knn

April 12, 2025

[ ]:
```
# 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/assignment1/'
FOLDERNAME = 'cs231n/assignments/assignment1/'
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
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
/content/drive/My Drive/cs231n/assignments/assignment1/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment1
```

1 k-Nearest Neighbor (kNN) exercise

Complete and hand in this completed worksheet (including its outputs and any supporting code outside of the worksheet) with your assignment submission. For more details see the assignments page on the course website.

The kNN classifier consists of two stages:

- During training, the classifier takes the training data and simply remembers it
- During testing, kNN classifies every test image by comparing to all training images and transfering the labels of the k most similar training examples
- The value of k is cross-validated

In this exercise you will implement these steps and understand the basic Image Classification pipeline, cross-validation, and gain proficiency in writing efficient, vectorized code.

[ ]:
```
# Run some setup code for this notebook.

import random
import numpy as np
from cs231n.data_utils import load_CIFAR10
import matplotlib.pyplot as plt

# This is a bit of magic to make matplotlib figures appear inline in the␣
 ↪notebook
# rather than in a new window.
%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'

# Some more magic so that the notebook will reload external python modules;
# see http://stackoverflow.com/questions/1907993/
 ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

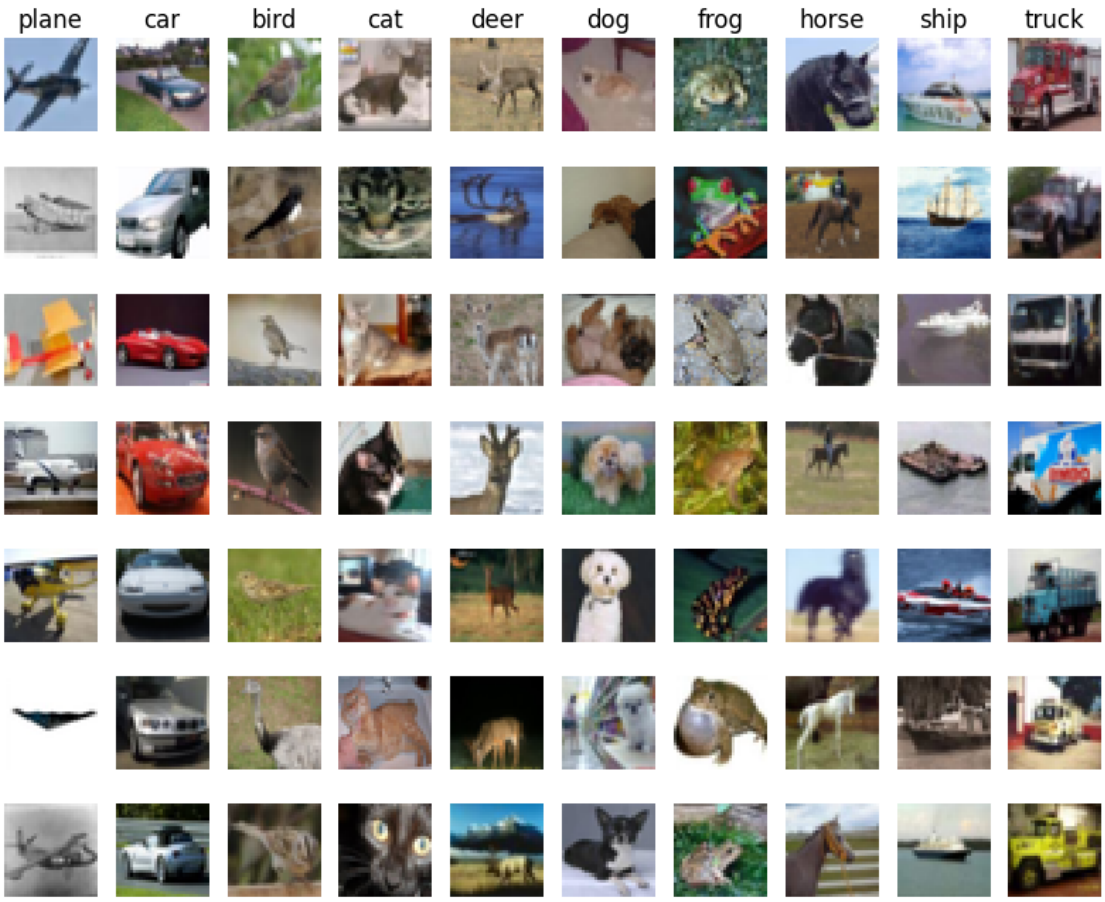
```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

[ ]:
```
# Load the raw CIFAR-10 data.
cifar10_dir = 'cs231n/datasets/cifar-10-batches-py'

# Cleaning up variables to prevent loading data multiple times (which may cause␣
 ↪memory issue)
try:
   del X_train, y_train
   del X_test, y_test
   print('Clear previously loaded data.')
except:
   pass

X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

# As a sanity check, we print out the size of the training and test data.
print('Training data shape: ', X_train.shape)
print('Training labels shape: ', y_train.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
```

```
Clear previously loaded data.
```

```
Training data shape:  (50000, 32, 32, 3)
Training labels shape:  (50000,)
Test data shape:  (10000, 32, 32, 3)
Test labels shape:  (10000,)
```

[ ]:
```
# Visualize some examples from the dataset.
# We show a few examples of training images from each class.
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse',␣
 ↪'ship', 'truck']
num_classes = len(classes)
samples_per_class = 7
for y, cls in enumerate(classes):
    idxs = np.flatnonzero(y_train == y)
    idxs = np.random.choice(idxs, samples_per_class, replace=False)
    for i, idx in enumerate(idxs):
        plt_idx = i * num_classes + y + 1
        plt.subplot(samples_per_class, num_classes, plt_idx)
        plt.imshow(X_train[idx].astype('uint8'))
        plt.axis('off')
        if i == 0:
            plt.title(cls)
plt.show()
```



[ ]:
```
# Subsample the data for more efficient code execution in this exercise
num_training = 5000
mask = list(range(num_training))
X_train = X_train[mask]
y_train = y_train[mask]

num_test = 500
mask = list(range(num_test))
X_test = X_test[mask]
y_test = y_test[mask]

# Reshape the image data into rows
X_train = np.reshape(X_train, (X_train.shape[0], -1))
X_test = np.reshape(X_test, (X_test.shape[0], -1))
print(X_train.shape, X_test.shape)
```

```
(5000, 3072) (500, 3072)
```

[ ]:
```
from cs231n.classifiers import KNearestNeighbor

# Create a kNN classifier instance.
# Remember that training a kNN classifier is a noop:
# the Classifier simply remembers the data and does no further processing
classifier = KNearestNeighbor()
classifier.train(X_train, y_train)
```

We would now like to classify the test data with the kNN classifier. Recall that we can break down this process into two steps:

1. First we must compute the distances between all test examples and all train examples.
2. Given these distances, for each test example we find the k nearest examples and have them vote for the label

Lets begin with computing the distance matrix between all training and test examples. For example, if there are **Ntr** training examples and **Nte** test examples, this stage should result in a **Nte x Ntr** matrix where each element (i,j) is the distance between the i-th test and j-th train example.

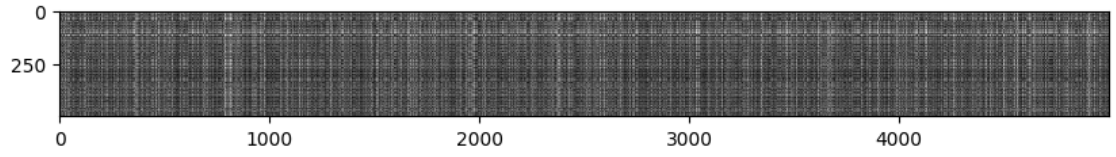
Note: For the three distance computations that we require you to implement in this notebook, you may not use the np.linalg.norm() function that numpy provides.

First, open `cs231n/classifiers/k_nearest_neighbor.py` and implement the function `compute_distances_two_loops` that uses a (very inefficient) double loop over all pairs of (test, train) examples and computes the distance matrix one element at a time.

[ ]:
```
# Open cs231n/classifiers/k_nearest_neighbor.py and implement
# compute_distances_two_loops.

# Test your implementation:
dists = classifier.compute_distances_two_loops(X_test)
print(dists.shape)
```

```
(500, 5000)
```

[ ]:
```
# We can visualize the distance matrix: each row is a single test example and
# its distances to training examples
plt.imshow(dists, interpolation='none')
plt.show()
```



Inline Question 1

Notice the structured patterns in the distance matrix, where some rows or columns are visibly brighter. (Note that with the default color scheme black indicates low distances while white indicates high distances.)

- What in the data is the cause behind the distinctly bright rows?
- What causes the columns?

YourAnswer : 1. **Bright Rows:** These indicate that a particular training example is far from many other training examples. This could be due to that example being an outlier or having features that are quite different from the majority of the training data. 2. **Bright Columns:** Similar to bright rows, a bright column indicates that a particular training example is far from many other training examples. This is also likely due to the example being an outlier or having unique features that set it apart from the rest of the training data.

[ ]:
```
# Now implement the function predict_labels and run the code below:
# We use k = 1 (which is Nearest Neighbor).
y_test_pred = classifier.predict_labels(dists, k=1)

# Compute and print the fraction of correctly predicted examples
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test, accuracy))
```

```
Got 137 / 500 correct => accuracy: 0.274000
```

You should expect to see approximately 27% accuracy. Now lets try out a larger `k`, say `k = 5`:

[ ]:
```
y_test_pred = classifier.predict_labels(dists, k=5)
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test, accuracy))
```

```
Got 145 / 500 correct => accuracy: 0.290000
```

You should expect to see a slightly better performance than with `k = 1`.

Inline Question 2

We can also use other distance metrics such as L1 distance. For pixel values $p_{ij}^{(k)}$ at location (i, j) of some image I_k,

the mean μ across all pixels over all images is

$$\mu = \frac{1}{nhw} \sum_{k=1}^{n} \sum_{i=1}^{h} \sum_{j=1}^{w} p_{ij}^{(k)}$$

And the pixel-wise mean μ_{ij} across all images is

$$\mu_{ij} = \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)}.$$

The general standard deviation σ and pixel-wise standard deviation σ_{ij} is defined similarly.

Which of the following preprocessing steps will not change the performance of a Nearest Neighbor classifier that uses L1 distance? Select all that apply. To clarify, both training and test examples are preprocessed in the same way.

1. Subtracting the mean μ ($\tilde{p}_{ij}^{(k)} = p_{ij}^{(k)} - \mu$.)
2. Subtracting the per pixel mean μ_{ij} ($\tilde{p}_{ij}^{(k)} = p_{ij}^{(k)} - \mu_{ij}$.)
3. Subtracting the mean μ and dividing by the standard deviation σ.
4. Subtracting the pixel-wise mean μ_{ij} and dividing by the pixel-wise standard deviation σ_{ij}.
5. Rotating the coordinate axes of the data, which means rotating all the images by the same angle. Empty regions in the image caused by rotation are padded with a same pixel value and no interpolation is performed.

YourAnswer : 1, 2, 5

YourExplanation : 1. Subtracting the mean μ: This shifts all pixel values by the same constant μ. Since L1 distance measures the absolute difference between pixel values, subtracting a constant from all values does not change the relative distances between images. **Performance remains unchanged.**

2. Subtracting the per pixel mean μ_{ij}: This shifts each pixel value by a different constant specific to that pixel location. This changes the absolute values but not the relative differences between images, so the L1 distance remains the same. **Performance remains unchanged.**
3. Subtracting the mean μ and dividing by the standard deviation σ: This step normalizes the data by both shifting and scaling. Scaling changes the relative distances between images, which affects the L1 distance. **Performance changes.**
4. Subtracting the pixel-wise mean μ_{ij} and dividing by the pixel-wise standard deviation σ_{ij}: This step normalizes each pixel individually by both shifting and scaling. Scaling changes the relative differences between images, which affects the L1 distance. **Performance changes.**
5. Rotating the coordinate axes of the data: Rotating all images by the same angle and padding empty regions with a constant pixel value does not change the relative distances between images in terms of L1 distance. The absolute positions of the pixels change, but the differences between corresponding pixels remain the same. **Performance remains unchanged.**

[ ]:
```
# Now lets speed up distance matrix computation by using partial vectorization
# with one loop. Implement the function compute_distances_one_loop and run the
# code below:
dists_one = classifier.compute_distances_one_loop(X_test)

# To ensure that our vectorized implementation is correct, we make sure that it
# agrees with the naive implementation. There are many ways to decide whether
# two matrices are similar; one of the simplest is the Frobenius norm. In case
# you haven't seen it before, the Frobenius norm of two matrices is the square
# root of the squared sum of differences of all elements; in other words, reshape
# the matrices into vectors and compute the Euclidean distance between them.
difference = np.linalg.norm(dists - dists_one, ord='fro')
print('One loop difference was: %f' % (difference, ))
if difference < 0.001:
```

```
    print('Good! The distance matrices are the same')
else:
    print('Uh-oh! The distance matrices are different')
```

```
One loop difference was: 0.000000
Good! The distance matrices are the same
```

[26]:
```
# Now implement the fully vectorized version inside compute_distances_no_loops
# and run the code
dists_two = classifier.compute_distances_no_loops(X_test)

# check that the distance matrix agrees with the one we computed before:
difference = np.linalg.norm(dists - dists_two, ord='fro')
print('No loop difference was: %f' % (difference, ))
if difference < 0.001:
    print('Good! The distance matrices are the same')
else:
    print('Uh-oh! The distance matrices are different')
```

```
No loop difference was: 0.000000
Good! The distance matrices are the same
```

[27]:
```
# Let's compare how fast the implementations are
def time_function(f, *args):
    """
    Call a function f with args and return the time (in seconds) that it took␣
 ↪to execute.
    """
    import time
    tic = time.time()
    f(*args)
    toc = time.time()
    return toc - tic

two_loop_time = time_function(classifier.compute_distances_two_loops, X_test)
print('Two loop version took %f seconds' % two_loop_time)

one_loop_time = time_function(classifier.compute_distances_one_loop, X_test)
print('One loop version took %f seconds' % one_loop_time)

no_loop_time = time_function(classifier.compute_distances_no_loops, X_test)
print('No loop version took %f seconds' % no_loop_time)

# You should see significantly faster performance with the fully vectorized␣
 ↪implementation!

# NOTE: depending on what machine you're using,
```

```
# you might not see a speedup when you go from two loops to one loop,
# and might even see a slow-down.
```

```
Two loop version took 40.836344 seconds
One loop version took 56.832551 seconds
No loop version took 0.598151 seconds
```

1.0.1 Cross-validation

We have implemented the k-Nearest Neighbor classifier but we set the value k = 5 arbitrarily. We will now determine the best value of this hyperparameter with cross-validation.

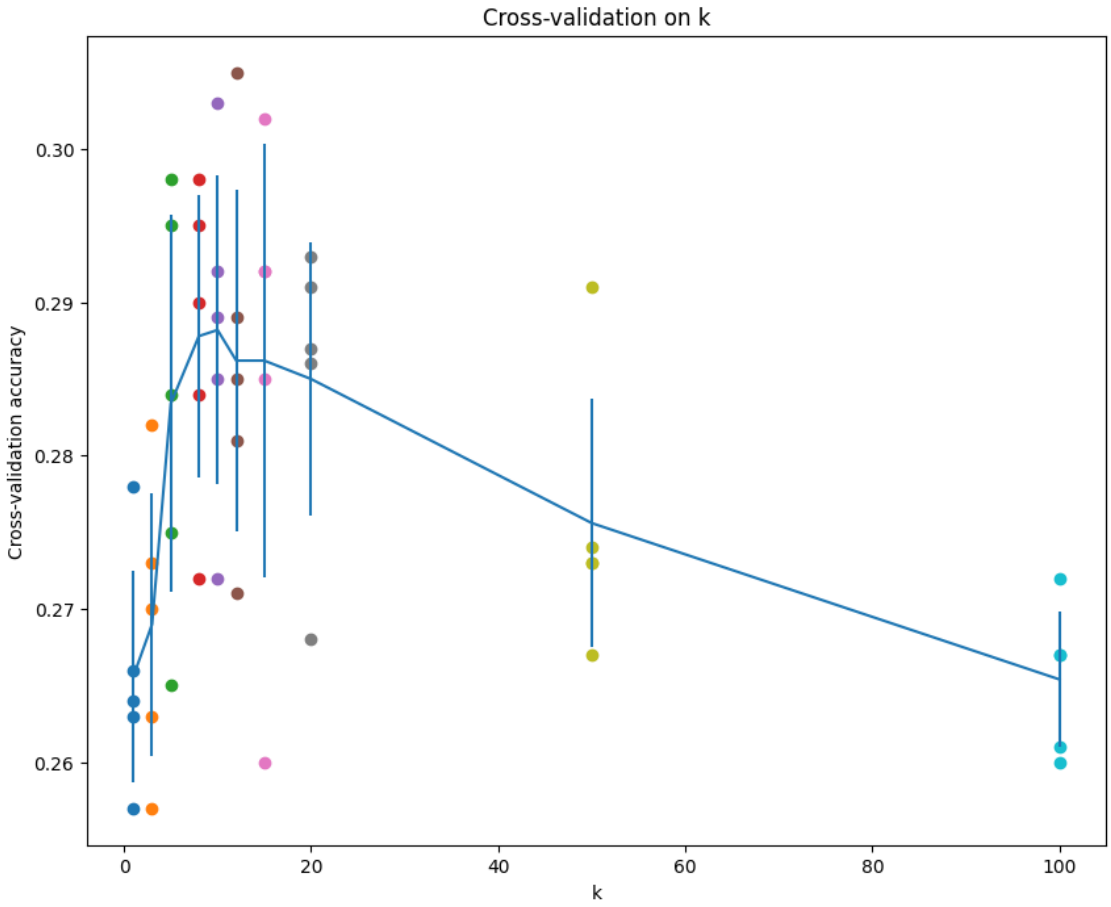
[42]:
```
num_folds = 5
k_choices = [1, 3, 5, 8, 10, 12, 15, 20, 50, 100]

X_train_folds = np.split(X_train, num_folds)
y_train_folds = np.split(y_train, num_folds)
################################################################################
# TODO:                                                                        #
# Split up the training data into folds. After splitting, X_train_folds and    #
# y_train_folds should each be lists of length num_folds, where                #
# y_train_folds[i] is the label vector for the points in X_train_folds[i].     #
# Hint: Look up the numpy array_split function.                                #
################################################################################


# A dictionary holding the accuracies for different values of k that we find
# when running cross-validation. After running cross-validation,
# k_to_accuracies[k] should be a list of length num_folds giving the different
# accuracy values that we found when using that value of k.
k_to_accuracies = {}


################################################################################
# TODO:                                                                        #
# Perform k-fold cross validation to find the best value of k. For each        #
# possible value of k, run the k-nearest-neighbor algorithm num_folds times,   #
# where in each case you use all but one of the folds as training data and the #
# last fold as a validation set. Store the accuracies for all fold and all     #
# values of k in the k_to_accuracies dictionary.                               #
################################################################################
for k in k_choices:
    k_to_accuracies[k] = []
    for fold in range(num_folds):
        # Prepare the training and validation sets
        X_train_cv = np.concatenate([X_train_folds[i] for i in range(num_folds)␣
 ↪if i != fold])
```

```
        y_train_cv = np.concatenate([y_train_folds[i] for i in range(num_folds)␣
 ↪if i != fold])
        X_val_cv = X_train_folds[fold]
        y_val_cv = y_train_folds[fold]

        # Train the k-Nearest Neighbor classifier
        classifier = KNearestNeighbor()
        classifier.train(X_train_cv, y_train_cv)

        # Evaluate the classifier on the validation set
        y_val_pred = classifier.predict(X_val_cv, k)
        accuracy = np.mean(y_val_pred == y_val_cv)
        k_to_accuracies[k].append(accuracy)

# Print out the computed accuracies
for k in sorted(k_to_accuracies):
    for accuracy in k_to_accuracies[k]:
        print('k = %d, accuracy = %f' % (k, accuracy))
```

```
k = 1, accuracy = 0.263000
k = 1, accuracy = 0.257000
k = 1, accuracy = 0.264000
k = 1, accuracy = 0.278000
k = 1, accuracy = 0.266000
k = 3, accuracy = 0.257000
k = 3, accuracy = 0.263000
k = 3, accuracy = 0.273000
k = 3, accuracy = 0.282000
k = 3, accuracy = 0.270000
k = 5, accuracy = 0.265000
k = 5, accuracy = 0.275000
k = 5, accuracy = 0.295000
k = 5, accuracy = 0.298000
k = 5, accuracy = 0.284000
k = 8, accuracy = 0.272000
k = 8, accuracy = 0.295000
k = 8, accuracy = 0.284000
k = 8, accuracy = 0.298000
k = 8, accuracy = 0.290000
k = 10, accuracy = 0.272000
k = 10, accuracy = 0.303000
k = 10, accuracy = 0.289000
k = 10, accuracy = 0.292000
k = 10, accuracy = 0.285000
k = 12, accuracy = 0.271000
k = 12, accuracy = 0.305000
k = 12, accuracy = 0.285000
```

```
k = 12, accuracy = 0.289000
k = 12, accuracy = 0.281000
k = 15, accuracy = 0.260000
k = 15, accuracy = 0.302000
k = 15, accuracy = 0.292000
k = 15, accuracy = 0.292000
k = 15, accuracy = 0.285000
k = 20, accuracy = 0.268000
k = 20, accuracy = 0.293000
k = 20, accuracy = 0.291000
k = 20, accuracy = 0.287000
k = 20, accuracy = 0.286000
k = 50, accuracy = 0.273000
k = 50, accuracy = 0.291000
k = 50, accuracy = 0.274000
k = 50, accuracy = 0.267000
k = 50, accuracy = 0.273000
k = 100, accuracy = 0.261000
k = 100, accuracy = 0.272000
k = 100, accuracy = 0.267000
k = 100, accuracy = 0.260000
k = 100, accuracy = 0.267000
```

[43]:
```
# plot the raw observations
for k in k_choices:
    accuracies = k_to_accuracies[k]
    plt.scatter([k] * len(accuracies), accuracies)

# plot the trend line with error bars that correspond to standard deviation
accuracies_mean = np.array([np.mean(v) for k,v in sorted(k_to_accuracies.
 ↪items())])
accuracies_std = np.array([np.std(v) for k,v in sorted(k_to_accuracies.
 ↪items())])
plt.errorbar(k_choices, accuracies_mean, yerr=accuracies_std)
plt.title('Cross-validation on k')
plt.xlabel('k')
plt.ylabel('Cross-validation accuracy')
plt.show()
```



[47]:
```
# Based on the cross-validation results above, choose the best value for k,
# retrain the classifier using all the training data, and test it on the test
# data. You should be able to get above 28% accuracy on the test data.
best_k = 8

classifier = KNearestNeighbor()
classifier.train(X_train, y_train)
y_test_pred = classifier.predict(X_test, k=best_k)

# Compute and display the accuracy
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test, accuracy))
```

```
Got 147 / 500 correct => accuracy: 0.294000
```

Inline Question 3

Which of the following statements about k-Nearest Neighbor (k-NN) are true in a classification setting, and for all k? Select all that apply. 1. The decision boundary of the k-NN classifier is

linear. 2. The training error of a 1-NN will always be lower than or equal to that of 5-NN. 3. The test error of a 1-NN will always be lower than that of a 5-NN. 4. The time needed to classify a test example with the k-NN classifier grows with the size of the training set. 5. None of the above.

YourAnswer :

1, 4

YourExplanation : 1. **False:** The decision boundary of the k-NN classifier is not necessarily linear. In fact, it can be highly non-linear and complex, depending on the distribution of the training data and the value of k. The decision boundary is determined by the regions around the training points and how they vote for the class labels.

2. **True:** The training error for 1-NN is typically very low because each training point is its own nearest neighbor, leading to perfect classification on the training set. For 5-NN, the training error can be higher because the decision is based on the majority vote of the 5 nearest neighbors, which might not always include the point itself.
3. **False:** The test error of 1-NN is not always lower than that of 5-NN. While 1-NN might perform better on the training data, it can overfit and perform poorly on unseen test data. 5-NN, by considering more neighbors, can generalize better and have lower test error in many cases.
4. **True:** The time complexity of the k-NN classifier for classifying a test example is O(n d), where n is the number of training examples and d is the number of features. As the size of the training set increases, the time needed to compute the distances to all training points and find the k nearest neighbors also increases.

softmax

April 12, 2025

[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/assignment1/'
FOLDERNAME = 'cs231n/assignments/assignment1/'
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/assignment1/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment1
```

1 Softmax Classifier exercise

Complete and hand in this completed worksheet (including its outputs and any supporting code outside of the worksheet) with your assignment submission. For more details see the assignments page on the course website.

In this exercise you will:

- implement a fully-vectorized **loss function** for the Softmax classifier.
- implement the fully-vectorized expression for its **analytic gradient**
- **check your implementation** using numerical gradient
- use a validation set to **tune the learning rate and regularization** strength
- **optimize** the loss function with **SGD**
- **visualize** the final learned weights

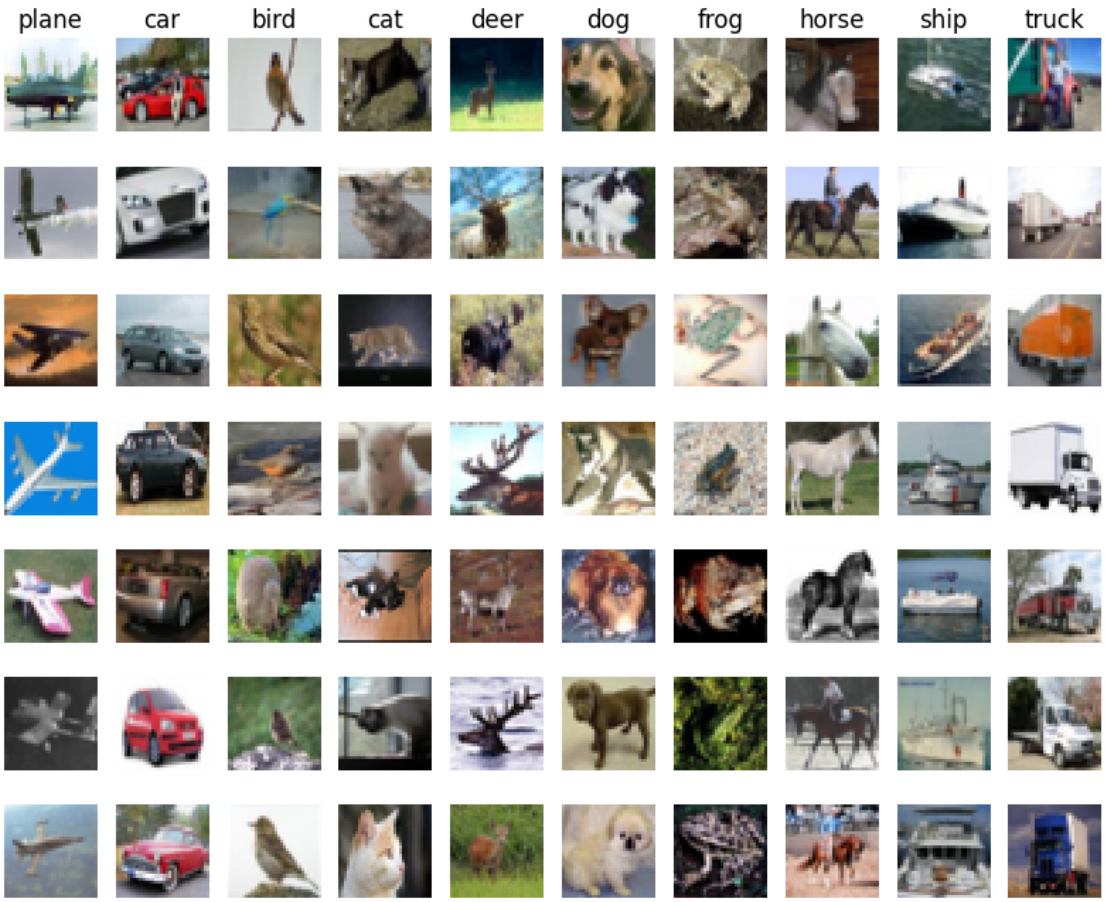
[2]:
```
# Run some setup code for this notebook.
import random
import numpy as np
from cs231n.data_utils import load_CIFAR10
import matplotlib.pyplot as plt

# This is a bit of magic to make matplotlib figures appear inline in the
# notebook rather than in a new window.
%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'

# Some more magic so that the notebook will reload external python modules;
# see http://stackoverflow.com/questions/1907993/
 ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

1.1 CIFAR-10 Data Loading and Preprocessing

[3]:
```
# Load the raw CIFAR-10 data.
cifar10_dir = 'cs231n/datasets/cifar-10-batches-py'

# Cleaning up variables to prevent loading data multiple times (which may cause␣
 ↪memory issue)
try:
   del X_train, y_train
   del X_test, y_test
   print('Clear previously loaded data.')
except:
   pass

X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

# As a sanity check, we print out the size of the training and test data.
print('Training data shape: ', X_train.shape)
print('Training labels shape: ', y_train.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
```

```
Training data shape:  (50000, 32, 32, 3)
Training labels shape:  (50000,)
Test data shape:  (10000, 32, 32, 3)
Test labels shape:  (10000,)
```

[4]:
```
# Visualize some examples from the dataset.
# We show a few examples of training images from each class.
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse',␣
 ↪'ship', 'truck']
num_classes = len(classes)
samples_per_class = 7
for y, cls in enumerate(classes):
    idxs = np.flatnonzero(y_train == y)
    idxs = np.random.choice(idxs, samples_per_class, replace=False)
    for i, idx in enumerate(idxs):
        plt_idx = i * num_classes + y + 1
        plt.subplot(samples_per_class, num_classes, plt_idx)
        plt.imshow(X_train[idx].astype('uint8'))
        plt.axis('off')
        if i == 0:
            plt.title(cls)
plt.show()
```



[5]:
```
# Split the data into train, val, and test sets. In addition we will
# create a small development set as a subset of the training data;
# we can use this for development so our code runs faster.
num_training = 49000
num_validation = 1000
num_test = 1000
num_dev = 500

# Our validation set will be num_validation points from the original
# training set.
mask = range(num_training, num_training + num_validation)
X_val = X_train[mask]
y_val = y_train[mask]

# Our training set will be the first num_train points from the original
# training set.
mask = range(num_training)
X_train = X_train[mask]
y_train = y_train[mask]

# We will also make a development set, which is a small subset of
# the training set.
mask = np.random.choice(num_training, num_dev, replace=False)
X_dev = X_train[mask]
y_dev = y_train[mask]

# We use the first num_test points of the original test set as our
# test set.
mask = range(num_test)
X_test = X_test[mask]
y_test = y_test[mask]

print('Train data shape: ', X_train.shape)
print('Train labels shape: ', y_train.shape)
print('Validation data shape: ', X_val.shape)
print('Validation labels shape: ', y_val.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
```

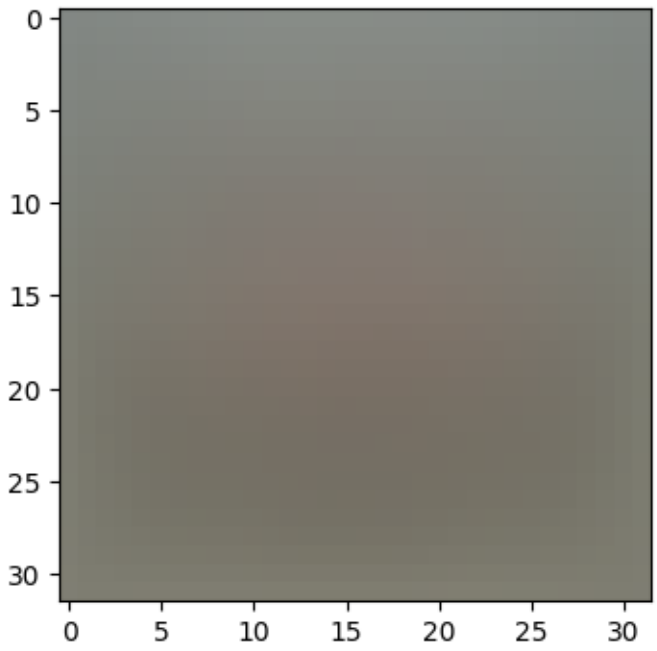
```
Train data shape:  (49000, 32, 32, 3)
Train labels shape:  (49000,)
Validation data shape:  (1000, 32, 32, 3)
Validation labels shape:  (1000,)
Test data shape:  (1000, 32, 32, 3)
Test labels shape:  (1000,)
```

[6]:
```
# Preprocessing: reshape the image data into rows
X_train = np.reshape(X_train, (X_train.shape[0], -1))
X_val = np.reshape(X_val, (X_val.shape[0], -1))
X_test = np.reshape(X_test, (X_test.shape[0], -1))
X_dev = np.reshape(X_dev, (X_dev.shape[0], -1))

# As a sanity check, print out the shapes of the data
print('Training data shape: ', X_train.shape)
print('Validation data shape: ', X_val.shape)
print('Test data shape: ', X_test.shape)
print('dev data shape: ', X_dev.shape)
```

```
Training data shape:  (49000, 3072)
Validation data shape:  (1000, 3072)
Test data shape:  (1000, 3072)
dev data shape:  (500, 3072)
```

[7]:
```
# Preprocessing: subtract the mean image
# first: compute the image mean based on the training data
mean_image = np.mean(X_train, axis=0)
print(mean_image[:10]) # print a few of the elements
plt.figure(figsize=(4,4))
plt.imshow(mean_image.reshape((32,32,3)).astype('uint8')) # visualize the mean␣
 ↪image
plt.show()

# second: subtract the mean image from train and test data
X_train -= mean_image
X_val -= mean_image
X_test -= mean_image
X_dev -= mean_image

# third: append the bias dimension of ones (i.e. bias trick) so that our␣
 ↪classifier
# only has to worry about optimizing a single weight matrix W.
X_train = np.hstack([X_train, np.ones((X_train.shape[0], 1))])
X_val = np.hstack([X_val, np.ones((X_val.shape[0], 1))])
X_test = np.hstack([X_test, np.ones((X_test.shape[0], 1))])
X_dev = np.hstack([X_dev, np.ones((X_dev.shape[0], 1))])

print(X_train.shape, X_val.shape, X_test.shape, X_dev.shape)
```

```
[130.64189796 135.98173469 132.47391837 130.05569388 135.34804082
 131.75402041 130.96055102 136.14328571 132.47636735 131.48467347]
```




```
(49000, 3073) (1000, 3073) (1000, 3073) (500, 3073)
```

1.2 Softmax Classifier

Your code for this section will all be written inside `cs231n/classifiers/softmax.py`.

As you can see, we have prefilled the function `softmax_loss_naive` which uses for loops to evaluate the softmax loss function.

[8]:
```
# Evaluate the naive implementation of the loss we provided for you:
from cs231n.classifiers.softmax import softmax_loss_naive
import time

# generate a random Softmax classifier weight matrix of small numbers
W = np.random.randn(3073, 10) * 0.0001

loss, grad = softmax_loss_naive(W, X_dev, y_dev, 0.000005)
print('loss: %f' % (loss, ))

# As a rough sanity check, our loss should be something close to -log(0.1).
print('loss: %f' % loss)
print('sanity check: %f' % (-np.log(0.1)))
```

```
loss: 2.339684
loss: 2.339684
sanity check: 2.302585
```

Inline Question 1

Why do we expect our loss to be close to -log(0.1)? **Explain briefly.**

Your Answer :

1. The model is initialized with small random weights, leading to random guesses.
2. For a 10-class classification problem, the probability of the correct class is 0.1.
3. The expected loss for random guessing is $-\log(0.1)=1$.

The `grad` returned from the function above is right now all zero. Derive and implement the gradient for the softmax loss function and implement it inline inside the function `softmax_loss_naive`. You will find it helpful to interleave your new code inside the existing function.

To check that you have correctly implemented the gradient, you can numerically estimate the gradient of the loss function and compare the numeric estimate to the gradient that you computed. We have provided code that does this for you:

[9]:
```
# Once you've implemented the gradient, recompute it with the code below
# and gradient check it with the function we provided for you

# Compute the loss and its gradient at W.
loss, grad = softmax_loss_naive(W, X_dev, y_dev, 0.0)

# Numerically compute the gradient along several randomly chosen dimensions, and
# compare them with your analytically computed gradient. The numbers should␣
 ↪match
# almost exactly along all dimensions.
from cs231n.gradient_check import grad_check_sparse
f = lambda w: softmax_loss_naive(w, X_dev, y_dev, 0.0)[0]
grad_numerical = grad_check_sparse(f, W, grad)

# do the gradient check once again with regularization turned on
# you didn't forget the regularization gradient did you?
loss, grad = softmax_loss_naive(W, X_dev, y_dev, 5e1)
f = lambda w: softmax_loss_naive(w, X_dev, y_dev, 5e1)[0]
grad_numerical = grad_check_sparse(f, W, grad)
```

```
numerical: -0.137457 analytic: -0.137457, relative error: 1.730748e-07
numerical: 2.402560 analytic: 2.402559, relative error: 2.880371e-08
numerical: -1.711513 analytic: -1.711513, relative error: 5.963209e-09
numerical: -3.985068 analytic: -3.985068, relative error: 7.788099e-09
numerical: -0.236315 analytic: -0.236315, relative error: 2.652051e-07
numerical: 1.898034 analytic: 1.898034, relative error: 2.042329e-08
numerical: 1.074734 analytic: 1.074734, relative error: 1.783367e-08
numerical: -3.177368 analytic: -3.177368, relative error: 1.058983e-08
numerical: 0.859727 analytic: 0.859727, relative error: 3.508874e-08
numerical: 3.442533 analytic: 3.442533, relative error: 1.519492e-08
numerical: -0.380668 analytic: -0.380668, relative error: 5.603591e-08
```

```
numerical: -0.941712 analytic: -0.941712, relative error: 9.656213e-09
numerical: -1.239438 analytic: -1.239438, relative error: 2.016861e-08
numerical: -0.410752 analytic: -0.410752, relative error: 6.116438e-09
numerical: -1.804399 analytic: -1.804399, relative error: 3.681555e-09
numerical: 1.798550 analytic: 1.798550, relative error: 1.284486e-08
numerical: -0.673201 analytic: -0.673201, relative error: 7.030877e-08
numerical: 0.582125 analytic: 0.582125, relative error: 8.047786e-09
numerical: 1.987867 analytic: 1.987867, relative error: 2.235821e-08
numerical: 1.289384 analytic: 1.289384, relative error: 5.476163e-09
```

Inline Question 2

Although gradcheck is reliable softmax loss, it is possible that for SVM loss, once in a while, a dimension in the gradcheck will not match exactly. What could such a discrepancy be caused by? Is it a reason for concern? What is a simple example in one dimension where a svm loss gradient check could fail? How would change the margin affect of the frequency of this happening?

Note that SVM loss for a sample (x_i, y_i) is defined as:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

where j iterates over all classes except the correct class y_i and s_j denotes the classifier score for j^{th} class. Δ is a scalar margin. For more information, refer to 'Multiclass Support Vector Machine loss' on this page.

Hint: the SVM loss function is not strictly speaking differentiable.

YourAnswer : 1. The discrepancy in gradient checks for the SVM loss function can be caused by the non-differentiability of the SVM loss function at certain points. Specifically, the SVM loss function is not differentiable when $s_j - s_{y_i} + \Delta = 0$. At these points, the gradient is not well-defined, leading to potential mismatches between the numerical and analytical gradients.

2. While the non-differentiability of the SVM loss function can lead to occasional gradient check failures, these points are typically rare in practice. As long as the number of such points is not significant, it generally does not affect the overall training of the model. Therefore, these discrepancies are usually not a major cause for concern.

3. Consider a simple one-dimensional example where the SVM loss function is evaluated at a point where $s_j - s_{y_i} + \Delta = 0$. For instance:

   $$s_j = 1$$

   $$s_{y_i} = 1$$

   $$\Delta = 0$$

   At this point, the loss function is:

   $$L_i = \max(0, 1 - 1 + 0) = 0$$

   Here, the analytical gradient is 0, but the numerical gradient might differ due to the choice of h. If h is too large, the numerical gradient might not accurately reflect the true gradient.

4. Increasing the margin Δ increases the number of points where the loss function is non-differentiable. This is because more terms $s_j - s_{y_i} + \Delta$ will equal zero, leading to more points where the gradient is not well-defined. Consequently, this increases the likelihood of gradient check failures.

[10]:
```
# Next implement the function softmax_loss_vectorized; for now only compute the loss;
# we will implement the gradient in a moment.
tic = time.time()
loss_naive, grad_naive = softmax_loss_naive(W, X_dev, y_dev, 0.000005)
toc = time.time()
print('Naive loss: %e computed in %fs' % (loss_naive, toc - tic))

from cs231n.classifiers.softmax import softmax_loss_vectorized
tic = time.time()
loss_vectorized, _ = softmax_loss_vectorized(W, X_dev, y_dev, 0.000005)
toc = time.time()
print('Vectorized loss: %e computed in %fs' % (loss_vectorized, toc - tic))

# The losses should match but your vectorized implementation should be much faster.
print('difference: %f' % (loss_naive - loss_vectorized))
```

```
Naive loss: 2.339684e+00 computed in 0.055363s
Vectorized loss: 2.339684e+00 computed in 0.010352s
difference: 0.000000
```

[11]:
```
# Complete the implementation of softmax_loss_vectorized, and compute the gradient
# of the loss function in a vectorized way.

# The naive implementation and the vectorized implementation should match, but
# the vectorized version should still be much faster.
tic = time.time()
_, grad_naive = softmax_loss_vectorized(W, X_dev, y_dev, 0.000005)
toc = time.time()
print('Naive loss and gradient: computed in %fs' % (toc - tic))

tic = time.time()
_, grad_vectorized = softmax_loss_vectorized(W, X_dev, y_dev, 0.000005)
toc = time.time()
print('Vectorized loss and gradient: computed in %fs' % (toc - tic))

# The loss is a single number, so it is easy to compare the values computed
# by the two implementations. The gradient on the other hand is a matrix, so
# we use the Frobenius norm to compare them.
difference = np.linalg.norm(grad_naive - grad_vectorized, ord='fro')
print('difference: %f' % difference)
```

```
Naive loss and gradient: computed in 0.006123s
Vectorized loss and gradient: computed in 0.006442s
difference: 0.000000
```

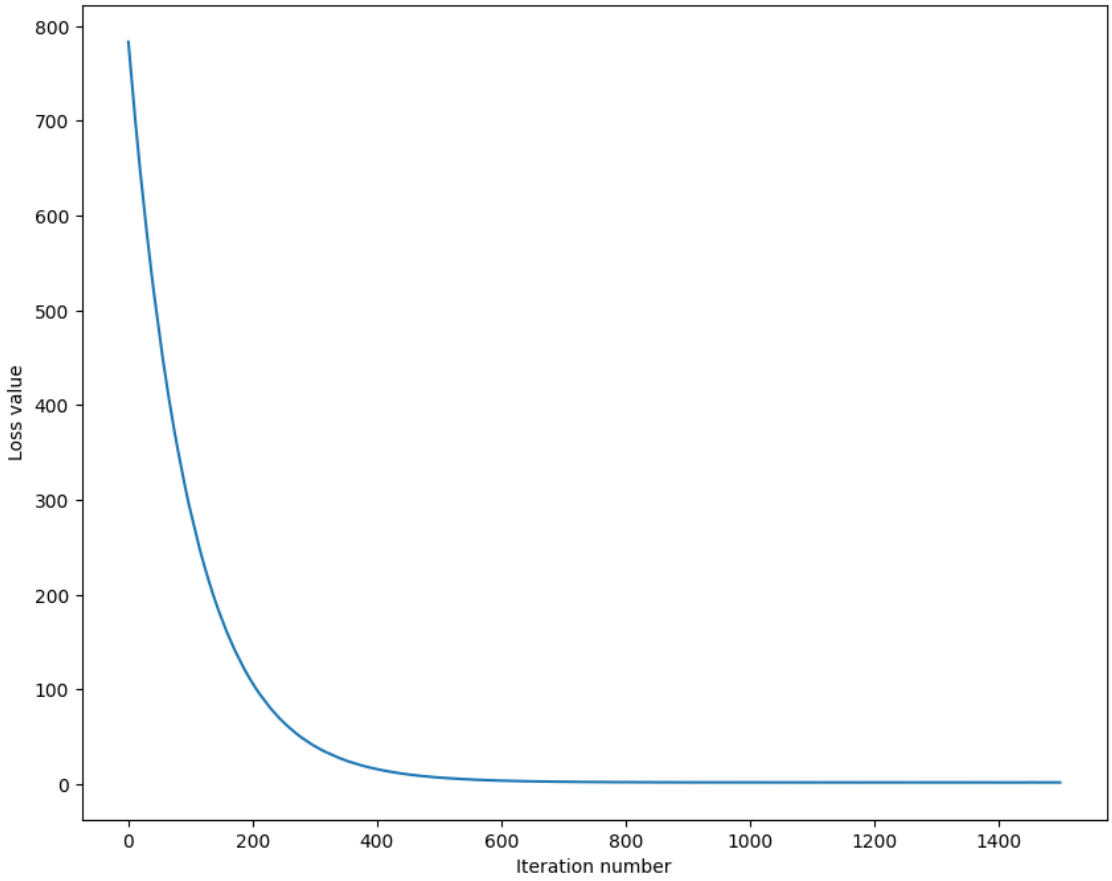
1.2.1 Stochastic Gradient Descent

We now have vectorized and efficient expressions for the loss, the gradient and our gradient matches the numerical gradient. We are therefore ready to do SGD to minimize the loss. Your code for this part will be written inside `cs231n/classifiers/linear_classifier.py`.

[12]:
```
# In the file linear_classifier.py, implement SGD in the function
# LinearClassifier.train() and then run it with the code below.
from cs231n.classifiers import Softmax
softmax = Softmax()
tic = time.time()
loss_hist = softmax.train(X_train, y_train, learning_rate=1e-7, reg=2.5e4,
                      num_iters=1500, verbose=True)
toc = time.time()
print('That took %fs' % (toc - tic))
```

```
iteration 0 / 1500: loss 783.388182
iteration 100 / 1500: loss 287.646097
iteration 200 / 1500: loss 106.446221
iteration 300 / 1500: loss 40.297482
iteration 400 / 1500: loss 16.183427
iteration 500 / 1500: loss 7.165475
iteration 600 / 1500: loss 3.937327
iteration 700 / 1500: loss 2.706771
iteration 800 / 1500: loss 2.378952
iteration 900 / 1500: loss 2.180463
iteration 1000 / 1500: loss 2.108417
iteration 1100 / 1500: loss 2.075895
iteration 1200 / 1500: loss 2.110284
iteration 1300 / 1500: loss 2.090869
iteration 1400 / 1500: loss 2.079732
That took 5.863721s
```

[13]:
```
# A useful debugging strategy is to plot the loss as a function of
# iteration number:
plt.plot(loss_hist)
plt.xlabel('Iteration number')
plt.ylabel('Loss value')
plt.show()
```



[14]:
```
# Write the LinearClassifier.predict function and evaluate the performance on
# both the training and validation set
# You should get validation accuracy of about 0.34 (> 0.33).
y_train_pred = softmax.predict(X_train)
print('training accuracy: %f' % (np.mean(y_train == y_train_pred), ))
y_val_pred = softmax.predict(X_val)
print('validation accuracy: %f' % (np.mean(y_val == y_val_pred), ))
```

```
training accuracy: 0.330347
validation accuracy: 0.352000
```

[15]:
```
# Save the trained model for autograder.
softmax.save("softmax.npy")
```

```
softmax.npy saved.
```

[16]:
```
# Use the validation set to tune hyperparameters (regularization strength and
# learning rate). You should experiment with different ranges for the learning
# rates and regularization strengths; if you are careful you should be able to
```

```
# get a classification accuracy of about 0.365 (> 0.36) on the validation set.

# Note: you may see runtime/overflow warnings during hyper-parameter search.
# This may be caused by extreme values, and is not a bug.

# results is dictionary mapping tuples of the form
# (learning_rate, regularization_strength) to tuples of the form
# (training_accuracy, validation_accuracy). The accuracy is simply the fraction
# of data points that are correctly classified.
results = {}
best_val = -1   # The highest validation accuracy that we have seen so far.
best_softmax = None # The Softmax object that achieved the highest validation␣
 ↪rate.

################################################################################
# TODO:                                                                        #
# Write code that chooses the best hyperparameters by tuning on the validation #
# set. For each combination of hyperparameters, train a Softmax on the         #
# training set, compute its accuracy on the training and validation sets, and  #
# store these numbers in the results dictionary. In addition, store the best   #
# validation accuracy in best_val and the Softmax object that achieves this    #
# accuracy in best_softmax.                                                    #
#                                                                              #
# Hint: You should use a small value for num_iters as you develop your         #
# validation code so that the classifiers don't take much time to train; once  #
# you are confident that your validation code works, you should rerun the      #
# code with a larger value for num_iters.                                      #
################################################################################

# Provided as a reference. You may or may not want to change these␣
 ↪hyperparameters
learning_rates = [1e-8, 5e-8, 1e-7, 5e-7, 1e-6, 5e-6, 1e-5]
regularization_strengths = [1e3, 2.5e3, 5e3, 1e4, 2.5e4, 5e4, 1e5]

# Use a small number of iterations for initial debugging
num_iters =  2000  # You can increase this once your code is working

# Loop over all combinations of hyperparameters
for lr in learning_rates:
    for reg in regularization_strengths:
        # Create a new Softmax classifier
        softmax = Softmax()

        # Train the classifier with the current hyperparameters
        loss_history = softmax.train(X_train, y_train, learning_rate=lr,␣
 ↪reg=reg, num_iters=num_iters, verbose=False)
```

```
        # Compute training accuracy
        y_train_pred = softmax.predict(X_train)
        train_accuracy = np.mean(y_train == y_train_pred)

        # Compute validation accuracy
        y_val_pred = softmax.predict(X_val)
        val_accuracy = np.mean(y_val == y_val_pred)

        # Store the results
        results[(lr, reg)] = (train_accuracy, val_accuracy)

        # Update the best validation accuracy and the corresponding classifier
        if val_accuracy > best_val:
            best_val = val_accuracy
            best_softmax = softmax

# Print out results.
for lr, reg in sorted(results):
    train_accuracy, val_accuracy = results[(lr, reg)]
    print('lr %e reg %e train accuracy: %f val accuracy: %f' % (
                lr, reg, train_accuracy, val_accuracy))

print('best validation accuracy achieved during cross-validation: %f' %␣
  ↪best_val)
```

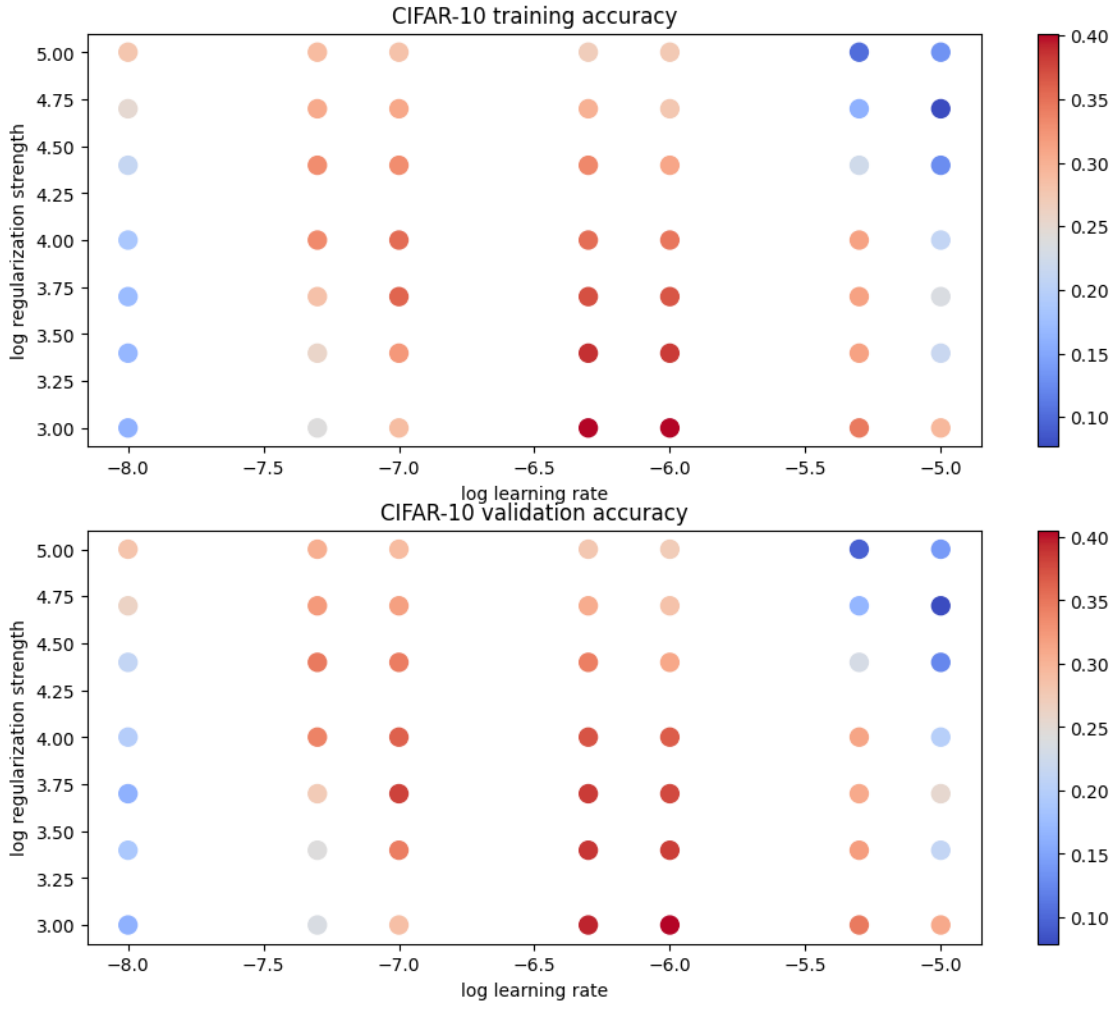
```
/content/drive/My
Drive/cs231n/assignments/assignment1/cs231n/classifiers/softmax.py:93:
RuntimeWarning: divide by zero encountered in log
  correct_logprobs = -np.log(probs[np.arange(num_train), y])
```

```
lr 1.000000e-08 reg 1.000000e+03 train accuracy: 0.159653 val accuracy: 0.163000
lr 1.000000e-08 reg 2.500000e+03 train accuracy: 0.165653 val accuracy: 0.189000
lr 1.000000e-08 reg 5.000000e+03 train accuracy: 0.171571 val accuracy: 0.163000
lr 1.000000e-08 reg 1.000000e+04 train accuracy: 0.184510 val accuracy: 0.198000
lr 1.000000e-08 reg 2.500000e+04 train accuracy: 0.211796 val accuracy: 0.212000
lr 1.000000e-08 reg 5.000000e+04 train accuracy: 0.249490 val accuracy: 0.263000
lr 1.000000e-08 reg 1.000000e+05 train accuracy: 0.278245 val accuracy: 0.284000
lr 5.000000e-08 reg 1.000000e+03 train accuracy: 0.240306 val accuracy: 0.236000
lr 5.000000e-08 reg 2.500000e+03 train accuracy: 0.256163 val accuracy: 0.243000
lr 5.000000e-08 reg 5.000000e+03 train accuracy: 0.282490 val accuracy: 0.274000
lr 5.000000e-08 reg 1.000000e+04 train accuracy: 0.332898 val accuracy: 0.341000
lr 5.000000e-08 reg 2.500000e+04 train accuracy: 0.329633 val accuracy: 0.347000
lr 5.000000e-08 reg 5.000000e+04 train accuracy: 0.306102 val accuracy: 0.323000
lr 5.000000e-08 reg 1.000000e+05 train accuracy: 0.289082 val accuracy: 0.305000
lr 1.000000e-07 reg 1.000000e+03 train accuracy: 0.287735 val accuracy: 0.288000
lr 1.000000e-07 reg 2.500000e+03 train accuracy: 0.321204 val accuracy: 0.344000
lr 1.000000e-07 reg 5.000000e+03 train accuracy: 0.357551 val accuracy: 0.382000
lr 1.000000e-07 reg 1.000000e+04 train accuracy: 0.352878 val accuracy: 0.364000
```

```
lr 1.000000e-07 reg 2.500000e+04 train accuracy: 0.330204 val accuracy: 0.345000
lr 1.000000e-07 reg 5.000000e+04 train accuracy: 0.307694 val accuracy: 0.318000
lr 1.000000e-07 reg 1.000000e+05 train accuracy: 0.281122 val accuracy: 0.292000
lr 5.000000e-07 reg 1.000000e+03 train accuracy: 0.401633 val accuracy: 0.396000
lr 5.000000e-07 reg 2.500000e+03 train accuracy: 0.386796 val accuracy: 0.388000
lr 5.000000e-07 reg 5.000000e+03 train accuracy: 0.370898 val accuracy: 0.385000
lr 5.000000e-07 reg 1.000000e+04 train accuracy: 0.351122 val accuracy: 0.371000
lr 5.000000e-07 reg 2.500000e+04 train accuracy: 0.333347 val accuracy: 0.343000
lr 5.000000e-07 reg 5.000000e+04 train accuracy: 0.299837 val accuracy: 0.307000
lr 5.000000e-07 reg 1.000000e+05 train accuracy: 0.267143 val accuracy: 0.280000
lr 1.000000e-06 reg 1.000000e+03 train accuracy: 0.400918 val accuracy: 0.405000
lr 1.000000e-06 reg 2.500000e+03 train accuracy: 0.381347 val accuracy: 0.384000
lr 1.000000e-06 reg 5.000000e+03 train accuracy: 0.366796 val accuracy: 0.378000
lr 1.000000e-06 reg 1.000000e+04 train accuracy: 0.345816 val accuracy: 0.366000
lr 1.000000e-06 reg 2.500000e+04 train accuracy: 0.309265 val accuracy: 0.312000
lr 1.000000e-06 reg 5.000000e+04 train accuracy: 0.276061 val accuracy: 0.285000
lr 1.000000e-06 reg 1.000000e+05 train accuracy: 0.273857 val accuracy: 0.273000
lr 5.000000e-06 reg 1.000000e+03 train accuracy: 0.343816 val accuracy: 0.347000
lr 5.000000e-06 reg 2.500000e+03 train accuracy: 0.312980 val accuracy: 0.320000
lr 5.000000e-06 reg 5.000000e+03 train accuracy: 0.313490 val accuracy: 0.309000
lr 5.000000e-06 reg 1.000000e+04 train accuracy: 0.313735 val accuracy: 0.314000
lr 5.000000e-06 reg 2.500000e+04 train accuracy: 0.223286 val accuracy: 0.232000
lr 5.000000e-06 reg 5.000000e+04 train accuracy: 0.159490 val accuracy: 0.168000
lr 5.000000e-06 reg 1.000000e+05 train accuracy: 0.100816 val accuracy: 0.095000
lr 1.000000e-05 reg 1.000000e+03 train accuracy: 0.292653 val accuracy: 0.310000
lr 1.000000e-05 reg 2.500000e+03 train accuracy: 0.215878 val accuracy: 0.212000
lr 1.000000e-05 reg 5.000000e+03 train accuracy: 0.234980 val accuracy: 0.254000
lr 1.000000e-05 reg 1.000000e+04 train accuracy: 0.209837 val accuracy: 0.200000
lr 1.000000e-05 reg 2.500000e+04 train accuracy: 0.126531 val accuracy: 0.124000
lr 1.000000e-05 reg 5.000000e+04 train accuracy: 0.076571 val accuracy: 0.079000
lr 1.000000e-05 reg 1.000000e+05 train accuracy: 0.132020 val accuracy: 0.141000
best validation accuracy achieved during cross-validation: 0.405000
```

[20]:
```python
# Visualize the cross-validation results
import math
import pdb

# pdb.set_trace()

x_scatter = [math.log10(x[0]) for x in results]
y_scatter = [math.log10(x[1]) for x in results]

# plot training accuracy
marker_size = 100
colors = [results[x][0] for x in results]
plt.subplot(2, 1, 1)
plt.tight_layout(pad=3)
```

```
plt.scatter(x_scatter, y_scatter, marker_size, c=colors, cmap=plt.cm.coolwarm)
plt.colorbar()
plt.xlabel('log learning rate')
plt.ylabel('log regularization strength')
plt.title('CIFAR-10 training accuracy')

# plot validation accuracy
colors = [results[x][1] for x in results] # default size of markers is 20
plt.subplot(2, 1, 2)
plt.scatter(x_scatter, y_scatter, marker_size, c=colors, cmap=plt.cm.coolwarm)
plt.colorbar()
plt.xlabel('log learning rate')
plt.ylabel('log regularization strength')
plt.title('CIFAR-10 validation accuracy')
plt.show()
```



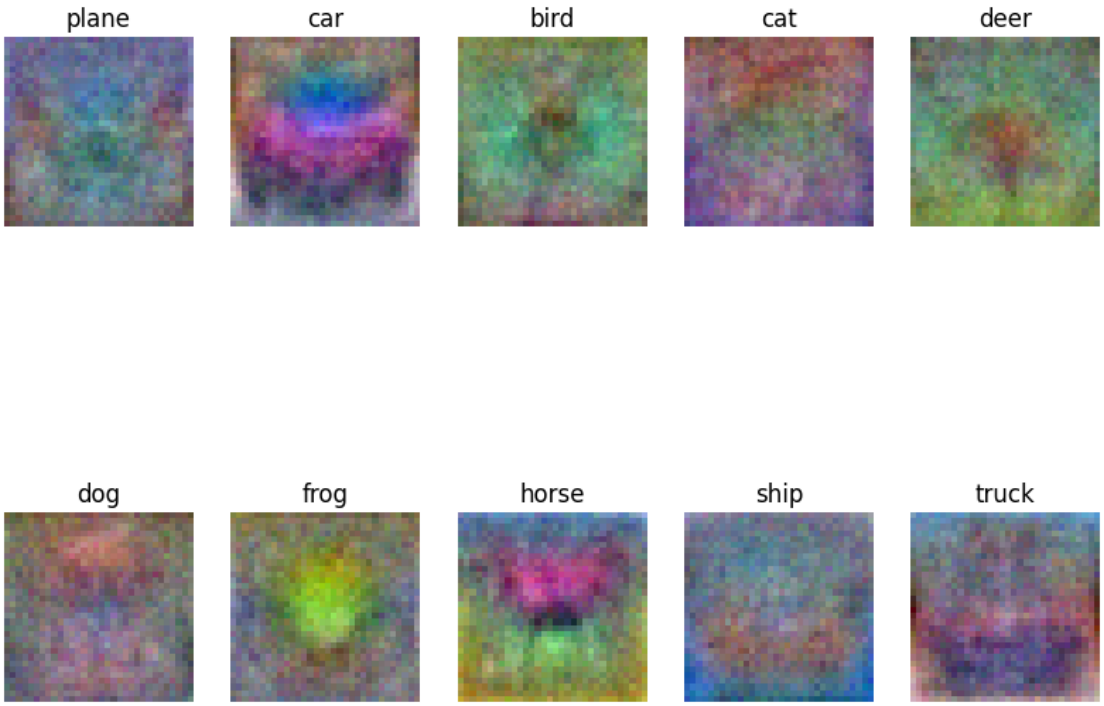
[23]:
```
# Evaluate the best softmax on test set
y_test_pred = best_softmax.predict(X_test)
test_accuracy = np.mean(y_test == y_test_pred)
print('Softmax classifier on raw pixels final test set accuracy: %f' %␣
 ↪test_accuracy)
```

```
Softmax classifier on raw pixels final test set accuracy: 0.381000
```

[24]:
```
# Save best softmax model
best_softmax.save("best_softmax.npy")
```

```
best_softmax.npy saved.
```

[25]:
```
# Visualize the learned weights for each class.
# Depending on your choice of learning rate and regularization strength, these␣
 ↪may
# or may not be nice to look at.
w = best_softmax.W[:-1,:] # strip out the bias
w = w.reshape(32, 32, 3, 10)
w_min, w_max = np.min(w), np.max(w)
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse',␣
 ↪'ship', 'truck']
for i in range(10):
    plt.subplot(2, 5, i + 1)

    # Rescale the weights to be between 0 and 255
    wimg = 255.0 * (w[:, :, :, i].squeeze() - w_min) / (w_max - w_min)
    plt.imshow(wimg.astype('uint8'))
    plt.axis('off')
    plt.title(classes[i])
```



Inline question 3

Describe what your visualized Softmax classifier weights look like, and offer a brief explanation for why they look the way they do.

YourAnswer : 1. **Visualized Weights:** The weights of a Softmax classifier, when reshaped to the original image dimensions, show colorful, blurred, and abstract patterns that resemble the average features of each class.

2. **Explanation:** These patterns emerge because the classifier learns to recognize general features that are indicative of each class, rather than memorizing specific instances. The weights represent the average features learned from the training data, which helps the classifier generalize to new examples.

Inline Question 4 - *True or False*

Suppose the overall training loss is defined as the sum of the per-datapoint loss over all training examples. It is possible to add a new datapoint to a training set that would change the softmax loss, but leave the SVM loss unchanged.

YourAnswer : True

YourExplanation :

1. **Softmax Loss**:
   - The Softmax loss depends on the relative scores of all classes for each data point. Adding a new data point can change the overall distribution of scores, thereby affecting the Softmax loss.

2. **SVM Loss**:
    - The SVM loss is computed based on the margin between the score of the correct class and the scores of the incorrect classes. It only considers the scores that violate the margin condition.

1.2.2 Scenario:

Consider adding a new data point (x_{new}, y_{new}) to the training set:

- **Softmax Loss**:
  - The new data point will affect the overall distribution of scores for all classes. Even if the new data point does not change the relative order of scores significantly, it can still affect the normalization factor in the Softmax loss calculation. Therefore, the Softmax loss will change.
- **SVM Loss**:
  - If the new data point (x_{new}, y_{new}) does not violate the margin condition for any class, it will not contribute to the SVM loss. Specifically, if for all $j \neq y_{\text{new}}$, $s_j - s_{y_{\text{new}}} + \Delta \leq 0$, then the SVM loss for this new data point will be zero and will not affect the overall SVM loss.

1.2.3 Conclusion:

It is possible to add a new data point to a training set that would change the Softmax loss but leave the SVM loss unchanged. This is due to the different nature of how the Softmax and SVM losses are computed.

two_layer_net

April 12, 2025

[2]:
```
# 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/assignment1/'
FOLDERNAME = 'cs231n/assignments/assignment1/'
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
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
/content/drive/My Drive/cs231n/assignments/assignment1/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment1
```

1 Fully-Connected Neural Nets

In this exercise we will implement fully-connected networks using a 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.

[ ]:
```
# As usual, a bit of setup
from __future__ import print_function
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'

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

def rel_error(x, y):
```

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

[ ]:
```
# Load the (preprocessed) CIFAR10 data.

data = get_CIFAR10_data()
for k, 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,))
```

2 Affine layer: forward

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:

[ ]:
```
# 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
```

3 Affine layer: backward

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

[ ]:
```
# 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
```

4 ReLU activation: forward

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

[ ]:
```
# 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
```

5 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:

[ ]:
```
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
```

5.1 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

YourAnswer : Sigmoid and ReLU have the problem of zero (or close to zero) gradient flow during backpropagation.

1. **Sigmoid:** Inputs with large absolute values: When x is large (e.g., x>5 or x<−5), the gradient of the sigmoid function becomes very small, close to zero. This is because the sigmoid function saturates at these values, leading to vanishing gradients.
2. **ReLU:** Negative inputs: When x 0, the gradient of the ReLU function is exactly zero. This means that for any negative input, the gradient flow during backpropagation will be zero, potentially leading to dead neurons that do not update during training.

6 "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:

[ ]:
```
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:  2.299579177309368e-11
dw error:  8.162011105764925e-11
db error:  7.826724021458994e-12
```

7 Loss layers: Softmax

Now implement the loss and gradient for softmax in the `softmax_loss` function in `cs231n/layers.py`. These should be similar to what you implemented in `cs231n/classifiers/softmax.py`. Other loss functions (e.g. `svm_loss`) can also be implemented in a modular way, however, it is not required for this assignment.

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

[ ]:
```
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: 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 softmax_loss:
loss:  2.3025458445007376
dx error:  8.234144091578429e-09
```

8 Two-layer network

Open the file `cs231n/classifiers/fc_net.py` and complete the implementation of the `TwoLayerNet` class. Read through it to make sure you understand the API. You can run the cell below to test your implementation.

[ ]:
```
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)
Running numeric gradient check with reg =  0.0
W1 relative error: 1.83e-08
W2 relative error: 3.20e-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: 9.09e-10
```

9 Solver

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 about `36%` accuracy on the validation set.

[ ]:
```
input_size = 32 * 32 * 3
hidden_size = 50
num_classes = 10
model = TwoLayerNet(input_size, hidden_size, num_classes)
solver = None

################################################################################
# TODO: Use a Solver instance to train a TwoLayerNet that achieves about 36% #
# accuracy on the validation set.                                              #
################################################################################
# Create a Solver instance
solver = Solver(model, data={
    'X_train': data["X_train"],
    'y_train': data["y_train"],
    'X_val': data["X_val"],
    'y_val': data["y_val"]
}, update_rule='sgd', optim_config={'learning_rate': 1e-3}, lr_decay=0.95,␣
 ↪batch_size=200, num_epochs=10, print_every=100)

# Train the model
solver.train()

# Print the final validation accuracy
print('Final validation accuracy: ', solver.val_acc_history[-1])
################################################################################
#                              END OF YOUR CODE                                #
################################################################################
```

```
(Iteration 1 / 2450) loss: 2.304688
(Epoch 0 / 10) train acc: 0.112000; val_acc: 0.149000
(Iteration 101 / 2450) loss: 1.966108
(Iteration 201 / 2450) loss: 1.655363
(Epoch 1 / 10) train acc: 0.416000; val_acc: 0.418000
(Iteration 301 / 2450) loss: 1.508005
(Iteration 401 / 2450) loss: 1.583474
(Epoch 2 / 10) train acc: 0.458000; val_acc: 0.456000
(Iteration 501 / 2450) loss: 1.497746
(Iteration 601 / 2450) loss: 1.385734
(Iteration 701 / 2450) loss: 1.484091
(Epoch 3 / 10) train acc: 0.458000; val_acc: 0.472000
```

```
(Iteration 801 / 2450) loss: 1.541688
(Iteration 901 / 2450) loss: 1.600099
(Epoch 4 / 10) train acc: 0.513000; val_acc: 0.485000
(Iteration 1001 / 2450) loss: 1.557297
(Iteration 1101 / 2450) loss: 1.441770
(Iteration 1201 / 2450) loss: 1.391543
(Epoch 5 / 10) train acc: 0.500000; val_acc: 0.493000
(Iteration 1301 / 2450) loss: 1.438863
(Iteration 1401 / 2450) loss: 1.376444
(Epoch 6 / 10) train acc: 0.511000; val_acc: 0.501000
(Iteration 1501 / 2450) loss: 1.357388
(Iteration 1601 / 2450) loss: 1.444965
(Iteration 1701 / 2450) loss: 1.449189
(Epoch 7 / 10) train acc: 0.546000; val_acc: 0.507000
(Iteration 1801 / 2450) loss: 1.167436
(Iteration 1901 / 2450) loss: 1.285801
(Epoch 8 / 10) train acc: 0.529000; val_acc: 0.495000
(Iteration 2001 / 2450) loss: 1.323224
(Iteration 2101 / 2450) loss: 1.290187
(Iteration 2201 / 2450) loss: 1.474980
(Epoch 9 / 10) train acc: 0.552000; val_acc: 0.510000
(Iteration 2301 / 2450) loss: 1.296648
(Iteration 2401 / 2450) loss: 1.224151
(Epoch 10 / 10) train acc: 0.542000; val_acc: 0.487000
Final validation accuracy:  0.487
```

10 Debug the training

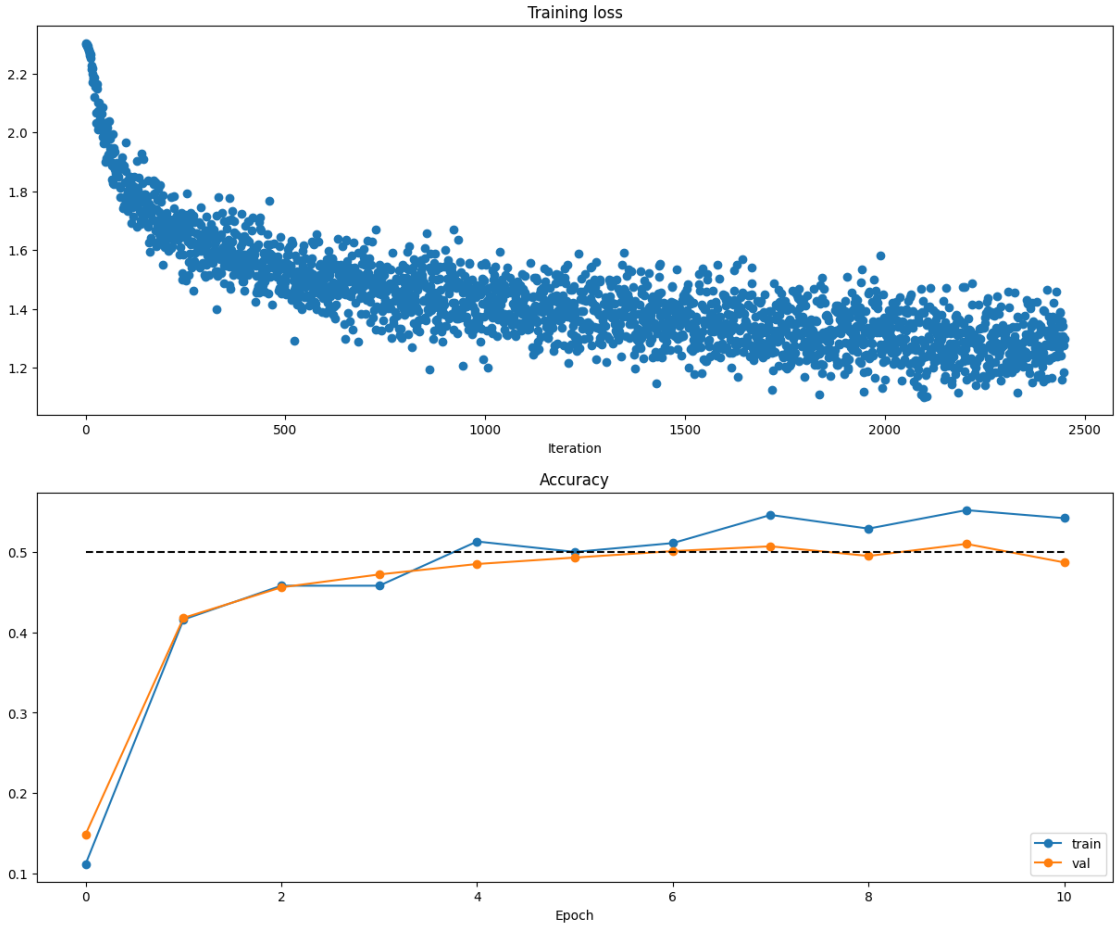
With the default parameters we provided above, you should get a validation accuracy of about 0.36 on the validation set. This isn't very good.

One strategy for getting insight into what's wrong is to plot the loss function and the accuracies on the training and validation sets during optimization.

Another strategy is to visualize the weights that were learned in the first layer of the network. In most neural networks trained on visual data, the first layer weights typically show some visible structure when visualized.

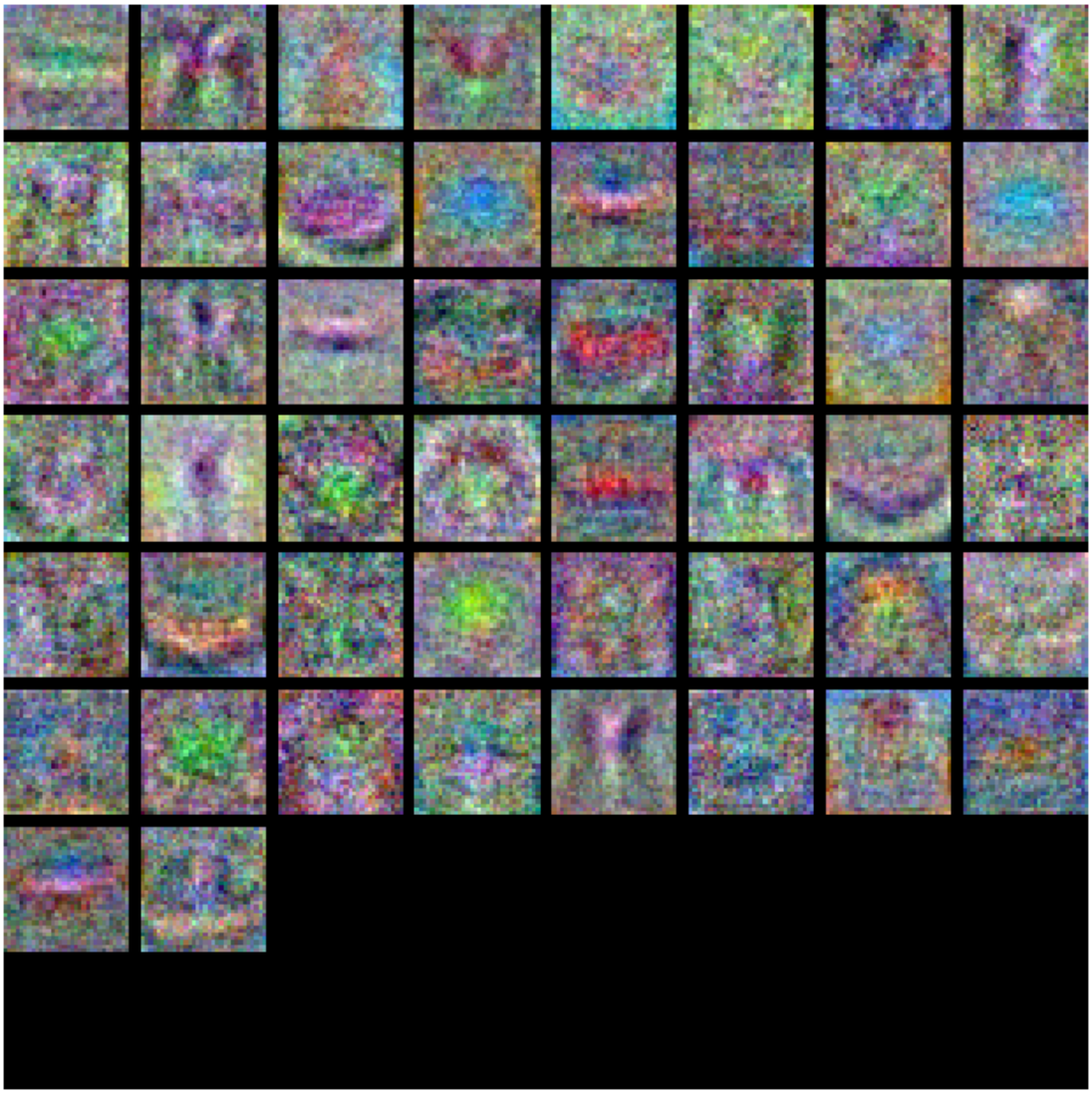
[ ]:
```
# Run this cell to visualize training loss and train / val accuracy

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()
```



[ ]:
```
from cs231n.vis_utils import visualize_grid

# Visualize the weights of the network

def show_net_weights(net):
    W1 = net.params['W1']
    W1 = W1.reshape(3, 32, 32, -1).transpose(3, 1, 2, 0)
    plt.imshow(visualize_grid(W1, padding=3).astype('uint8'))
    plt.gca().axis('off')
    plt.show()
```

```
show_net_weights(model)
```



11 Tune your hyperparameters

What's wrong?. Looking at the visualizations above, we see that the loss is decreasing more or less linearly, which seems to suggest that the learning rate may be too low. Moreover, there is no gap between the training and validation accuracy, suggesting that the model we used has low capacity, and that we should increase its size. On the other hand, with a very large model we would expect to see more overfitting, which would manifest itself as a very large gap between the training and validation accuracy.

Tuning. Tuning the hyperparameters and developing intuition for how they affect the final performance is a large part of using Neural Networks, so we want you to get a lot of practice. Below,

you should experiment with different values of the various hyperparameters, including hidden layer size, learning rate, numer of training epochs, and regularization strength. You might also consider tuning the learning rate decay, but you should be able to get good performance using the default value.

Approximate results. You should be aim to achieve a classification accuracy of greater than 48% on the validation set. Our best network gets over 52% on the validation set.

Experiment: You goal in this exercise is to get as good of a result on CIFAR-10 as you can (52% could serve as a reference), with a fully-connected Neural Network. Feel free implement your own techniques (e.g. PCA to reduce dimensionality, or adding dropout, or adding features to the solver, etc.).

[ ]:
```
best_model = None
best_accuracy = 0.0

from tqdm.auto import tqdm
from sklearn.model_selection import ParameterGrid

################################################################################
# TODO: Tune hyperparameters using the validation set. Store your best trained #
# model in best_model.                                                         #
#                                                                              #
# To help debug your network, it may help to use visualizations similar to the #
# ones we used above; these visualizations will have significant qualitative   #
# differences from the ones we saw above for the poorly tuned network.         #
#                                                                              #
# Tweaking hyperparameters by hand can be fun, but you might find it useful to #
# write code to sweep through possible combinations of hyperparameters         #
# automatically like we did on thexs previous exercises.                       #
################################################################################
param_grid = {
    'learning_rate': [1e-4, 1e-3],
    'batch_size': [100, 200],
    'hidden_size': [30, 50],
    'num_epochs': [10, 20],
    'lr_decay': [0.9, 0.95],
}
```

```
param_combinations = list(ParameterGrid(param_grid))


for params in tqdm(param_combinations, desc="Hyperparameter Tuning",␣
 ↪position=0, leave=True):
    model = TwoLayerNet(input_size, params['hidden_size'], num_classes)

    solver = Solver(model, data={
        'X_train': data["X_train"],
        'y_train': data["y_train"],
        'X_val': data["X_val"],
        'y_val': data["y_val"]
    }, update_rule='sgd', optim_config={'learning_rate':␣
 ↪params['learning_rate']},
        lr_decay=params['lr_decay'], batch_size=params['batch_size'],␣
 ↪num_epochs=params['num_epochs'], print_every=100, verbose=False)

    solver.train()

    final_accuracy = solver.val_acc_history[-1]
    print(f"Training with params: {params}; Final validation accuracy:␣
 ↪{final_accuracy:.4f}")

    if final_accuracy > best_accuracy:
        best_accuracy = final_accuracy
        best_model = model

print(f"Best Validation Accuracy: {best_accuracy:.4f}")
################################################################################
#                               END OF YOUR CODE                               #
################################################################################
```

```
Hyperparameter Tuning:   0%|          | 0/32 [00:00<?, ?it/s]
```

```
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.4470
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.4470
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4510
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4550
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.4960
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.5030
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
```

0.001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4700
Training with params: {'batch_size': 100, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4950
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.4620
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.4580
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4530
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4880
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.5040
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.5170
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.5010
Training with params: {'batch_size': 100, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4990
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.3890
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.4160
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4080
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4540
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.4870
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.4890
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4990
Training with params: {'batch_size': 200, 'hidden_size': 30, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4840
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.3970
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.4340
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4270
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.0001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.4580
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 10}; Final validation accuracy: 0.5190
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.9, 'num_epochs': 20}; Final validation accuracy: 0.5220
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':

```
0.001, 'lr_decay': 0.95, 'num_epochs': 10}; Final validation accuracy: 0.4800
Training with params: {'batch_size': 200, 'hidden_size': 50, 'learning_rate':
0.001, 'lr_decay': 0.95, 'num_epochs': 20}; Final validation accuracy: 0.5040
Best Validation Accuracy: 0.5220
```

12 Test your model!

Run your best model on the validation and test sets. You should achieve above 48% accuracy on the validation set and the test set.

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

```
Validation set accuracy:  0.523
```

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

```
Test set accuracy:  0.537
```

[ ]:
```
# Save best model
best_model.save("best_two_layer_net.npy")
```

```
best_two_layer_net.npy saved.
```

12.1 Inline Question 2:

Now that you have trained a Neural Network classifier, you may find that your testing accuracy is much lower than the training accuracy. In what ways can we decrease this gap? Select all that apply.

1. Train on a larger dataset.
2. Add more hidden units.
3. Increase the regularization strength.
4. None of the above.

YourAnswer : 1, 3

YourExplanation : 1. **Train on a larger dataset.** - Explanation: - A larger dataset can help the model generalize better to unseen data. More data provides more examples for the model to learn from, reducing the likelihood of overfitting to the training set. - Effect: This can improve both training and testing accuracy and reduce the gap between them. 2. **Add more hidden units.** - Explanation: - Adding more hidden units can increase the model's capacity to learn more complex patterns. However, this can also lead to overfitting if the model becomes too complex relative to the amount of training data. - Effect: This might improve training accuracy but could potentially worsen the gap if not managed properly (e.g., with regularization). 3. **Increase the regularization strength.** - Explanation: - Regularization techniques (such as L2 regularization, dropout, or early stopping) help prevent overfitting by adding constraints to the model. Increasing the regularization strength can make the model simpler and more generalizable. - Effect: This can

reduce overfitting and improve testing accuracy, thereby decreasing the gap between training and testing accuracy.

features

April 12, 2025

[ ]:
```
# 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/assignment1/'
FOLDERNAME = 'cs231n/assignments/assignment1/'
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/assignment1/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment1
```

1 Image features exercise

Complete and hand in this completed worksheet (including its outputs and any supporting code outside of the worksheet) with your assignment submission. For more details see the assignments page on the course website.

We have seen that we can achieve reasonable performance on an image classification task by training a linear classifier on the pixels of the input image. In this exercise we will show that we can improve our classification performance by training linear classifiers not on raw pixels but on features that are computed from the raw pixels.

All of your work for this exercise will be done in this notebook.

[18]:
```
import random
import numpy as np
from cs231n.data_utils import load_CIFAR10
import matplotlib.pyplot as plt


%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 extenrnal modules
# see http://stackoverflow.com/questions/1907993/
 ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

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

1.1 Load data

Similar to previous exercises, we will load CIFAR-10 data from disk.

[19]:
```
from cs231n.features import color_histogram_hsv, hog_feature

def get_CIFAR10_data(num_training=49000, num_validation=1000, num_test=1000):
    # Load the raw CIFAR-10 data
    cifar10_dir = 'cs231n/datasets/cifar-10-batches-py'

    # Cleaning up variables to prevent loading data multiple times (which may␣
 ↪cause memory issue)
    try:
       del X_train, y_train
       del X_test, y_test
       print('Clear previously loaded data.')
    except:
       pass

    X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

    # Subsample the data
    mask = list(range(num_training, num_training + num_validation))
    X_val = X_train[mask]
    y_val = y_train[mask]
    mask = list(range(num_training))
    X_train = X_train[mask]
    y_train = y_train[mask]
```

```
    mask = list(range(num_test))
    X_test = X_test[mask]
    y_test = y_test[mask]

    return X_train, y_train, X_val, y_val, X_test, y_test

X_train, y_train, X_val, y_val, X_test, y_test = get_CIFAR10_data()
```

1.2 Extract Features

For each image we will compute a Histogram of Oriented Gradients (HOG) as well as a color histogram using the hue channel in HSV color space. We form our final feature vector for each image by concatenating the HOG and color histogram feature vectors.

Roughly speaking, HOG should capture the texture of the image while ignoring color information, and the color histogram represents the color of the input image while ignoring texture. As a result, we expect that using both together ought to work better than using either alone. Verifying this assumption would be a good thing to try for your own interest.

The `hog_feature` and `color_histogram_hsv` functions both operate on a single image and return a feature vector for that image. The extract_features function takes a set of images and a list of feature functions and evaluates each feature function on each image, storing the results in a matrix where each column is the concatenation of all feature vectors for a single image.

[20]:
```
from cs231n.features import *

# num_color_bins = 10 # Number of bins in the color histogram
num_color_bins = 25 # Number of bins in the color histogram
feature_fns = [hog_feature, lambda img: color_histogram_hsv(img,␣
 ↪nbin=num_color_bins)]
X_train_feats = extract_features(X_train, feature_fns, verbose=True)
X_val_feats = extract_features(X_val, feature_fns)
X_test_feats = extract_features(X_test, feature_fns)

# Preprocessing: Subtract the mean feature
mean_feat = np.mean(X_train_feats, axis=0, keepdims=True)
X_train_feats -= mean_feat
X_val_feats -= mean_feat
X_test_feats -= mean_feat

# Preprocessing: Divide by standard deviation. This ensures that each feature
# has roughly the same scale.
std_feat = np.std(X_train_feats, axis=0, keepdims=True)
X_train_feats /= std_feat
X_val_feats /= std_feat
X_test_feats /= std_feat

# Preprocessing: Add a bias dimension
```

```
X_train_feats = np.hstack([X_train_feats, np.ones((X_train_feats.shape[0], 1))])
X_val_feats = np.hstack([X_val_feats, np.ones((X_val_feats.shape[0], 1))])
X_test_feats = np.hstack([X_test_feats, np.ones((X_test_feats.shape[0], 1))])
```

```
Done extracting features for 1000 / 49000 images
Done extracting features for 2000 / 49000 images
Done extracting features for 3000 / 49000 images
Done extracting features for 4000 / 49000 images
Done extracting features for 5000 / 49000 images
Done extracting features for 6000 / 49000 images
Done extracting features for 7000 / 49000 images
Done extracting features for 8000 / 49000 images
Done extracting features for 9000 / 49000 images
Done extracting features for 10000 / 49000 images
Done extracting features for 11000 / 49000 images
Done extracting features for 12000 / 49000 images
Done extracting features for 13000 / 49000 images
Done extracting features for 14000 / 49000 images
Done extracting features for 15000 / 49000 images
Done extracting features for 16000 / 49000 images
Done extracting features for 17000 / 49000 images
Done extracting features for 18000 / 49000 images
Done extracting features for 19000 / 49000 images
Done extracting features for 20000 / 49000 images
Done extracting features for 21000 / 49000 images
Done extracting features for 22000 / 49000 images
Done extracting features for 23000 / 49000 images
Done extracting features for 24000 / 49000 images
Done extracting features for 25000 / 49000 images
Done extracting features for 26000 / 49000 images
Done extracting features for 27000 / 49000 images
Done extracting features for 28000 / 49000 images
Done extracting features for 29000 / 49000 images
Done extracting features for 30000 / 49000 images
Done extracting features for 31000 / 49000 images
Done extracting features for 32000 / 49000 images
Done extracting features for 33000 / 49000 images
Done extracting features for 34000 / 49000 images
Done extracting features for 35000 / 49000 images
Done extracting features for 36000 / 49000 images
Done extracting features for 37000 / 49000 images
Done extracting features for 38000 / 49000 images
Done extracting features for 39000 / 49000 images
Done extracting features for 40000 / 49000 images
Done extracting features for 41000 / 49000 images
Done extracting features for 42000 / 49000 images
Done extracting features for 43000 / 49000 images
```

```
Done extracting features for 44000 / 49000 images
Done extracting features for 45000 / 49000 images
Done extracting features for 46000 / 49000 images
Done extracting features for 47000 / 49000 images
Done extracting features for 48000 / 49000 images
Done extracting features for 49000 / 49000 images
```

1.3 Train Softmax classifier on features

Using the Softmax code developed earlier in the assignment, train Softmax classifiers on top of the features extracted above; this should achieve better results than training them directly on top of raw pixels.

[31]:
```
# Use the validation set to tune the learning rate and regularization strength

from cs231n.classifiers.linear_classifier import Softmax

learning_rates = [1e-7, 1e-6, 1e-5, 1e-4]
regularization_strengths = [1e4, 1e5, 1e6, 1e7, 5e4, 5e5, 5e6, 5e7]

results = {}
best_val = -1
best_softmax = None

################################################################################
# TODO:                                                                        #
# Use the validation set to set the learning rate and regularization strength. #
# This should be identical to the validation that you did for the Softmax; save#
# the best trained classifer in best_softmax. If you carefully tune the model, #
# you should be able to get accuracy of above 0.42 on the validation set.      #
################################################################################

# Loop over all combinations of learning rates and regularization strengths
for lr in learning_rates:
    for reg in regularization_strengths:
        # Create a new Softmax classifier
        softmax = Softmax()

        # Train the classifier
        loss_hist = softmax.train(X_train_feats, y_train, learning_rate=lr,␣
 ↪reg=reg, num_iters=200, verbose=False)

        # Evaluate the classifier on the training set
        y_train_pred = softmax.predict(X_train_feats)
        train_accuracy = np.mean(y_train == y_train_pred)

        # Evaluate the classifier on the validation set
        y_val_pred = softmax.predict(X_val_feats)
```

```
        val_accuracy = np.mean(y_val == y_val_pred)

        # Save the results
        results[(lr, reg)] = (train_accuracy, val_accuracy)

        # Check if this is the best model so far
        if val_accuracy > best_val:
            best_val = val_accuracy
            best_softmax = softmax

# Print out results.
for lr, reg in sorted(results):
    train_accuracy, val_accuracy = results[(lr, reg)]
    print('lr %e reg %e train accuracy: %f val accuracy: %f' % (
                lr, reg, train_accuracy, val_accuracy))

print('best validation accuracy achieved: %f' % best_val)
```

```
/content/drive/My
Drive/cs231n/assignments/assignment1/cs231n/classifiers/softmax.py:86:
RuntimeWarning: invalid value encountered in subtract
  scores -= np.max(scores, axis=1, keepdims=True)
```

```
lr 1.000000e-07 reg 1.000000e+04 train accuracy: 0.104633 val accuracy: 0.110000
lr 1.000000e-07 reg 5.000000e+04 train accuracy: 0.088857 val accuracy: 0.090000
lr 1.000000e-07 reg 1.000000e+05 train accuracy: 0.094184 val accuracy: 0.103000
lr 1.000000e-07 reg 5.000000e+05 train accuracy: 0.422122 val accuracy: 0.433000
lr 1.000000e-07 reg 1.000000e+06 train accuracy: 0.407020 val accuracy: 0.406000
lr 1.000000e-07 reg 5.000000e+06 train accuracy: 0.323571 val accuracy: 0.321000
lr 1.000000e-07 reg 1.000000e+07 train accuracy: 0.096143 val accuracy: 0.095000
lr 1.000000e-07 reg 5.000000e+07 train accuracy: 0.118796 val accuracy: 0.103000
lr 1.000000e-06 reg 1.000000e+04 train accuracy: 0.179327 val accuracy: 0.162000
lr 1.000000e-06 reg 5.000000e+04 train accuracy: 0.411776 val accuracy: 0.414000
lr 1.000000e-06 reg 1.000000e+05 train accuracy: 0.399061 val accuracy: 0.388000
lr 1.000000e-06 reg 5.000000e+05 train accuracy: 0.307918 val accuracy: 0.291000
lr 1.000000e-06 reg 1.000000e+06 train accuracy: 0.096388 val accuracy: 0.092000
lr 1.000000e-06 reg 5.000000e+06 train accuracy: 0.115878 val accuracy: 0.129000
lr 1.000000e-06 reg 1.000000e+07 train accuracy: 0.100245 val accuracy: 0.118000
lr 1.000000e-06 reg 5.000000e+07 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-05 reg 1.000000e+04 train accuracy: 0.400449 val accuracy: 0.413000
lr 1.000000e-05 reg 5.000000e+04 train accuracy: 0.336837 val accuracy: 0.345000
lr 1.000000e-05 reg 1.000000e+05 train accuracy: 0.110673 val accuracy: 0.113000
lr 1.000000e-05 reg 5.000000e+05 train accuracy: 0.086306 val accuracy: 0.078000
lr 1.000000e-05 reg 1.000000e+06 train accuracy: 0.095388 val accuracy: 0.101000
lr 1.000000e-05 reg 5.000000e+06 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-05 reg 1.000000e+07 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-05 reg 5.000000e+07 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-04 reg 1.000000e+04 train accuracy: 0.088878 val accuracy: 0.078000
```

```
lr 1.000000e-04 reg 5.000000e+04 train accuracy: 0.104163 val accuracy: 0.089000
lr 1.000000e-04 reg 1.000000e+05 train accuracy: 0.098878 val accuracy: 0.095000
lr 1.000000e-04 reg 5.000000e+05 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-04 reg 1.000000e+06 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-04 reg 5.000000e+06 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-04 reg 1.000000e+07 train accuracy: 0.100265 val accuracy: 0.087000
lr 1.000000e-04 reg 5.000000e+07 train accuracy: 0.100265 val accuracy: 0.087000
best validation accuracy achieved: 0.433000
```

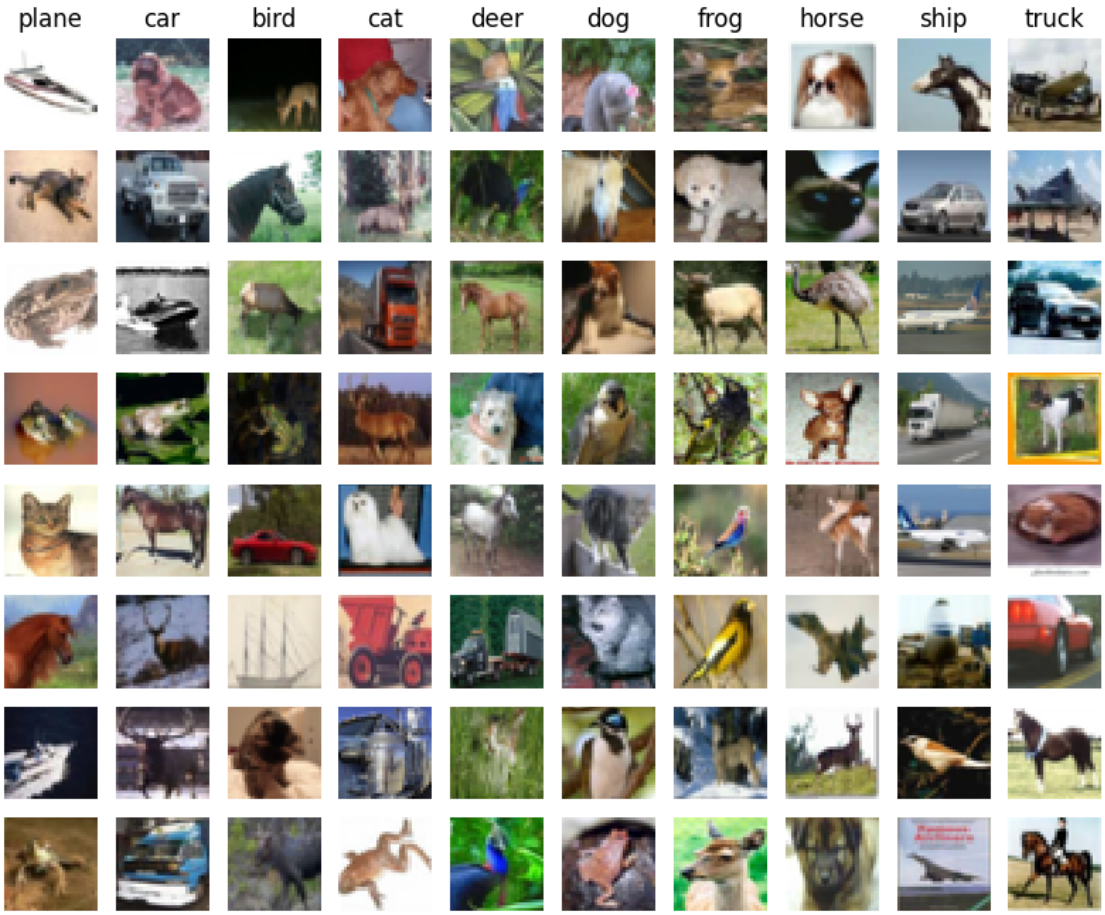
[32]:
```
# Evaluate your trained Softmax on the test set: you should be able to get at␣
 ↪least 0.42
y_test_pred = best_softmax.predict(X_test_feats)
test_accuracy = np.mean(y_test == y_test_pred)
print(test_accuracy)
```

```
0.428
```

[33]:
```
# Save best softmax model
best_softmax.save("best_softmax_features.npy")
```

```
best_softmax_features.npy saved.
```

[34]:
```
# An important way to gain intuition about how an algorithm works is to
# visualize the mistakes that it makes. In this visualization, we show examples
# of images that are misclassified by our current system. The first column
# shows images that our system labeled as "plane" but whose true label is
# something other than "plane".

examples_per_class = 8
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse',␣
 ↪'ship', 'truck']
for cls, cls_name in enumerate(classes):
    idxs = np.where((y_test != cls) & (y_test_pred == cls))[0]
    idxs = np.random.choice(idxs, examples_per_class, replace=False)
    for i, idx in enumerate(idxs):
        plt.subplot(examples_per_class, len(classes), i * len(classes) + cls +␣
 ↪1)
        plt.imshow(X_test[idx].astype('uint8'))
        plt.axis('off')
        if i == 0:
            plt.title(cls_name)
plt.show()
```



1.3.1 Inline question 1:

Describe the misclassification results that you see. Do they make sense?

YourAnswer :

The visualization shows images misclassified as "plane" but actually belonging to other classes. These errors likely make sense due to similarities in shapes or patterns between classes, such as mistaking a "bird" for a "plane" because of their elongated forms.

1.4 Neural Network on image features

Earlier in this assigment we saw that training a two-layer neural network on raw pixels achieved better classification performance than linear classifiers on raw pixels. In this notebook we have seen that linear classifiers on image features outperform linear classifiers on raw pixels.

For completeness, we should also try training a neural network on image features. This approach should outperform all previous approaches: you should easily be able to achieve over 55% classification accuracy on the test set; our best model achieves about 60% classification accuracy.

[35]:
```
# Preprocessing: Remove the bias dimension
# Make sure to run this cell only ONCE
print(X_train_feats.shape)
X_train_feats = X_train_feats[:, :-1]
X_val_feats = X_val_feats[:, :-1]
X_test_feats = X_test_feats[:, :-1]

print(X_train_feats.shape)
```

```
(49000, 170)
(49000, 169)
```

[91]:
```
from cs231n.classifiers.fc_net import TwoLayerNet
from cs231n.solver import Solver

input_dim = X_train_feats.shape[1]
hidden_dim = 500
num_classes = 10

data = {
    'X_train': X_train_feats,
    'y_train': y_train,
    'X_val': X_val_feats,
    'y_val': y_val,
    'X_test': X_test_feats,
    'y_test': y_test,
}

net = TwoLayerNet(input_dim, hidden_dim, num_classes)
best_net = None
best_val_acc = -1
best_params = None

################################################################################
# TODO: Train a two-layer neural network on image features. You may want to    #
# cross-validate various parameters as in previous sections. Store your best   #
# model in the best_net variable.                                              #
################################################################################

from tqdm.auto import tqdm
from sklearn.model_selection import ParameterGrid

# Define the hyperparameter search space
param_grid = {
    'learning_rate': [0.4295, 0.4296, 0.4294],
    'reg': [8.0021e7, 8.0019e7, 8.0020e7],
}
```

```
# Generate all combinations of hyperparameters
param_combinations = list(ParameterGrid(param_grid))

# Hyperparameter search with tqdm progress bar
for params in tqdm(param_combinations, desc="Hyperparameter Tuning",␣
 ↪position=0, leave=True):
    # Create a new TwoLayerNet instance
    net = TwoLayerNet(input_dim, hidden_dim, num_classes)

    # Create a Solver instance to train the network
    solver = Solver(net, data,
                    update_rule='sgd',
                    optim_config={
                        'learning_rate': params['learning_rate'],
                    },
                    lr_decay=0.95,
                    num_epochs=5,
                    batch_size=200,
                    print_every=100,
                    verbose=False)

    # Train the network
    solver.train()

    # Get the validation accuracy
    val_acc = solver.check_accuracy(data['X_val'], data['y_val'])

    # Print the results
    print(f"lr: {params['learning_rate']}, reg: {params['reg']}, val_acc:␣
 ↪{val_acc}")

    # Check if this model is the best so far
    if val_acc > best_val_acc:
        best_val_acc = val_acc
        best_net = net
        best_params = params

# Print the best validation accuracy and corresponding parameters
print(f'Best validation accuracy: {best_val_acc}')
print(f'Best parameters: {best_params}')
```

```
Hyperparameter Tuning:   0%|          | 0/9 [00:00<?, ?it/s]
```

```
lr: 0.4295, reg: 80021000.0, val_acc: 0.585
lr: 0.4295, reg: 80019000.0, val_acc: 0.593
lr: 0.4295, reg: 80020000.0, val_acc: 0.588
lr: 0.4296, reg: 80021000.0, val_acc: 0.584
```

```
lr: 0.4296, reg: 80019000.0, val_acc: 0.579
lr: 0.4296, reg: 80020000.0, val_acc: 0.588
lr: 0.4294, reg: 80021000.0, val_acc: 0.598
lr: 0.4294, reg: 80019000.0, val_acc: 0.595
lr: 0.4294, reg: 80020000.0, val_acc: 0.577
Best validation accuracy: 0.598
Best parameters: {'learning_rate': 0.4294, 'reg': 80021000.0}
```

[92]:
```python
# Run your best neural net classifier on the test set. You should be able
# to get more than 58% accuracy. It is also possible to get >60% accuracy
# with careful tuning.

y_test_pred = np.argmax(best_net.loss(data['X_test']), axis=1)
test_acc = (y_test_pred == data['y_test']).mean()
print(test_acc)
```

```
0.581
```

[93]:
```python
# Save best model
best_net.save("best_two_layer_net_features.npy")
```

```
best_two_layer_net_features.npy saved.
```

FullyConnectedNets

April 12, 2025

[2]:
```
# 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/assignment1/'
FOLDERNAME = 'cs231n/assignments/assignment1/'
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/$FOLDERNAME/cs231n/datasets/
# %cd /content/drive/My\ Drive/$FOLDERNAME
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
Mounted at /content/drive
/content/drive/My Drive/cs231n/assignments/assignment1/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment1
```

1 Multi-Layer Fully Connected Network

In this exercise, you will implement a fully connected network with an arbitrary number of hidden layers.

[3]:
```
# from google.colab import drive
# drive.mount('/content/drive')
```

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 before. For right now, don't worry about implementing dropout or batch/layer normalization yet, as you will add those features later.

[4]:
```
# 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))))
```

[5]:
```
# 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,)
```

1.1 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.

[6]:
```
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])}")
```

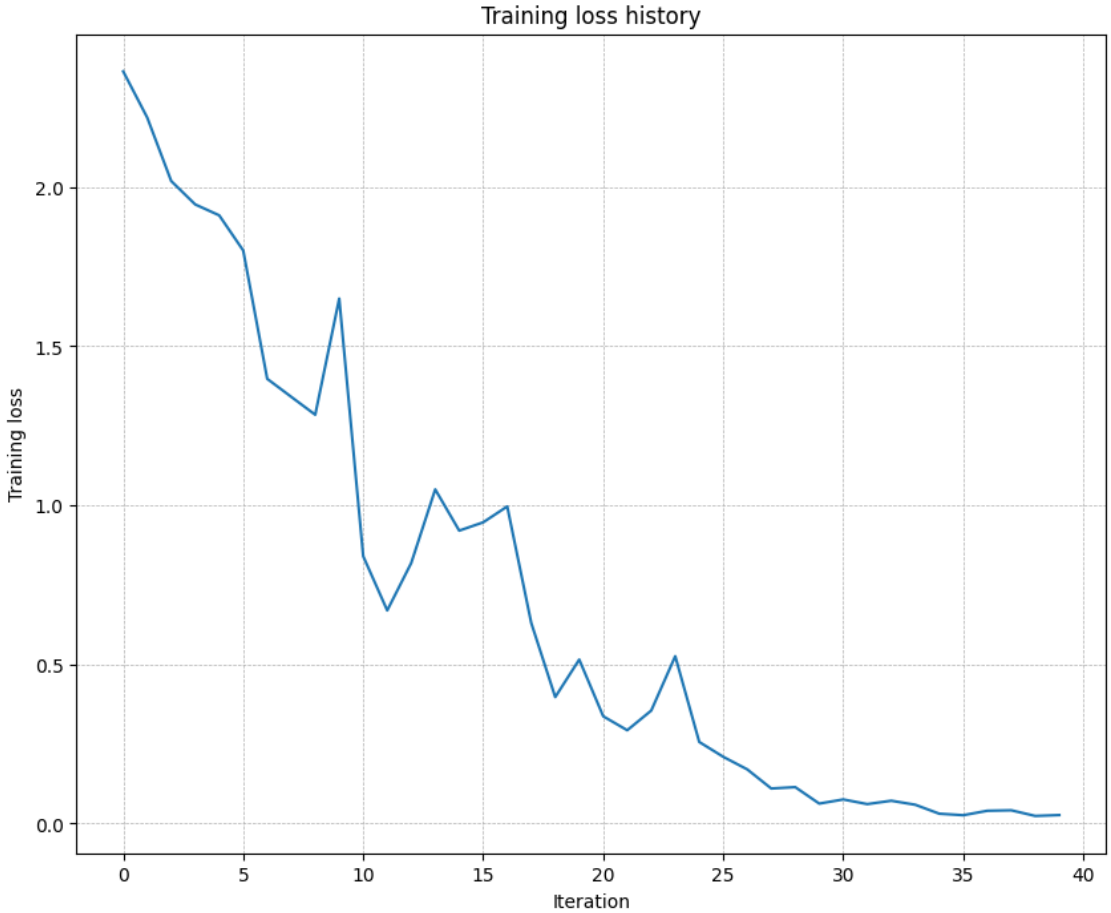
```
Running check with reg =  0
Initial loss:  2.300479089768492
W1 relative error: 1.0252674471656573e-07
W2 relative error: 2.2120479295080622e-05
W3 relative error: 4.5623278736665505e-07
b1 relative error: 4.6600944653202505e-09
b2 relative error: 2.085654276112763e-09
b3 relative error: 1.689724888469736e-10
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: 2.6171457283983532e-08
b1 relative error: 1.4752427965311745e-08
b2 relative error: 1.7223751746766738e-09
b3 relative error: 2.378772438198909e-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.

[7]:
```
# 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)
plt.title("Training loss history")
plt.xlabel("Iteration")
plt.ylabel("Training loss")
plt.grid(linestyle='--', linewidth=0.5)
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
```



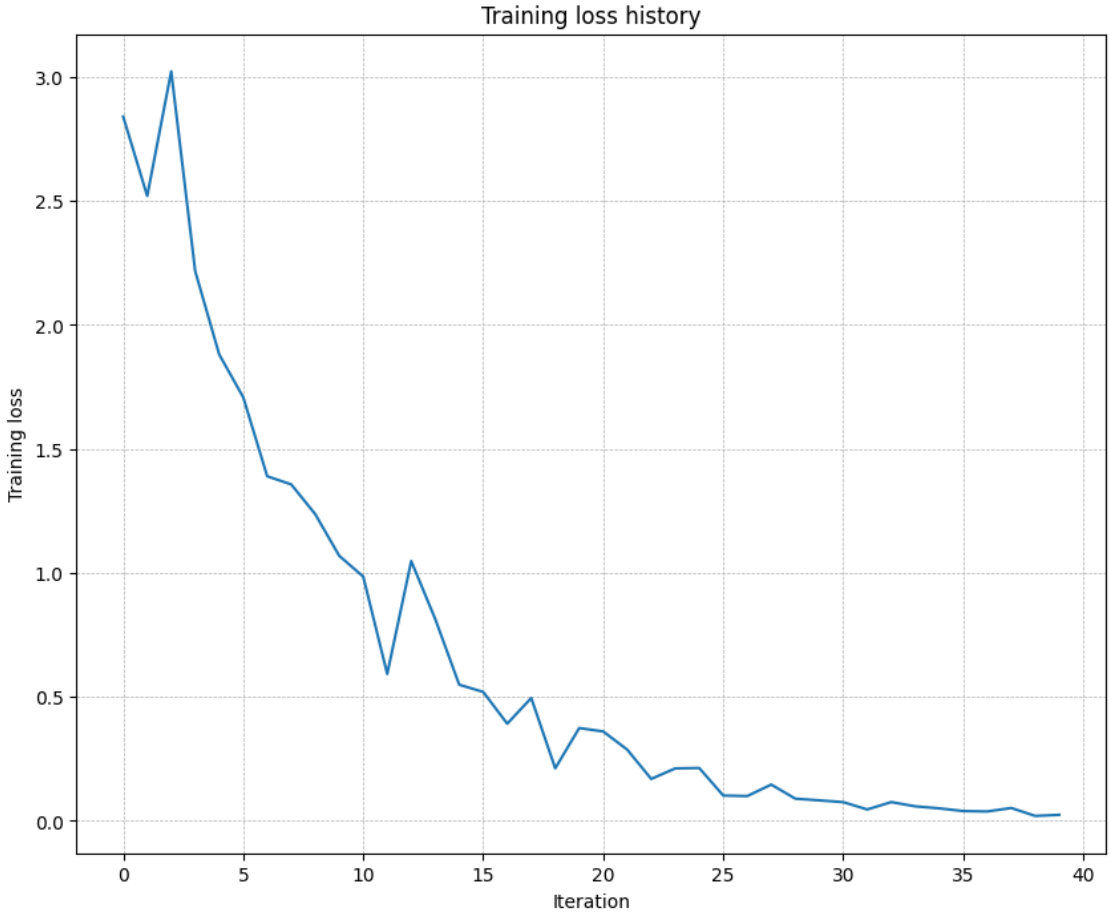
Now, try to use a five-layer network with 100 units on each layer to overfit on 50 training examples. Again, you will have to adjust the learning rate and weight initialization scale, but you should be

able to achieve 100% training accuracy within 20 epochs.

[8]:
```
# 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-2  # Experiment with this!
weight_scale = 4e-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: 2.839886
(Epoch 0 / 20) train acc: 0.220000; val_acc: 0.111000
(Epoch 1 / 20) train acc: 0.240000; val_acc: 0.084000
(Epoch 2 / 20) train acc: 0.420000; val_acc: 0.137000
(Epoch 3 / 20) train acc: 0.440000; val_acc: 0.128000
(Epoch 4 / 20) train acc: 0.700000; val_acc: 0.148000
(Epoch 5 / 20) train acc: 0.740000; val_acc: 0.150000
(Iteration 11 / 40) loss: 0.985188
```

```
(Epoch 6 / 20) train acc: 0.780000; val_acc: 0.166000
(Epoch 7 / 20) train acc: 0.900000; val_acc: 0.178000
(Epoch 8 / 20) train acc: 0.940000; val_acc: 0.166000
(Epoch 9 / 20) train acc: 0.940000; val_acc: 0.169000
(Epoch 10 / 20) train acc: 0.900000; val_acc: 0.185000
(Iteration 21 / 40) loss: 0.360608
(Epoch 11 / 20) train acc: 1.000000; val_acc: 0.182000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.188000
(Epoch 13 / 20) train acc: 0.980000; val_acc: 0.200000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.188000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.179000
(Iteration 31 / 40) loss: 0.075659
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.185000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.193000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.189000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.186000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.180000
```



1.2 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?

1.3 Answer:

1. **Three-Layer Network:** Less sensitive to initialization, easier to train.
2. **Five-Layer Network:** More sensitive to initialization, harder to train, needs careful setup.

Why Deeper Networks Are More Sensitive

- Vanishing/Exploding Gradients:
  - More layers mean gradients can get very small (vanish) or very large (explode) more easily.
- Complexity:
  - More parameters and a more complex loss landscape make it harder to find good solutions.

2 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.

2.1 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.

[9]:
```
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
```

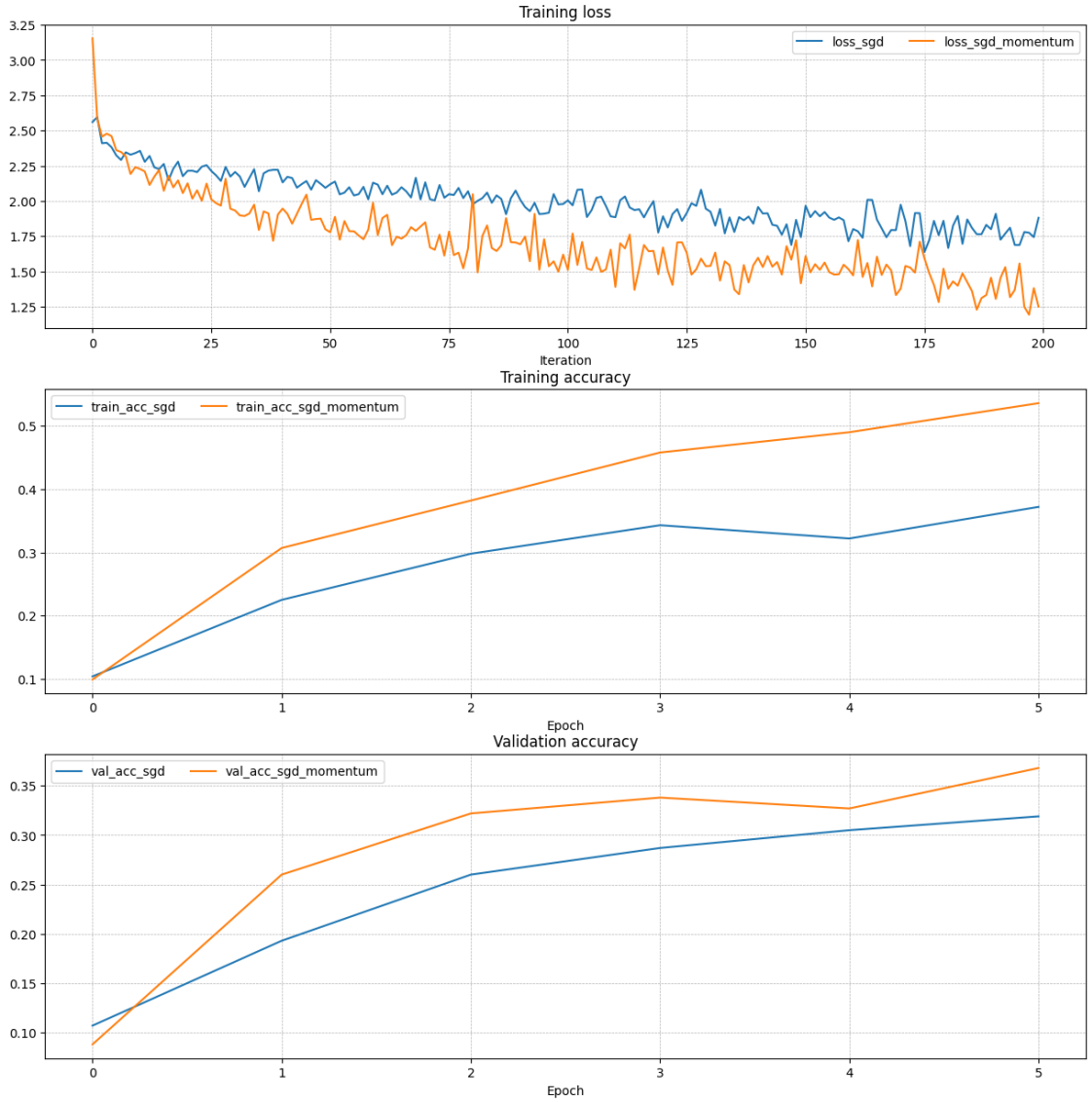
Once you have done so, run the following to train a six-layer network with both SGD and SGD+momentum. You should see the SGD+momentum update rule converge faster.

[10]:
```
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()

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: 1.891516
(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: 1.666573
(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: 1.778583
(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: 1.305934
(Epoch 5 / 5) train acc: 0.536000; val_acc: 0.368000
```



2.2 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.

[11]:
```
# 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
```

[12]:
```
# 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
```

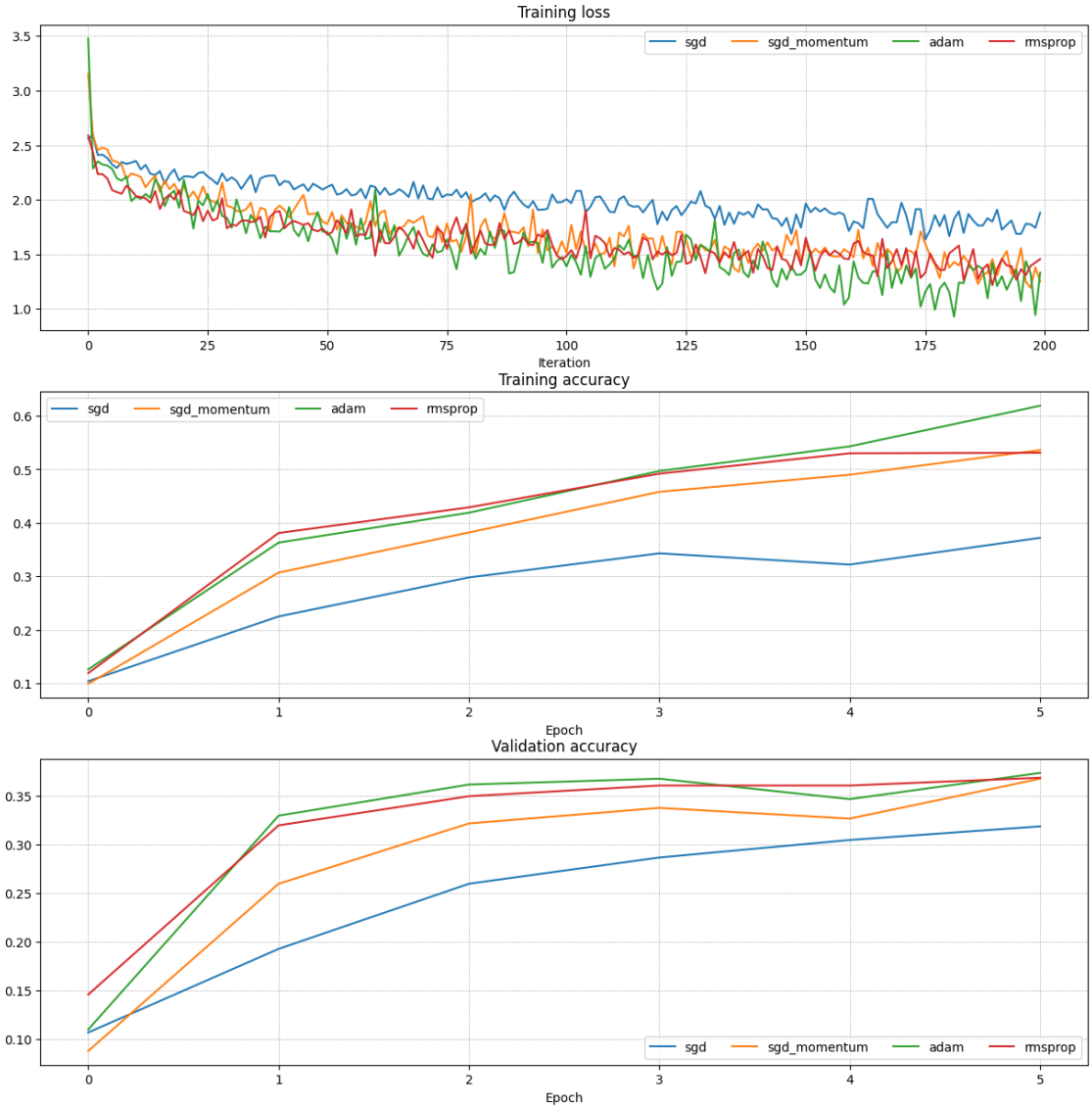
Once you have debugged your RMSProp and Adam implementations, run the following to train a pair of deep networks using these new update rules:

[13]:
```
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()

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: 1.505558
(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: 1.520198
(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: 1.436402
(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: 1.365772
(Epoch 5 / 5) train acc: 0.531000; val_acc: 0.369000
```



2.3 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?

2.4 Answer:

- **AdaGrad:** Updates become small because it keeps adding squared gradients to a cache, making the denominator in the update rule grow larger over time. This reduces the effective learning rate.
- **Adam:** Doesn't have this issue because it uses a moving average of squared gradients (with decay) instead of accumulating them indefinitely. This keeps the updates more stable.

3 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 next 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: In the next assignment, you will learn techniques like BatchNormalization and Dropout which can help you train powerful models.

[84]:
```
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)*****

num_train = 40000
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'],
}

# Define the hyperparameters
hidden_dims = [200, 200, 200, 200, 200]  # 5 hidden layers with 100 units each
weight_scale = 2.9e-2
learning_rate = 8.25e-4
reg = 1.85e-2
num_epochs = 20
batch_size = 2000

# Initialize the model
model = FullyConnectedNet(hidden_dims, input_dim=32*32*3, num_classes=10,
```

```
                          weight_scale=weight_scale, reg=reg, dtype=np.float64)

# Train the model using Adam optimizer
solver = Solver(model,
                small_data,
                update_rule='adam',
                optim_config={'learning_rate': learning_rate},
                lr_decay=0.95,
                num_epochs=num_epochs,
                print_every=100,
                batch_size=batch_size,
                verbose=True)

solver.train()

# Store the best model
best_model = model

# Print the best validation accuracy
print('Best validation accuracy: ', max(solver.val_acc_history))

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

```
(Iteration 1 / 400) loss: 8.375623
(Epoch 0 / 20) train acc: 0.180000; val_acc: 0.183000
(Epoch 1 / 20) train acc: 0.373000; val_acc: 0.374000
(Epoch 2 / 20) train acc: 0.440000; val_acc: 0.433000
(Epoch 3 / 20) train acc: 0.472000; val_acc: 0.453000
(Epoch 4 / 20) train acc: 0.518000; val_acc: 0.487000
(Epoch 5 / 20) train acc: 0.519000; val_acc: 0.472000
(Iteration 101 / 400) loss: 2.153963
(Epoch 6 / 20) train acc: 0.521000; val_acc: 0.479000
(Epoch 7 / 20) train acc: 0.530000; val_acc: 0.481000
(Epoch 8 / 20) train acc: 0.564000; val_acc: 0.487000
(Epoch 9 / 20) train acc: 0.542000; val_acc: 0.513000
(Epoch 10 / 20) train acc: 0.589000; val_acc: 0.508000
(Iteration 201 / 400) loss: 1.640151
(Epoch 11 / 20) train acc: 0.640000; val_acc: 0.527000
(Epoch 12 / 20) train acc: 0.632000; val_acc: 0.517000
(Epoch 13 / 20) train acc: 0.615000; val_acc: 0.500000
(Epoch 14 / 20) train acc: 0.627000; val_acc: 0.511000
(Epoch 15 / 20) train acc: 0.657000; val_acc: 0.526000
(Iteration 301 / 400) loss: 1.500681
(Epoch 16 / 20) train acc: 0.642000; val_acc: 0.521000
```

```
(Epoch 17 / 20) train acc: 0.655000; val_acc: 0.530000
(Epoch 18 / 20) train acc: 0.649000; val_acc: 0.541000
(Epoch 19 / 20) train acc: 0.661000; val_acc: 0.552000
(Epoch 20 / 20) train acc: 0.695000; val_acc: 0.539000
Best validation accuracy:  0.552
```

4 Test Your Model!

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

[85]:
```
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:  0.552
Test set accuracy:  0.528
```