/fc_net.py`.

Implement the initialization, the forward pass, and the backward pass. For the moment don't worry about implementing dropout or 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 [12]:

```
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])))
```

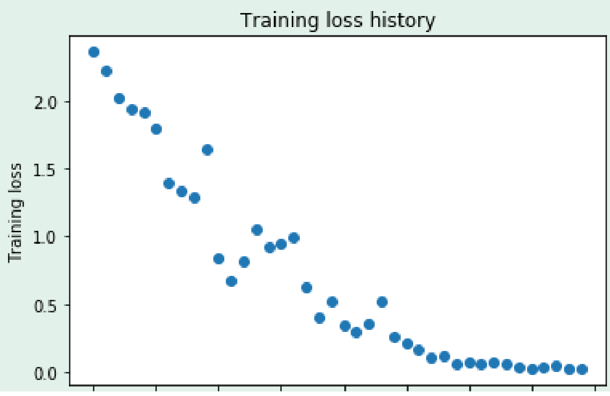
```
Running check with reg =  0
Initial loss:  2.3004790897684924
W1 relative error: 1.48e-07
W2 relative error: 2.21e-05
W3 relative error: 3.53e-07
b1 relative error: 5.38e-09
b2 relative error: 2.09e-09
b3 relative error: 5.80e-11
Running check with reg =  3.14
Initial loss:  7.052114776533016
W1 relative error: 1.14e-08
W2 relative error: 6.87e-08
W3 relative error: 3.48e-08
b1 relative error: 1.48e-08
b2 relative error: 1.72e-09
b3 relative error: 1.80e-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 **weight initialization scale** to overfit and achieve 100% training accuracy within 20 epochs.

In [13]:

```
# 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 = 1e-2   # Experiment with this!
learning_rate = 1e-2  # Experiment with this!
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: 2.363364
(Epoch 0 / 20) train acc: 0.180000; val_acc: 0.108000
(Epoch 1 / 20) train acc: 0.320000; val_acc: 0.127000
(Epoch 2 / 20) train acc: 0.440000; val_acc: 0.172000
(Epoch 3 / 20) train acc: 0.500000; val_acc: 0.184000
(Epoch 4 / 20) train acc: 0.540000; val_acc: 0.181000
(Epoch 5 / 20) train acc: 0.740000; val_acc: 0.190000
(Iteration 11 / 40) loss: 0.839976
(Epoch 6 / 20) train acc: 0.740000; val_acc: 0.187000
(Epoch 7 / 20) train acc: 0.740000; val_acc: 0.183000
(Epoch 8 / 20) train acc: 0.820000; val_acc: 0.177000
(Epoch 9 / 20) train acc: 0.860000; val_acc: 0.200000
(Epoch 10 / 20) train acc: 0.920000; val_acc: 0.191000
(Iteration 21 / 40) loss: 0.337174
(Epoch 11 / 20) train acc: 0.960000; val_acc: 0.189000
(Epoch 12 / 20) train acc: 0.940000; val_acc: 0.180000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.199000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.199000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.195000
(Iteration 31 / 40) loss: 0.075911
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.182000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.201000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.207000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.185000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.192000
```

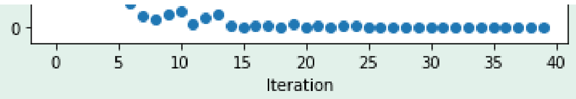


Now try to use a five-layer network with 100 units on each layer to overfit 50 training examples. Again, you will have to adjust the learning rate and weight initialization scale, but you should be able to achieve 100% training accuracy within 20 epochs.

In [14]:

```
# 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 = 3e-4  # Experiment with this!
weight_scale = 1e-1   # Experiment with this!
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: 166.501707
(Epoch 0 / 20) train acc: 0.160000; val_acc: 0.120000
(Epoch 1 / 20) train acc: 0.240000; val_acc: 0.095000
(Epoch 2 / 20) train acc: 0.320000; val_acc: 0.123000
(Epoch 3 / 20) train acc: 0.420000; val_acc: 0.120000
(Epoch 4 / 20) train acc: 0.640000; val_acc: 0.128000
(Epoch 5 / 20) train acc: 0.680000; val_acc: 0.112000
(Iteration 11 / 40) loss: 8.208668
(Epoch 6 / 20) train acc: 0.820000; val_acc: 0.128000
(Epoch 7 / 20) train acc: 0.860000; val_acc: 0.118000
(Epoch 8 / 20) train acc: 0.920000; val_acc: 0.118000
(Epoch 9 / 20) train acc: 0.960000; val_acc: 0.122000
(Epoch 10 / 20) train acc: 0.920000; val_acc: 0.120000
(Iteration 21 / 40) loss: 0.206859
(Epoch 11 / 20) train acc: 0.940000; val_acc: 0.111000
(Epoch 12 / 20) train acc: 0.960000; val_acc: 0.122000
(Epoch 13 / 20) train acc: 0.980000; val_acc: 0.128000
(Epoch 14 / 20) train acc: 0.960000; val_acc: 0.118000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.122000
(Iteration 31 / 40) loss: 0.000895
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.119000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.120000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.120000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.120000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.120000
```



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:

The five-layer net is harder to train and to find correct hyperparameters (learning rate and weight scale). And five-layer network is more sensitive to the initialization hyperparameter scale. Because five-layer net converges much faster and produces more precise result. It will be more sensitive to initial set up.

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. See the Momentum Update section at http://cs231n.github.io/neural-networks-3/#sgd for more information.

Open the file `cs231n/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 [15]:

```
from cs231n.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.

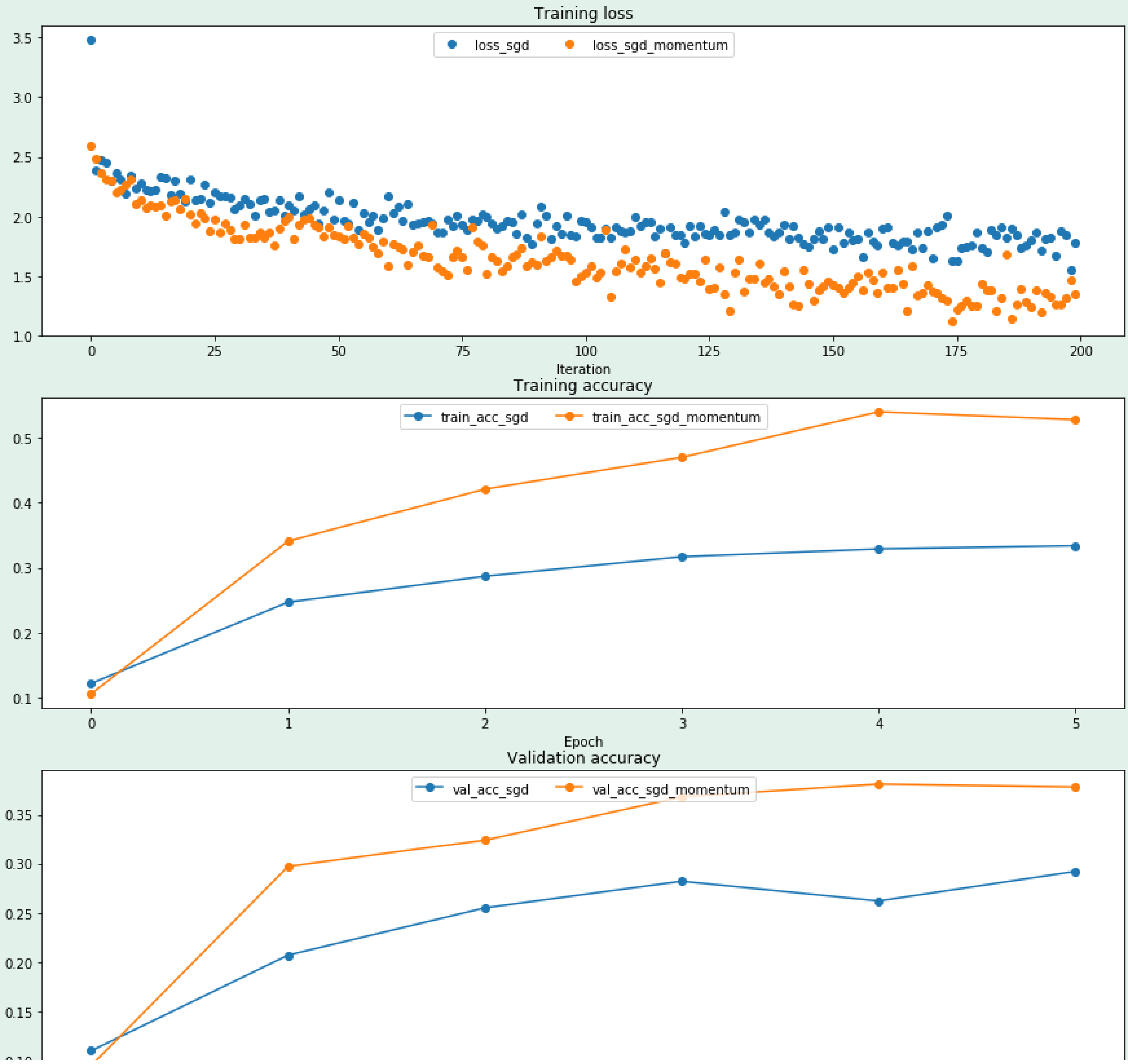
In [17]:

```
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': 5e-3,
                  },
                  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 solvers.items():
  plt.subplot(3, 1, 1)
  plt.plot(solver.loss_history, 'o', label="loss_%s" % update_rule)

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

  plt.subplot(3, 1, 3)
  plt.plot(solver.val_acc_history, '-o', label="val_acc_%s" % 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: 3.476928
(Epoch 0 / 5) train acc: 0.122000; val_acc: 0.110000
(Iteration 11 / 200) loss: 2.273918
(Iteration 21 / 200) loss: 2.306955
(Iteration 31 / 200) loss: 2.093337
(Epoch 1 / 5) train acc: 0.247000; val_acc: 0.207000
(Iteration 41 / 200) loss: 2.089881
(Iteration 51 / 200) loss: 2.142988
(Iteration 61 / 200) loss: 2.169558
(Iteration 71 / 200) loss: 1.870478
(Epoch 2 / 5) train acc: 0.287000; val_acc: 0.255000
(Iteration 81 / 200) loss: 1.994859
(Iteration 91 / 200) loss: 1.948483
(Iteration 101 / 200) loss: 1.956554
(Iteration 111 / 200) loss: 1.997367
(Epoch 3 / 5) train acc: 0.317000; val_acc: 0.282000
(Iteration 121 / 200) loss: 1.786599
(Iteration 131 / 200) loss: 1.863651
(Iteration 141 / 200) loss: 1.936477
(Iteration 151 / 200) loss: 1.732275
(Epoch 4 / 5) train acc: 0.329000; val_acc: 0.262000
(Iteration 161 / 200) loss: 1.903142
(Iteration 171 / 200) loss: 1.649538
(Iteration 181 / 200) loss: 1.742841
```

```
(Iteration 191 / 200) loss: 1.806653
(Epoch 5 / 5) train acc: 0.334000; val_acc: 0.292000

running with  sgd_momentum
(Iteration 1 / 200) loss: 2.589166
(Epoch 0 / 5) train acc: 0.106000; val_acc: 0.095000
(Iteration 11 / 200) loss: 2.139318
(Iteration 21 / 200) loss: 2.017108
(Iteration 31 / 200) loss: 1.813453
(Epoch 1 / 5) train acc: 0.341000; val_acc: 0.297000
(Iteration 41 / 200) loss: 2.002500
(Iteration 51 / 200) loss: 1.830173
(Iteration 61 / 200) loss: 1.584339
(Iteration 71 / 200) loss: 1.579533
(Epoch 2 / 5) train acc: 0.421000; val_acc: 0.324000
(Iteration 81 / 200) loss: 1.523871
(Iteration 91 / 200) loss: 1.597867
(Iteration 101 / 200) loss: 1.527771
(Iteration 111 / 200) loss: 1.638278
(Epoch 3 / 5) train acc: 0.470000; val_acc: 0.368000
(Iteration 121 / 200) loss: 1.482900
(Iteration 131 / 200) loss: 1.533665
(Iteration 141 / 200) loss: 1.547938
(Iteration 151 / 200) loss: 1.419212
(Epoch 4 / 5) train acc: 0.540000; val_acc: 0.381000
(Iteration 161 / 200) loss: 1.531513
(Iteration 171 / 200) loss: 1.371091
(Iteration 181 / 200) loss: 1.438618
(Iteration 191 / 200) loss: 1.241694
(Epoch 5 / 5) train acc: 0.528000; val_acc: 0.378000
```



RMSProp and Adam

RMSProp [1] and Adam [2] are update rules that set per-parameter learning rates by using a running average of the second moments of gradients.

In the file `cs231n/optim.py`, implement the RMSProp update rule in the `rmsprop` function and implement the Adam update rule in the `adam` function, and check your implementations using the tests below.

NOTE: Please implement the *complete* Adam update rule (with the bias correction mechanism), not the first simplified version mentioned in the course notes.

[1] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude." COURSERA: Neural Networks for Machine Learning 4 (2012).

[2] Diederik Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", ICLR 2015.

In [18]:

```
# Test RMSProp implementation
from cs231n.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)
cache = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-2, 'cache': cache}
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    ]])

# You should see relative errors around e-7 or less
print('next_w error: ', rel_error(expected_next_w, next_w))
print('cache error: ', rel_error(expected_cache, config['cache']))
```

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

In [19]:

```
# Test Adam implementation
from cs231n.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)
m = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)
v = np.linspace(0.7, 0.5, num=N*D).reshape(N, D)

config = {'learning_rate': 1e-2, 'm': m, 'v': v, '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_v = 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_m = 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      ]])

# You should see relative errors around e-7 or less
print('next_w error: ', rel_error(expected_next_w, next_w))
print('v error: ', rel_error(expected_v, config['v']))
print('m error: ', rel_error(expected_m, config['m']))
```

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

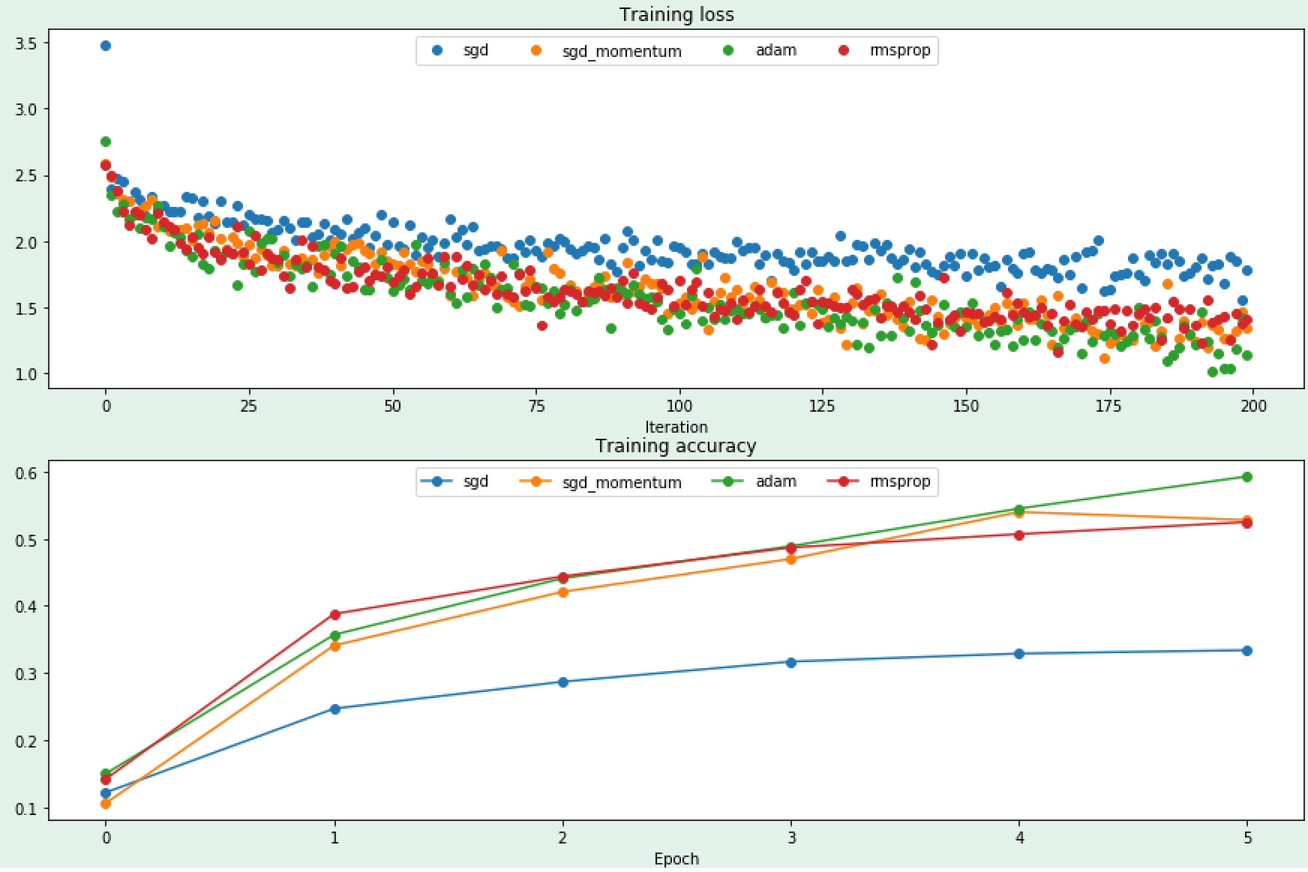
Once you have debugged your RMSProp and Adam implementations, run the following to train a pair of deep networks using these new update rules:

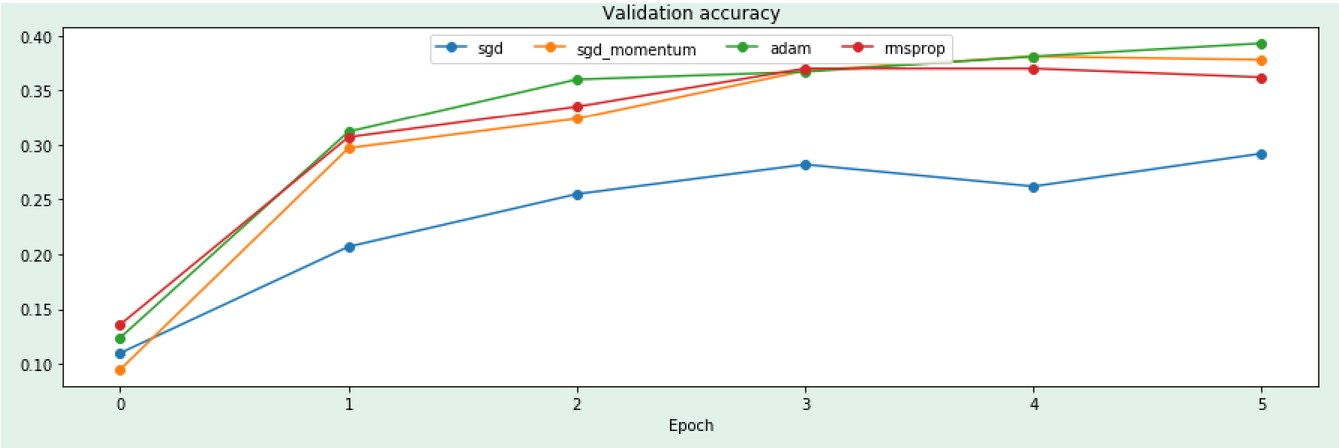
In [20]:

```
learning_rates = {'rmsprop': 1e-4, 'adam': 1e-3}
for update_rule in ['adam', 'rmsprop']:
  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': learning_rates[update_rule]
                  },
                  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  adam
(Iteration 1 / 200) loss: 2.760209
(Epoch 0 / 5) train acc: 0.150000; val_acc: 0.124000
(Iteration 11 / 200) loss: 2.110602
(Iteration 21 / 200) loss: 1.883176
(Iteration 31 / 200) loss: 1.842409
(Epoch 1 / 5) train acc: 0.357000; val_acc: 0.312000
(Iteration 41 / 200) loss: 1.912024
(Iteration 51 / 200) loss: 1.622724
(Iteration 61 / 200) loss: 1.598922
(Iteration 71 / 200) loss: 1.647785
(Epoch 2 / 5) train acc: 0.441000; val_acc: 0.360000
```

```
(Iteration 81 / 200) loss: 1.517582
(Iteration 91 / 200) loss: 1.621971
(Iteration 101 / 200) loss: 1.456345
(Iteration 111 / 200) loss: 1.491637
(Epoch 3 / 5) train acc: 0.489000; val_acc: 0.367000
(Iteration 121 / 200) loss: 1.607893
(Iteration 131 / 200) loss: 1.394200
(Iteration 141 / 200) loss: 1.321148
(Iteration 151 / 200) loss: 1.349208
(Epoch 4 / 5) train acc: 0.545000; val_acc: 0.381000
(Iteration 161 / 200) loss: 1.248493
(Iteration 171 / 200) loss: 1.153209
(Iteration 181 / 200) loss: 1.493787
(Iteration 191 / 200) loss: 1.214109
(Epoch 5 / 5) train acc: 0.593000; val_acc: 0.393000

running with  rmsprop
(Iteration 1 / 200) loss: 2.571494
(Epoch 0 / 5) train acc: 0.142000; val_acc: 0.136000
(Iteration 11 / 200) loss: 2.148984
(Iteration 21 / 200) loss: 1.858964
(Iteration 31 / 200) loss: 1.865406
(Epoch 1 / 5) train acc: 0.388000; val_acc: 0.307000
(Iteration 41 / 200) loss: 1.682452
(Iteration 51 / 200) loss: 1.690672
(Iteration 61 / 200) loss: 1.653560
(Iteration 71 / 200) loss: 1.602213
(Epoch 2 / 5) train acc: 0.444000; val_acc: 0.335000
(Iteration 81 / 200) loss: 1.633064
(Iteration 91 / 200) loss: 1.700801
(Iteration 101 / 200) loss: 1.698334
(Iteration 111 / 200) loss: 1.405926
(Epoch 3 / 5) train acc: 0.487000; val_acc: 0.370000
(Iteration 121 / 200) loss: 1.445123
(Iteration 131 / 200) loss: 1.634678
(Iteration 141 / 200) loss: 1.460120
(Iteration 151 / 200) loss: 1.462433
(Epoch 4 / 5) train acc: 0.507000; val_acc: 0.370000
(Iteration 161 / 200) loss: 1.462897
(Iteration 171 / 200) loss: 1.357323
(Iteration 181 / 200) loss: 1.451439
(Iteration 191 / 200) loss: 1.361932
(Epoch 5 / 5) train acc: 0.525000; val_acc: 0.362000
```





Inline Question 3:

AdaGrad, like Adam, is a per-parameter optimization method that uses the following update rule:

```
cache += dw**2
w += - learning_rate * dw / (np.sqrt(cache) + eps)
```

John notices that when he was training a network with AdaGrad that the updates became very small, and that his network was learning slowly. Using your knowledge of the AdaGrad update rule, why do you think the updates would become very small? Would Adam have the same issue?

Answer:

Because every time the square of dw is added to cache. As the learning progresses, the value of cache will become larger and larger, and in the process of w update, it needs to be divided by cache, which will cause the actual learnig rate to be smaller and smaller.

Adam would not have this kind of issue. Because in Adam, m and v are similar to cache but they are moving averages of squared gradients. Hyperparameters beta1 and beta2 will respectively make m and v to be leaky so the learning rate updates do not get monotonically smaller.

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` and `Dropout.ipynb` notebooks before completing this part, since those techniques can help you train powerful models.

In [23]:

```
best_model = None
################################################################################
# TODO: Train the best FullyConnectedNet that you can on CIFAR-10. You might   #
# find batch/layer normalization and dropout useful. Store your best model in  #
# the best_model variable.                                                     #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****

weight_scales = [1e-2, 1e-3]
learning_rates = [1e-5, 1e-4, 1e-3]
regulation_strengths = [1e-4, 1e-3]

best_val_accuracy = 0

for weight_scale in weight_scales:
    for lr in learning_rates:
```

```
        for reg in regulation_strengths:
            model = FullyConnectedNet([200, 100],
                                      weight_scale=weight_scale,
                                      reg=reg)

            solver = Solver(model,
                            data,
                            num_epochs=10,
                            batch_size=200,
                            update_rule='adam',
                            optim_config={
                                'learning_rate': lr
                            },
                            verbose=False)

            solver.train()

            val_accuracy = solver.best_val_acc

            if best_val_accuracy < val_accuracy:
                best_val_accuracy = val_accuracy
                best_model = model

            print('learning rate %e weight scale %e regulation strength %e val accuracy: %f' % (lr,
weight_scale, reg, val_accuracy))


print('Best validation accuracy achieved is: %f' % best_val_accuracy)


# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                              END OF YOUR CODE                                #
################################################################################
```

```
learning rate 1.000000e-05 weight scale 1.000000e-02 regulation strength 1.000000e-04 val
accuracy: 0.483000
learning rate 1.000000e-05 weight scale 1.000000e-02 regulation strength 1.000000e-03 val
accuracy: 0.468000
learning rate 1.000000e-04 weight scale 1.000000e-02 regulation strength 1.000000e-04 val
accuracy: 0.542000
learning rate 1.000000e-04 weight scale 1.000000e-02 regulation strength 1.000000e-03 val
accuracy: 0.531000
learning rate 1.000000e-03 weight scale 1.000000e-02 regulation strength 1.000000e-04 val
accuracy: 0.525000
learning rate 1.000000e-03 weight scale 1.000000e-02 regulation strength 1.000000e-03 val
accuracy: 0.509000
learning rate 1.000000e-05 weight scale 1.000000e-03 regulation strength 1.000000e-04 val
accuracy: 0.447000
learning rate 1.000000e-05 weight scale 1.000000e-03 regulation strength 1.000000e-03 val
accuracy: 0.445000
learning rate 1.000000e-04 weight scale 1.000000e-03 regulation strength 1.000000e-04 val
accuracy: 0.531000
learning rate 1.000000e-04 weight scale 1.000000e-03 regulation strength 1.000000e-03 val
accuracy: 0.522000
learning rate 1.000000e-03 weight scale 1.000000e-03 regulation strength 1.000000e-04 val
accuracy: 0.509000
learning rate 1.000000e-03 weight scale 1.000000e-03 regulation strength 1.000000e-03 val
accuracy: 0.505000
Best validation accuracy achieved is: 0.542000
```

Test your model!

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

In [25]:

```
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:  0.542
Test set accuracy:  0.534
```