In [1]:
```
# 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. 'cs231n/assignments/assignment2/'
FOLDERNAME = 'cs231n/assignments/assignment2/'
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/$FOLDERNAME/cs231n/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
Mounted at /content/drive
/content/drive/My Drive/cs231n/assignments/assignment2/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment2
```

Multi-Layer Fully Connected Network

In this exercise, 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 network initialization, forward pass, and backward pass. Throughout this assignment, you will be implementing layers in `cs231n/layers.py`. You can re-use your implementations for `affine_forward`, `affine_backward`, `relu_forward`, `relu_backward`, and `softmax_loss` from Assignment 1. For right now, don't worry about implementing dropout or batch/layer normalization yet, as you will add those features later.

In [2]:
```
# Setup cell.
import time
import numpy as np
import matplotlib.pyplot as plt
from cs231n.classifiers.fc_net import *
from cs231n.data_utils import get_CIFAR10_data
from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradient_array
from cs231n.solver import Solver

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

%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))))
```

```
=========== You can safely ignore the message below if you are NOT working on ConvolutionalNetworks.ipynb ======
=====
        You will need to compile a Cython extension for a portion of this assignment.
        The instructions to do this will be given in a section of the notebook below.
```

In [3]:
```
# Load the (preprocessed) CIFAR-10 data.
data = get_CIFAR10_data()
for k, v in list(data.items()):
    print(f"{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,)
```

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. This is a good way to see if the initial losses seem reasonable.

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

In [4]:
```
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(f"{name} relative error: {rel_error(grad_num, grads[name])}")
```

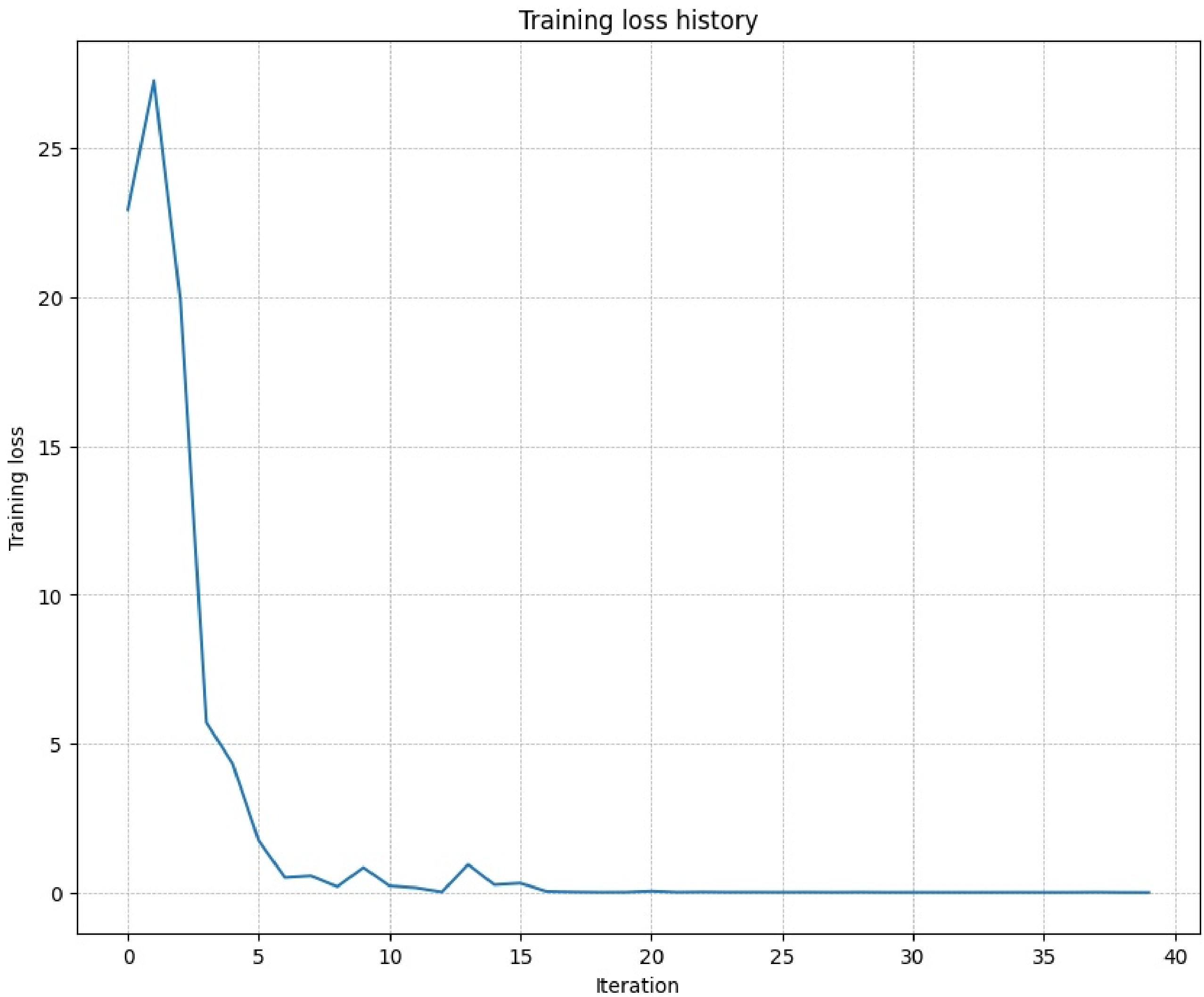
```
Running check with reg =  0
Initial loss:  2.3004790897684924
W1 relative error: 1.4839894098713283e-07
W2 relative error: 2.21204793107852e-05
W3 relative error: 3.527252851540647e-07
b1 relative error: 5.376386228531692e-09
b2 relative error: 2.085654200257447e-09
b3 relative error: 5.7957243458479405e-11
Running check with reg =  3.14
Initial loss:  7.052114776533016
W1 relative error: 6.862884860440611e-09
W2 relative error: 3.522821562176466e-08
W3 relative error: 1.3225242980747655e-08
b1 relative error: 1.4752428222134868e-08
b2 relative error: 1.7223750761525226e-09
b3 relative error: 1.801765144951982e-10
```

As another sanity check, make sure your network can overfit on 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 [5]:
```
# 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 =  4e-2    # Experiment with this!
learning_rate = 3e-3  # 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)
plt.title("Training loss history")
plt.xlabel("Iteration")
```

```
plt.ylabel("Training loss")
plt.grid(linestyle='--', linewidth=0.5)
plt.show()
```

```
(Iteration 1 / 40) loss: 22.942158
(Epoch 0 / 20) train acc: 0.200000; val_acc: 0.111000
(Epoch 1 / 20) train acc: 0.400000; val_acc: 0.136000
(Epoch 2 / 20) train acc: 0.560000; val_acc: 0.142000
(Epoch 3 / 20) train acc: 0.760000; val_acc: 0.124000
(Epoch 4 / 20) train acc: 0.900000; val_acc: 0.140000
(Epoch 5 / 20) train acc: 0.860000; val_acc: 0.135000
(Iteration 11 / 40) loss: 0.224576
(Epoch 6 / 20) train acc: 0.940000; val_acc: 0.150000
(Epoch 7 / 20) train acc: 0.960000; val_acc: 0.157000
(Epoch 8 / 20) train acc: 0.980000; val_acc: 0.152000
(Epoch 9 / 20) train acc: 0.980000; val_acc: 0.151000
(Epoch 10 / 20) train acc: 0.980000; val_acc: 0.151000
(Iteration 21 / 40) loss: 0.046719
(Epoch 11 / 20) train acc: 1.000000; val_acc: 0.148000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.149000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.149000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.150000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.150000
(Iteration 31 / 40) loss: 0.007623
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.151000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.151000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.150000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.150000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.150000
```

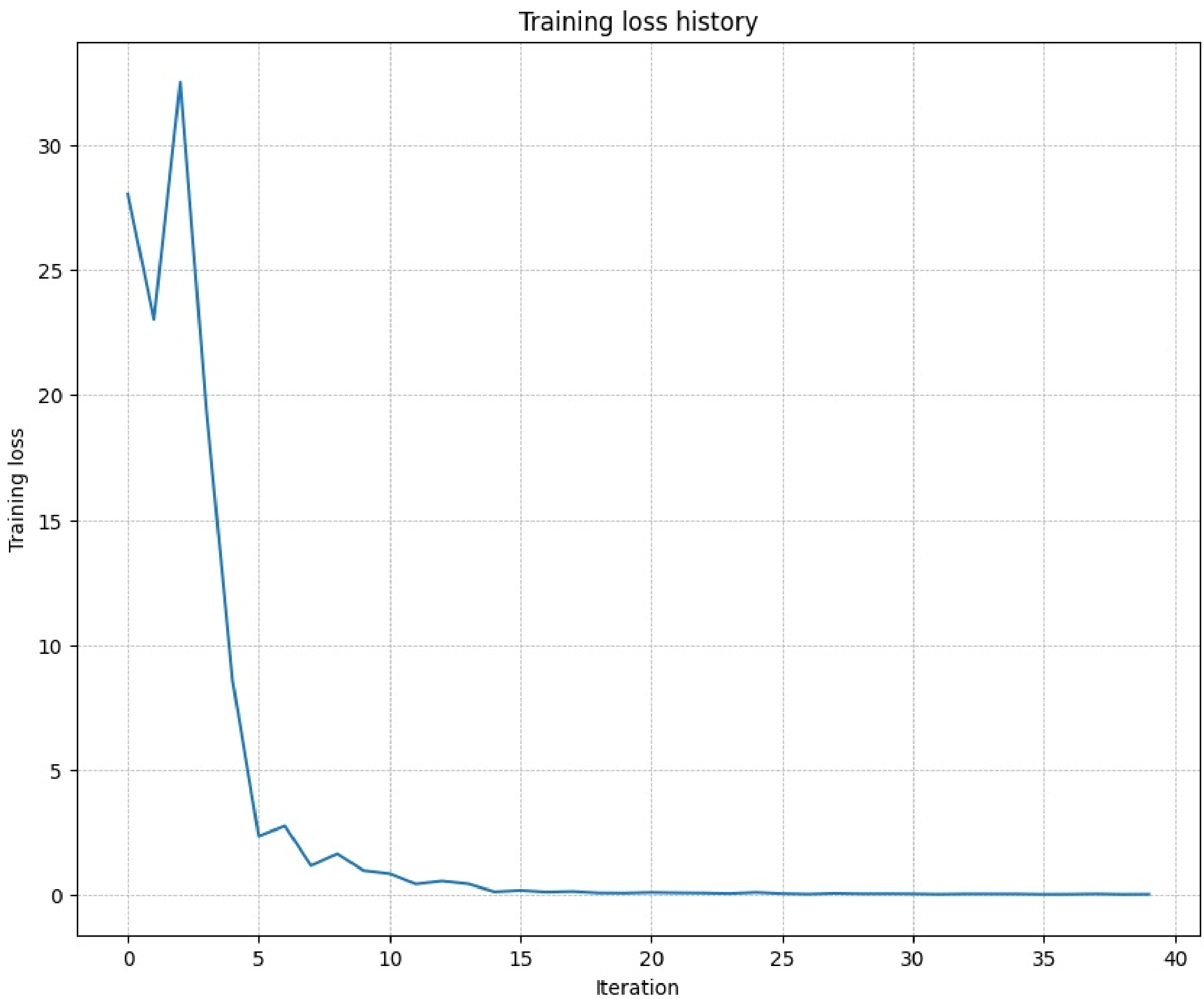


Now, try to use a five-layer network with 100 units on each layer to overfit on 50 training examples. Again, you will have to adjust the learning rate and weight initialization scale, but you should be able to achieve 100% training accuracy within 20 epochs.

In [6]:
```
# 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'],
```

```python
}

learning_rate = 5e-3  # Experiment with this!
weight_scale = 7e-2   # 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)
plt.title('Training loss history')
plt.xlabel('Iteration')
plt.ylabel('Training loss')
plt.grid(linestyle='--', linewidth=0.5)
plt.show()
```

```
(Iteration 1 / 40) loss: 28.032018
(Epoch 0 / 20) train acc: 0.120000; val_acc: 0.109000
(Epoch 1 / 20) train acc: 0.320000; val_acc: 0.094000
(Epoch 2 / 20) train acc: 0.160000; val_acc: 0.126000
(Epoch 3 / 20) train acc: 0.540000; val_acc: 0.123000
(Epoch 4 / 20) train acc: 0.720000; val_acc: 0.110000
(Epoch 5 / 20) train acc: 0.820000; val_acc: 0.108000
(Iteration 11 / 40) loss: 0.854613
(Epoch 6 / 20) train acc: 0.880000; val_acc: 0.122000
(Epoch 7 / 20) train acc: 0.960000; val_acc: 0.128000
(Epoch 8 / 20) train acc: 1.000000; val_acc: 0.119000
(Epoch 9 / 20) train acc: 1.000000; val_acc: 0.131000
(Epoch 10 / 20) train acc: 1.000000; val_acc: 0.129000
(Iteration 21 / 40) loss: 0.103280
(Epoch 11 / 20) train acc: 1.000000; val_acc: 0.130000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.133000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.130000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.131000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.130000
(Iteration 31 / 40) loss: 0.043595
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.133000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.131000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.128000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.129000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.131000
```



Inline Question 1:

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

Answer:

როგორც ექსპერიმენტებიდან დავინახე 5 layer-იანი უფრო მგრძნობიარე აღმოჩნდა 3-იანთან შედარებით. 5 layer-იანი უფრო მგრძნობიარეა იმიტომ რომ ზოგადად ღრმა ქსელები უფრო არიან მგრძნობიარეები, რადგან მათთვის გრადიენტის გაქრობა უფრო მარტივად შეიძლება. ეხა რო შევადაროთ 5 layerიანმა დაიწყო 18 იანი loss-ით ხოლო 3-იანმა 14 ით მაგრამ დაჭირდა 8 ეპოქა ისევე როგორც 3იანს დაჭირდა 8 ეპოქა. მაღალი დანაკარგი დასაწყისშ ნიშნავს რომ უფრო მგრძნობიარეა ინიციალიზაციისას. ასევე 3 შრიანი ბევრად სწრაფად მივიდა 5 შრიანტან შედარებით უფრო მაღალ სიზუსტეზე.

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 [7]:
```
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"]))
```

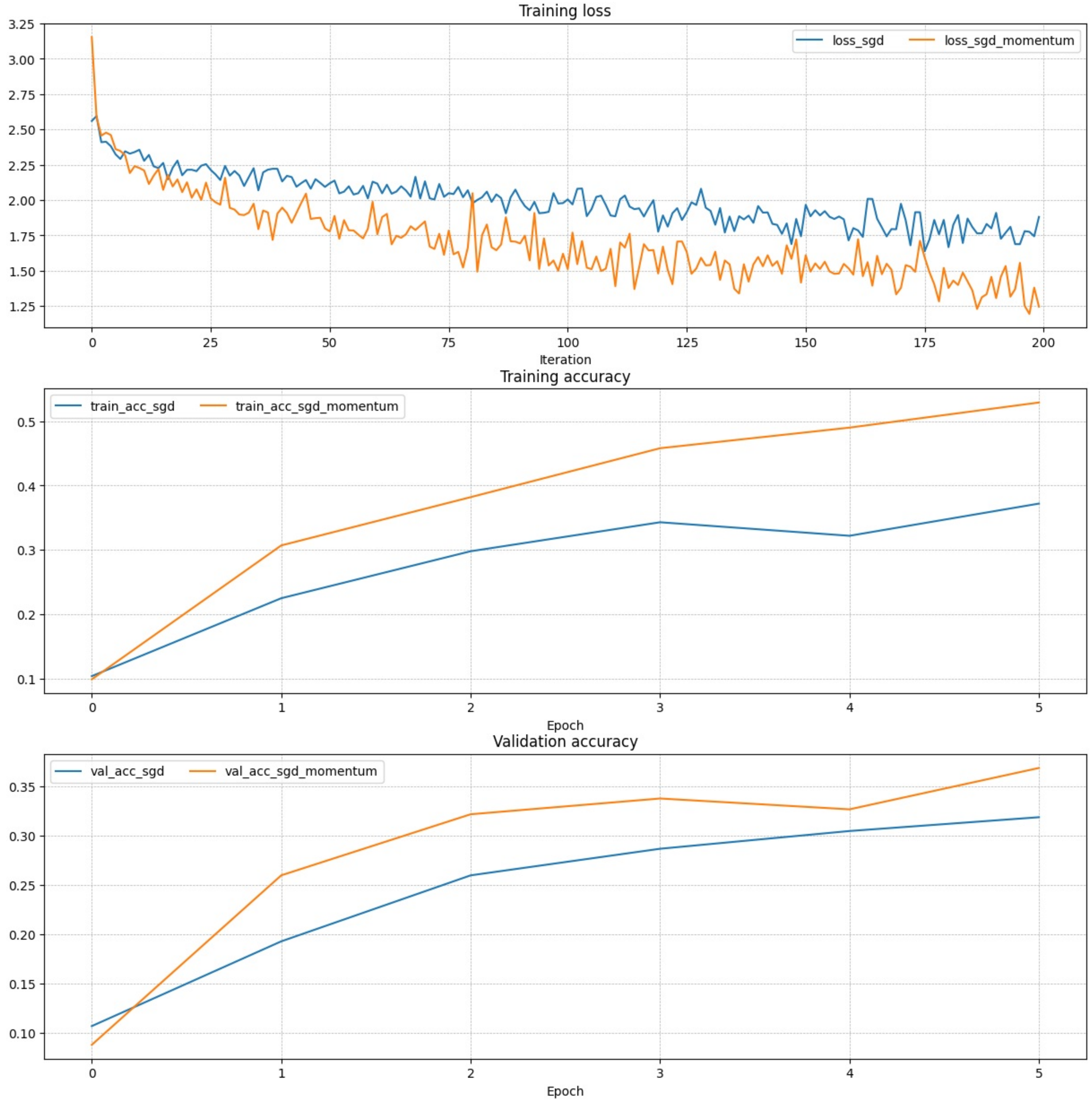
```
next_w error:  8.882347033505819e-09
velocity error:  4.269287743278663e-09
```

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 [8]:
```
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,
```

```python
    )
    solvers[update_rule] = solver
    solver.train()

fig, axes = plt.subplots(3, 1, figsize=(15, 15))

axes[0].set_title('Training loss')
axes[0].set_xlabel('Iteration')
axes[1].set_title('Training accuracy')
axes[1].set_xlabel('Epoch')
axes[2].set_title('Validation accuracy')
axes[2].set_xlabel('Epoch')

for update_rule, solver in solvers.items():
    axes[0].plot(solver.loss_history, label=f"loss_{update_rule}")
    axes[1].plot(solver.train_acc_history, label=f"train_acc_{update_rule}")
    axes[2].plot(solver.val_acc_history, label=f"val_acc_{update_rule}")

for ax in axes:
    ax.legend(loc="best", ncol=4)
    ax.grid(linestyle='--', linewidth=0.5)

plt.show()
```

```
Running with  sgd
(Iteration 1 / 200) loss: 2.559978
(Epoch 0 / 5) train acc: 0.104000; val_acc: 0.107000
(Iteration 11 / 200) loss: 2.356070
(Iteration 21 / 200) loss: 2.214091
(Iteration 31 / 200) loss: 2.205928
(Epoch 1 / 5) train acc: 0.225000; val_acc: 0.193000
(Iteration 41 / 200) loss: 2.132095
(Iteration 51 / 200) loss: 2.118950
(Iteration 61 / 200) loss: 2.116443
(Iteration 71 / 200) loss: 2.132549
(Epoch 2 / 5) train acc: 0.298000; val_acc: 0.260000
(Iteration 81 / 200) loss: 1.977227
(Iteration 91 / 200) loss: 2.007528
(Iteration 101 / 200) loss: 2.004762
(Iteration 111 / 200) loss: 1.885342
(Epoch 3 / 5) train acc: 0.343000; val_acc: 0.287000
(Iteration 121 / 200) loss: 1.891517
(Iteration 131 / 200) loss: 1.923677
(Iteration 141 / 200) loss: 1.957743
(Iteration 151 / 200) loss: 1.966736
(Epoch 4 / 5) train acc: 0.322000; val_acc: 0.305000
(Iteration 161 / 200) loss: 1.801483
(Iteration 171 / 200) loss: 1.973780
(Iteration 181 / 200) loss: 1.666572
(Iteration 191 / 200) loss: 1.909494
(Epoch 5 / 5) train acc: 0.372000; val_acc: 0.319000
Running with  sgd_momentum
(Iteration 1 / 200) loss: 3.153778
(Epoch 0 / 5) train acc: 0.099000; val_acc: 0.088000
(Iteration 11 / 200) loss: 2.227203
(Iteration 21 / 200) loss: 2.125706
(Iteration 31 / 200) loss: 1.932695
(Epoch 1 / 5) train acc: 0.307000; val_acc: 0.260000
(Iteration 41 / 200) loss: 1.946488
(Iteration 51 / 200) loss: 1.778584
(Iteration 61 / 200) loss: 1.758119
(Iteration 71 / 200) loss: 1.849137
(Epoch 2 / 5) train acc: 0.382000; val_acc: 0.322000
(Iteration 81 / 200) loss: 2.048671
(Iteration 91 / 200) loss: 1.693223
(Iteration 101 / 200) loss: 1.511693
(Iteration 111 / 200) loss: 1.390754
(Epoch 3 / 5) train acc: 0.458000; val_acc: 0.338000
(Iteration 121 / 200) loss: 1.670614
(Iteration 131 / 200) loss: 1.540271
(Iteration 141 / 200) loss: 1.597365
(Iteration 151 / 200) loss: 1.609851
(Epoch 4 / 5) train acc: 0.490000; val_acc: 0.327000
(Iteration 161 / 200) loss: 1.472687
(Iteration 171 / 200) loss: 1.378620
(Iteration 181 / 200) loss: 1.378175
(Iteration 191 / 200) loss: 1.306439
(Epoch 5 / 5) train acc: 0.529000; val_acc: 0.369000
```



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 [9]:
```
# 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 [10]:

```
# 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
```

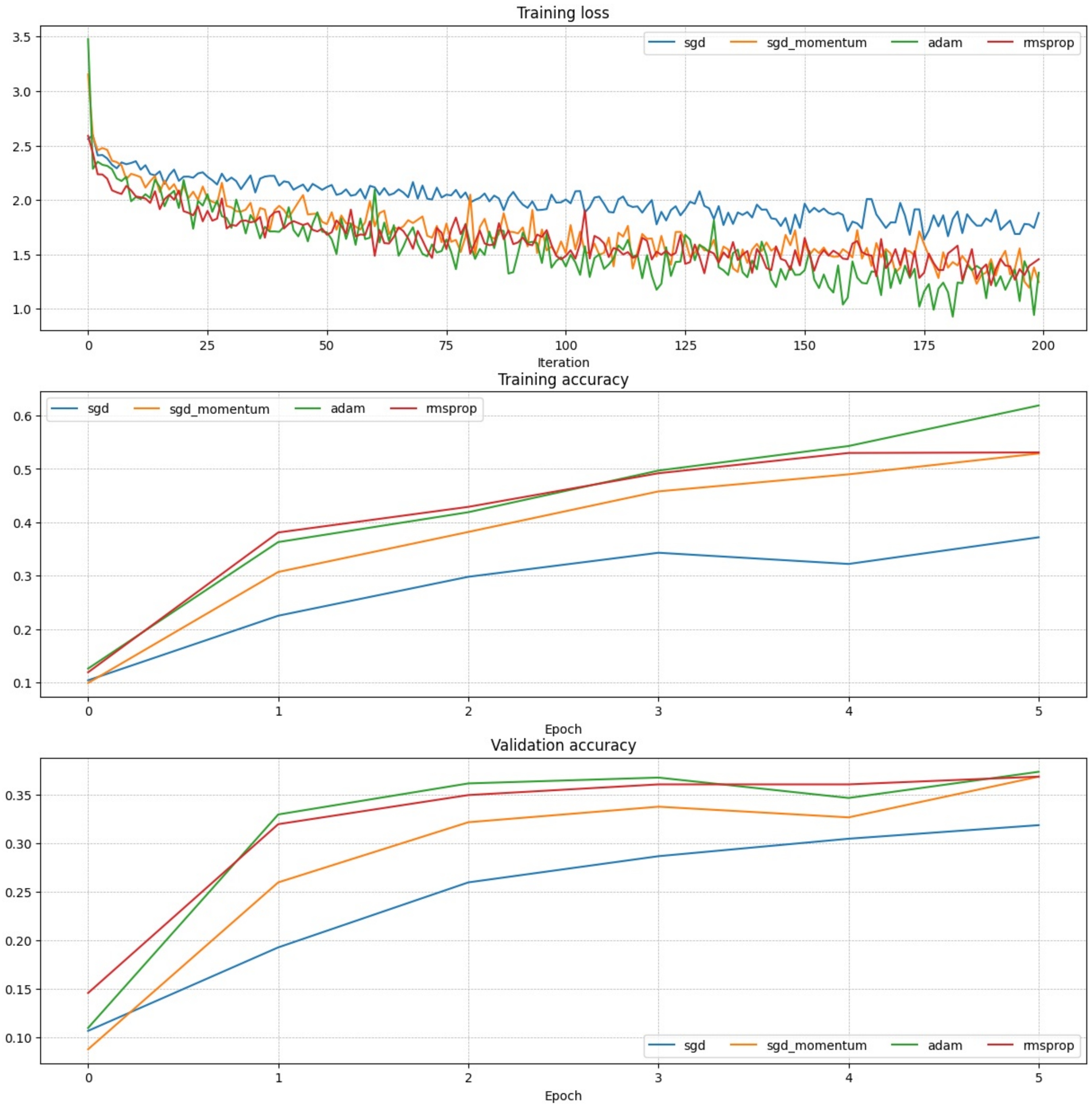
Once you have debugged your RMSProp and Adam implementations, run the following to train a pair of deep networks using these new update rules:

In [11]:

```
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()
```

```python
fig, axes = plt.subplots(3, 1, figsize=(15, 15))

axes[0].set_title('Training loss')
axes[0].set_xlabel('Iteration')
axes[1].set_title('Training accuracy')
axes[1].set_xlabel('Epoch')
axes[2].set_title('Validation accuracy')
axes[2].set_xlabel('Epoch')

for update_rule, solver in solvers.items():
    axes[0].plot(solver.loss_history, label=f"{update_rule}")
    axes[1].plot(solver.train_acc_history, label=f"{update_rule}")
    axes[2].plot(solver.val_acc_history, label=f"{update_rule}")

for ax in axes:
    ax.legend(loc='best', ncol=4)
    ax.grid(linestyle='--', linewidth=0.5)

plt.show()
```

```
Running with  adam
(Iteration 1 / 200) loss: 3.476928
(Epoch 0 / 5) train acc: 0.126000; val_acc: 0.110000
(Iteration 11 / 200) loss: 2.027712
(Iteration 21 / 200) loss: 2.183357
(Iteration 31 / 200) loss: 1.744257
(Epoch 1 / 5) train acc: 0.363000; val_acc: 0.330000
(Iteration 41 / 200) loss: 1.707951
(Iteration 51 / 200) loss: 1.703835
(Iteration 61 / 200) loss: 2.094758
(Iteration 71 / 200) loss: 1.505557
(Epoch 2 / 5) train acc: 0.419000; val_acc: 0.362000
(Iteration 81 / 200) loss: 1.594431
(Iteration 91 / 200) loss: 1.511452
(Iteration 101 / 200) loss: 1.389237
(Iteration 111 / 200) loss: 1.463575
(Epoch 3 / 5) train acc: 0.497000; val_acc: 0.368000
(Iteration 121 / 200) loss: 1.231313
(Iteration 131 / 200) loss: 1.520199
(Iteration 141 / 200) loss: 1.363221
(Iteration 151 / 200) loss: 1.355143
(Epoch 4 / 5) train acc: 0.543000; val_acc: 0.347000
(Iteration 161 / 200) loss: 1.436401
(Iteration 171 / 200) loss: 1.231426
(Iteration 181 / 200) loss: 1.153575
(Iteration 191 / 200) loss: 1.209479
(Epoch 5 / 5) train acc: 0.619000; val_acc: 0.374000

Running with  rmsprop
(Iteration 1 / 200) loss: 2.589166
(Epoch 0 / 5) train acc: 0.119000; val_acc: 0.146000
(Iteration 11 / 200) loss: 2.032921
(Iteration 21 / 200) loss: 1.897277
(Iteration 31 / 200) loss: 1.770793
(Epoch 1 / 5) train acc: 0.381000; val_acc: 0.320000
(Iteration 41 / 200) loss: 1.895731
(Iteration 51 / 200) loss: 1.681091
(Iteration 61 / 200) loss: 1.487204
(Iteration 71 / 200) loss: 1.629973
(Epoch 2 / 5) train acc: 0.429000; val_acc: 0.350000
(Iteration 81 / 200) loss: 1.506686
(Iteration 91 / 200) loss: 1.610742
(Iteration 101 / 200) loss: 1.486124
(Iteration 111 / 200) loss: 1.559454
(Epoch 3 / 5) train acc: 0.492000; val_acc: 0.361000
(Iteration 121 / 200) loss: 1.497406
(Iteration 131 / 200) loss: 1.530736
(Iteration 141 / 200) loss: 1.550957
(Iteration 151 / 200) loss: 1.652046
(Epoch 4 / 5) train acc: 0.530000; val_acc: 0.361000
(Iteration 161 / 200) loss: 1.599574
(Iteration 171 / 200) loss: 1.401073
(Iteration 181 / 200) loss: 1.509365
(Iteration 191 / 200) loss: 1.365773
(Epoch 5 / 5) train acc: 0.531000; val_acc: 0.369000
```



Inline Question 2:

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:

Each iteration adds squared gradients (cache += dw2) without removing old ones. after some time, the cache becomes very large, making the denominator (np.sqrt(cache) + eps) larger and larger. As a result, weight updates become extremely small, almost zero.

Adam doesn't have this problem because it uses a moving average for squared gradients.

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 network.

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 networks rather than fully connected networks.

Note: 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 [18]:

```
print("Starting hyperparameter search on CIFAR-10...")
best_val = -1
best_params = {}
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"],
}

print(f"Using {num_train} training examples for fast hyperparameter search")
print("Testing 15 different hyperparameter combinations with [2048, 1024] architecture")
print("-" * 80)

for i in range(15):
    print(f"Running experiment {i+1}/15...")
    lr = 10 ** np.random.uniform(-3, -2)
    ws = 10 ** np.random.uniform(-2, -1)
    reg = 10 ** np.random.uniform(-5, -3)
    kr = np.random.uniform(.3, .5)

    model = FullyConnectedNet([2048, 1024],
                              weight_scale=ws,
                              reg=reg,
                              dropout_keep_ratio=kr,
                              normalization='batchnorm')

    solver = Solver(model, small_data,
                    num_epochs=3, batch_size=512,
                    update_rule='adam',
                    optim_config={'learning_rate': lr},
                    lr_decay=0.8,
                    verbose=False)

    solver.train()
    new_val = solver.best_val_acc

    if new_val > best_val:
        best_val = new_val
        best_params = {'lr':lr, 'ws':ws, 'reg':reg, 'kr':kr}
        print(f"NEW BEST MODEL FOUND! Accuracy: {new_val:.5f}")

    print(f'lr: {lr:.5f} ws: {ws:.5f}, reg: {reg:.5f}, kr: {kr:.5f}, acc: {new_val:.5f}')
    print("-" * 60)

print(f'Best validation accuracy from search: {best_val:.5f}')
print(f'Best hyperparameters:')
print(f'  Learning rate: {best_params["lr"]:.5f}')
print(f'  Weight scale: {best_params["ws"]:.5f}')
print(f'  Regularization: {best_params["reg"]:.5f}')
print(f'  Dropout keep ratio: {best_params["kr"]:.5f}')
print("-" * 80)
print('Training final model with best parameters on full dataset...')

best_model = FullyConnectedNet([2048, 1024],
                               weight_scale=best_params['ws'],
                               reg=best_params['reg'],
                               dropout_keep_ratio=best_params['kr'],
                               normalization='batchnorm')

solver = Solver(best_model, data,
            num_epochs=5, batch_size=512,
            update_rule='adam',
            optim_config={'learning_rate': best_params['lr']},
            lr_decay=0.85,
```

```
            verbose=True,
            print_every=50)

print("Starting final model training...")
solver.train()

final_val_acc = (np.argmax(best_model.loss(data['X_val']), axis=1) == data['y_val']).mean()
final_test_acc = (np.argmax(best_model.loss(data['X_test']), axis=1) == data['y_test']).mean()

print("-" * 80)
print(f"FINAL RESULTS:")
print(f"Validation accuracy: {final_val_acc:.5f}")
print(f"Test accuracy: {final_test_acc:.5f}")
```

```
Starting hyperparameter search on CIFAR-10...
Using 1000 training examples for fast hyperparameter search
Testing 15 different hyperparameter combinations with [2048, 1024] architecture
--------------------------------------------------------------------------------
Running experiment 1/15...
NEW BEST MODEL FOUND! Accuracy: 0.23400
lr: 0.00567 ws: 0.02090, reg: 0.00011, kr: 0.47627, acc: 0.23400
------------------------------------------------------------
Running experiment 2/15...
lr: 0.00632 ws: 0.02806, reg: 0.00005, kr: 0.36358, acc: 0.22400
------------------------------------------------------------
Running experiment 3/15...
lr: 0.00870 ws: 0.03701, reg: 0.00033, kr: 0.42174, acc: 0.22300
------------------------------------------------------------
Running experiment 4/15...
NEW BEST MODEL FOUND! Accuracy: 0.26200
lr: 0.00830 ws: 0.03505, reg: 0.00057, kr: 0.33088, acc: 0.26200
------------------------------------------------------------
Running experiment 5/15...
NEW BEST MODEL FOUND! Accuracy: 0.31700
lr: 0.00181 ws: 0.01577, reg: 0.00018, kr: 0.46794, acc: 0.31700
------------------------------------------------------------
Running experiment 6/15...
lr: 0.00640 ws: 0.02827, reg: 0.00002, kr: 0.46625, acc: 0.20900
------------------------------------------------------------
Running experiment 7/15...
lr: 0.00901 ws: 0.04207, reg: 0.00006, kr: 0.49976, acc: 0.20300
------------------------------------------------------------
Running experiment 8/15...
lr: 0.00190 ws: 0.02205, reg: 0.00033, kr: 0.34611, acc: 0.27900
------------------------------------------------------------
Running experiment 9/15...
lr: 0.00371 ws: 0.02129, reg: 0.00056, kr: 0.44112, acc: 0.24200
------------------------------------------------------------
Running experiment 10/15...
lr: 0.00202 ws: 0.03377, reg: 0.00015, kr: 0.47819, acc: 0.26800
------------------------------------------------------------
Running experiment 11/15...
lr: 0.00216 ws: 0.03746, reg: 0.00003, kr: 0.49927, acc: 0.27600
------------------------------------------------------------
Running experiment 12/15...
lr: 0.00121 ws: 0.01178, reg: 0.00003, kr: 0.30023, acc: 0.26800
------------------------------------------------------------
Running experiment 13/15...
lr: 0.00575 ws: 0.02052, reg: 0.00004, kr: 0.47966, acc: 0.27200
------------------------------------------------------------
Running experiment 14/15...
lr: 0.00330 ws: 0.04330, reg: 0.00019, kr: 0.37859, acc: 0.24600
------------------------------------------------------------
Running experiment 15/15...
lr: 0.00313 ws: 0.01094, reg: 0.00002, kr: 0.48462, acc: 0.27800
------------------------------------------------------------
Best validation accuracy from search: 0.31700
Best hyperparameters:
  Learning rate: 0.00181
  Weight scale: 0.01577
  Regularization: 0.00018
  Dropout keep ratio: 0.46794
--------------------------------------------------------------------------------
Training final model with best parameters on full dataset...
Starting final model training...
(Iteration 1 / 475) loss: 2.573281
(Epoch 0 / 5) train acc: 0.206000; val_acc: 0.213000
(Iteration 51 / 475) loss: 1.958274
(Epoch 1 / 5) train acc: 0.482000; val_acc: 0.471000
(Iteration 101 / 475) loss: 1.859804
(Iteration 151 / 475) loss: 1.698061
(Epoch 2 / 5) train acc: 0.535000; val_acc: 0.475000
(Iteration 201 / 475) loss: 1.753984
(Iteration 251 / 475) loss: 1.598714
(Epoch 3 / 5) train acc: 0.509000; val_acc: 0.507000
(Iteration 301 / 475) loss: 1.681604
(Iteration 351 / 475) loss: 1.649953
(Epoch 4 / 5) train acc: 0.553000; val_acc: 0.517000
(Iteration 401 / 475) loss: 1.651559
(Iteration 451 / 475) loss: 1.610692
(Epoch 5 / 5) train acc: 0.549000; val_acc: 0.518000
--------------------------------------------------------------------------------
FINAL RESULTS:
Validation accuracy: 0.51800
Test accuracy: 0.53400
```

Test Your Model!

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

In [19]:

```
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:  0.518
Test set accuracy:  0.534
```