

# Homework 4

## Neural Networks Instructions

Pupipat Singkhorn

March 12, 2024

### Instructions

Answer the questions and upload your answers to courseville. Answers can be in Thai or English. Answers can be either typed or handwritten and scanned. the assignment is divided into several small tasks. Each task is weighted equally (marked with T). For this assignment, each task is awarded equally. There are also optional tasks (marked with OT) counts for half of the required task.

### The Basics

In this section, we will review some of the basic materials taught in class. These are simple tasks and integral to the understanding of deep neural networks, but many students seem to misunderstand.

#### T1.

Compute the forward and backward pass of the following computation. Note that this is a simplified residual connection.

$$\begin{aligned}x_1 &= ReLU(x_0 * w_0 + b_0) \\y_1 &= x_1 * w_1 + b_1 \\z &= ReLU(y_1 + x_0)\end{aligned}$$

Let  $x_0 = 1.0$ ,  $w_0 = 0.3$ ,  $w_1 = -0.2$ ,  $b_0 = 0.1$ ,  $b_1 = -0.3$ . Find the gradient of  $z$  with respect to  $w_0$ ,  $w_1$ ,  $b_0$ , and  $b_1$ .

**Answer:**

Forward Pass

$$\begin{aligned}x_1 &= ReLU(x_0 * w_0 + b_0) \\x_1(x_0 = 1.0, w_0 = 0.3, b_0 = 0.1) &= ReLU(1.0 * 0.3 + 0.1) \\&= ReLU(0.4) \\&= max(0, 0.4) \\\therefore x_1(x_0 = 1.0, w_0 = 0.3, b_0 = 0.1) &= 0.4\end{aligned}$$

$$\begin{aligned}y_1 &= x_1 * w_1 + b_1 \\y_1(x_1 = 0.4, w_1 = -0.2, b_1 = -0.3) &= 0.4 * -0.2 + -0.3 \\\therefore y_1(x_1 = 0.4, w_1 = -0.2, b_1 = -0.3) &= -0.38\end{aligned}$$

$$\begin{aligned}z &= ReLU(y_1 + x_0) \\z(y_1 = -0.38, x_0 = 1.0) &= ReLU(-0.38 + 1.0) \\&= ReLU(0.62) \\&= max(0, 0.62) \\\therefore z(y_1 = -0.38, x_0 = 1.0) &= 0.62\end{aligned}$$

Backward Pass

*function* :  $x_1(x_0, w_0, b_0), y_1(x_1, w_1, b_1), z(y_1, x_0)$

$$\begin{aligned}
 \frac{\partial z}{\partial w_0} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial x_1} * \frac{\partial x_1}{\partial w_0} \\
 &= \frac{\partial}{\partial y_1} ReLU(y_1 + x_0) \\
 &\quad * \frac{\partial}{\partial x_1} (x_1 * w_1 + b_1) \\
 &\quad * \frac{\partial}{\partial w_0} ReLU(x_0 * w_0 + b_0) \\
 \frac{\partial z}{\partial w_0} \Big|_{x_0=1.0, w_1=-0.2, b_0=0.1, b_1=-0.3} &= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0) \Big|_{x_0=1.0, y_1=-0.38} * w_1 * x_0
 \end{aligned}$$

$$\bullet \frac{\partial}{\partial x} ReLU(x) = \begin{cases} 1 & ; x > 0 \\ 0 & ; x \leq 0 \end{cases}$$

$$\begin{aligned}
 \frac{\partial z}{\partial w_0} \Big|_{x_0=1.0, w_1=-0.2, b_0=0.1, b_1=-0.3} &= 1 * -0.2 * 1 \\
 &= -0.2
 \end{aligned}$$

$\therefore$  Gradient of  $z$  with respect to  $w_0 = -0.2$

$$\begin{aligned}
 \frac{\partial z}{\partial w_1} \Big|_{x_0=1.0, w_0=0.3, b_0=0.1, b_1=-0.3} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial w_1} \\
 &= \frac{\partial}{\partial (y_1 + x_0)} ReLU(y_1 + x_0) \Big|_{x_0=1.0, y_1=-0.38} * x_1 \\
 &= 1 * 0.4 \\
 &= 0.4
 \end{aligned}$$

$\therefore$  Gradient of  $z$  with respect to  $w_1 = 0.4$

$$\begin{aligned}
\left. \frac{\partial z}{\partial b_0} \right|_{x_0=1.0, w_0=0.3, w_1=-0.2, b_1=-0.3} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial x_1} * \frac{\partial x_1}{\partial b_0} \\
&= \frac{\partial}{\partial(y_1 + x_0)} ReLU(y_1 + x_0) \Big|_{x_0, y_1} \\
&\quad * w_1 \\
&\quad * \frac{\partial}{\partial(x_0 * w_0 + b_0)} ReLU(x_0 * w_0 + b_0) \Big|_{x_0, w_0, b_0} \\
&= 1 * -0.2 * 1 \\
&= -0.2
\end{aligned}$$

$\therefore$  Gradient of  $z$  with respect to  $b_0 = -0.2$

$$\begin{aligned}
\left. \frac{\partial z}{\partial b_1} \right|_{x_0=1.0, w_0=0.3, w_1=-0.2, b_0=0.1} &= \frac{\partial z}{\partial y_1} * \frac{\partial y_1}{\partial b_1} \\
&= \frac{\partial}{\partial(y_1 + x_0)} ReLU(y_1 + x_0) \Big|_{x_0, w_0, w_1, b_0} * 1 \\
&= 1 * 1 \\
&= 1
\end{aligned}$$

$\therefore$  Gradient of  $z$  with respect to  $b_1 = 1$

## T2.

Given the following network architecture specifications, determine the size of the output A, B, and C.

∴ Size of the output A =  $32 * 1024 = 32,768$

∴ Size of the output B =  $32 * 512 = 16,384$

∴ Size of the output C =  $32 * 1 = 32$

## T3.

What is the total number of learnable parameters in this network? (Don't forget the bias term)

$$\begin{aligned} Total &= First\ layer + Second\ layer + Third\ layer \\ &= (30 + 1) * 1024 + (1024 + 1) * 512 + (512 + 1) * 1 \\ &= 557,057 \end{aligned}$$

∴ Total learnable parameters = 557,057

## Deep Learning from (almost) scratch

In this section we will code simple a neural network model from scratch (numpy). However, before we go into coding let's start with some loose ends, namely the gradient of the softmax layer.

Recall in class we define the softmax layer as:

$$P(y = j) = \frac{\exp(h_j)}{\sum_k \exp(h_k)}$$

where  $h_j$  is the output of the previous layer for class index  $j$ . The cross entropy loss is defined as:

$$L = - \sum_j y_j \log P(y = j)$$

where  $y_j$  is 1 if  $y$  is class  $j$ , and 0 otherwise.

Prove that the derivative of the loss with respect to  $h_i$  is  $P(y = i) - y_i$ . In other words, find  $\frac{\partial L}{\partial h_i}$  for  $i \in \{0, \dots, N - 1\}$  where  $N$  is the number of classes.

Hint: First find  $\frac{\partial P(y=j)}{\partial h_i}$  for the case where  $j = i$ , and the case where  $j \neq i$ . Then, use the results with the chain rule to find the derivative of the loss. To find the derivative

of the loss with respect to  $h_i$ , we need to compute  $\frac{\partial P(y=j)}{\partial h_i}$  for the case where  $j = i$  and the case where  $j \neq i$ , and then use the chain rule to find  $\frac{\partial L}{\partial h_i}$ .

**Answer:**

Compute  $\frac{\partial P(y=j)}{\partial h_i}$  for  $j = i$ :

$$\begin{aligned}
 \frac{\partial P(y=i)}{\partial h_i} &= \frac{\partial}{\partial h_i} \left( \frac{\exp(h_i)}{\sum_k \exp(h_k)} \right) \\
 &= \frac{\exp(h_i) \sum_k \exp(h_k) - \exp(h_i) \exp(h_i)}{(\sum_k \exp(h_k))^2} \\
 &= \frac{\exp(h_i)}{\sum_k \exp(h_k)} - \frac{(\exp(h_i))^2}{(\sum_k \exp(h_k))^2} \\
 &= P(y=i) - (P(y=i))^2 \\
 \therefore \frac{\partial P(y=i)}{\partial h_i} &= P(y=i) \cdot (1 - P(y=i))
 \end{aligned}$$

Compute  $\frac{\partial P(y=j)}{\partial h_i}$  for  $j \neq i$ :

$$\begin{aligned}
 \frac{\partial P(y=j)}{\partial h_i} &= \frac{\partial}{\partial h_i} \left( \frac{\exp(h_j)}{\sum_k \exp(h_k)} \right) \\
 &= -\frac{\exp(h_j) \exp(h_i)}{(\sum_k \exp(h_k))^2} \\
 \therefore \frac{\partial P(y=j)}{\partial h_i} &= -P(y=j) \cdot P(y=i)
 \end{aligned}$$

Now, we can compute  $\frac{\partial L}{\partial h_i}$

$$\begin{aligned}
\frac{\partial L}{\partial h_i} &= \frac{\partial}{\partial h_i} \left( - \sum_j y_j \log P(y = j) \right) \\
&= \sum_j - \frac{\partial}{\partial h_i} y_j \log P(y = j) \\
&= - \frac{\partial}{\partial h_i} y_i \log P(y = i) - \sum_{j \neq i} \frac{\partial}{\partial h_i} y_j \log P(y = j) \\
&= - y_i \frac{\partial}{\partial h_i} \log P(y = i) - \log P(y = i) \frac{\partial}{\partial h_i} y_i \\
&\quad - \sum_{j \neq i} \left( y_j \frac{\partial}{\partial h_i} \log P(y = j) + \log P(y = j) \frac{\partial}{\partial h_i} y_j \right) \\
&= - y_i \frac{\partial}{\partial h_i} \log P(y = i) - \sum_{j \neq i} y_j \frac{\partial}{\partial h_i} \log P(y = j) \\
&= - \frac{y_i}{P(y = i)} \frac{\partial}{\partial h_i} P(y = i) - \sum_{j \neq i} \frac{y_j}{P(y = j)} \frac{\partial}{\partial h_i} P(y = j) \\
&= - \frac{y_i}{P(y = i)} P(y = i) (1 - P(y = i)) - \sum_{j \neq i} \frac{y_j}{P(y = j)} \cdot (-P(y = j) P(y = i)) \\
&= -y_i + y_i P(y = i) + \sum_{j \neq i} y_j P(y = i) \\
&= -y_i + y_i P(y = i) + P(y = i) \sum_{j \neq i} y_j \\
&= -y_i + y_i P(y = i) + P(y = i) \left( \sum_j y_j - y_i \right) \\
&= -y_i + y_i P(y = i) + P(y = i) (1 - y_i) \\
&= -y_i + y_i P(y = i) + P(y = i) - y_i P(y = i) \\
\therefore \frac{\partial L}{\partial h_i} &= P(y = i) - y_i
\end{aligned}$$

Therefore, the derivative of the loss with respect to  $h_i$  is  $P(y = i) - y_i$ .

# HW4\_Simple\_Neural\_Network\_Lab

March 13, 2024

## 1 Two-Layer Neural Networks

In this part of the homework, we will work on building a simple Neural Network to classify digits using MNIST dataset. There are many powerful neural network frameworks nowadays that are relatively straightforward to use. However, we believe that in order to understand the theory behind neural networks, and to have the right intuitions in order to build, use, and debug more complex architectures, one must first start from the basics.

Your task is to complete the missing codes needed for training and testing of a simple fully-connected neural network.

The missing parts will be marked with `TODO#N`, where `N` represents the task number.

Many parts in this homework are modified from Stanford's cs231n assignments.

```
[ ]: import numpy as np
import matplotlib.pyplot as plt

from __future__ import print_function

%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'

from cattern.neural_net import TwoLayerNet

%load_ext autoreload
%autoreload 2

[ ]: # Just a function that verifies if your answers are correct

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))))
```

We will use the class `TwoLayerNet` in the file `cattern/neural_net.py` to represent instances of our network. The network parameters (weights and biases) are stored in the instance variable `self.params` where the keys are parameter names and values are numpy arrays. Below, we initialize some toy data and a toy model that will guide you with your implementation.



```
[ ]: # Create a small net and some toy data to check your implementations.
# Note that we set the random seed for repeatable experiments.

input_size = 4
hidden_size = 10
num_classes = 3
num_inputs = 5

def init_toy_model():
    np.random.seed(0)
    return TwoLayerNet(input_size, hidden_size, num_classes, std=1e-1)

def init_toy_data():
    np.random.seed(1)
    X = 10 * np.random.randn(num_inputs, input_size)
    y = np.array([0, 1, 2, 2, 1])
    return X, y

net = init_toy_model()
X, y = init_toy_data()
```

## 2 Forward pass: compute scores

Open the file `cattern/neural_net.py` and look at the method `TwoLayerNet.loss`. This function takes the data and weights and computes the class scores, the loss, and the gradients on the parameters.

Complete `TODO#1` in `TwoLayerNet.loss`, by implementing the first part of the forward pass which uses the weights and biases to compute the scores for all inputs. The scores refer to the output of the network just before the softmax layer.

```
[ ]: scores = net.loss(X)
print('Your scores:')
print(scores)
print()
print('correct scores:')
correct_scores = np.asarray([
    [-0.81233741, -1.27654624, -0.70335995],
    [-0.17129677, -1.18803311, -0.47310444],
    [-0.51590475, -1.01354314, -0.8504215 ],
    [-0.15419291, -0.48629638, -0.52901952],
    [-0.00618733, -0.12435261, -0.15226949]])
print(correct_scores)
print()

# The difference should be very small. We get < 1e-7
print('Difference between your scores and correct scores:')
```

```
print(np.sum(np.abs(scores - correct_scores)))
```

Your scores:

```
[[-0.81233741 -1.27654624 -0.70335995]
 [-0.17129677 -1.18803311 -0.47310444]
 [-0.51590475 -1.01354314 -0.8504215 ]
 [-0.15419291 -0.48629638 -0.52901952]
 [-0.00618733 -0.12435261 -0.15226949]]
```

correct scores:

```
[[-0.81233741 -1.27654624 -0.70335995]
 [-0.17129677 -1.18803311 -0.47310444]
 [-0.51590475 -1.01354314 -0.8504215 ]
 [-0.15419291 -0.48629638 -0.52901952]
 [-0.00618733 -0.12435261 -0.15226949]]
```

Difference between your scores and correct scores:  
3.6802720926321086e-08

### 3 Forward pass: compute loss

In the same function, complete TODO#2 by implementing the second part that computes the data and regularization loss.

```
[ ]: loss, _ = net.loss(X, y, reg=0.1)
      correct_loss = 1.30378789133

      # should be very small, we get < 1e-12
      print('Difference between your loss and correct loss:')
      print(np.sum(np.abs(loss - correct_loss)))
      print(loss, correct_loss)
```

Difference between your loss and correct loss:  
1.7985612998927536e-13  
1.3037878913298202 1.30378789133

### 4 Backward pass

Implement the rest of the function by completing TODO#3. This will compute the gradient of the loss with respect to the variables W1, b1, W2, and b2. Now that you (hopefully!) have a correctly implemented forward pass, you can debug your backward pass using a numeric gradient check:

```
[ ]: from cattern.gradient_check import eval_numerical_gradient

      # Use numeric gradient checking to check your implementation of the backward
      #    pass.
      # If your implementation is correct, the difference between the numeric and
```

```

# analytic gradients should be less than 1e-8 for each of W1, W2, b1, and b2.

loss, grads = net.loss(X, y, reg=0.05)

# these should all be less than 1e-8 or so
for param_name in grads:
    f = lambda W: net.loss(X, y, reg=0.05)[0]
    param_grad_num = eval_numerical_gradient(f, net.params[param_name],
    verbose=False)
    print('%s max relative error: %e' % (param_name, rel_error(param_grad_num,
    grads[param_name])))

```

```

W2 max relative error: 3.440708e-09
b2 max relative error: 3.865080e-11
W1 max relative error: 4.090897e-09
b1 max relative error: 2.738421e-09

```

## 5 Train the network

To train the network we will use stochastic gradient descent (SGD). Look at the function `TwoLayerNet.train` and fill in the missing sections to implement the training procedure (TODO#4-6). You will also have to implement `TwoLayerNet.predict` (TODO#7), as the training process periodically performs prediction to keep track of accuracy over time while the network trains.

Once you have implemented the method, run the code below to train a two-layer network on toy data. You should achieve a training loss less than 0.05.

```

[ ]: net = init_toy_model()
stats = net.train(X, y, X, y,
                  learning_rate=1e-1, reg=5e-6,
                  num_iters=100, verbose=False)

print('Final training loss: ', stats['loss_history'][-1])

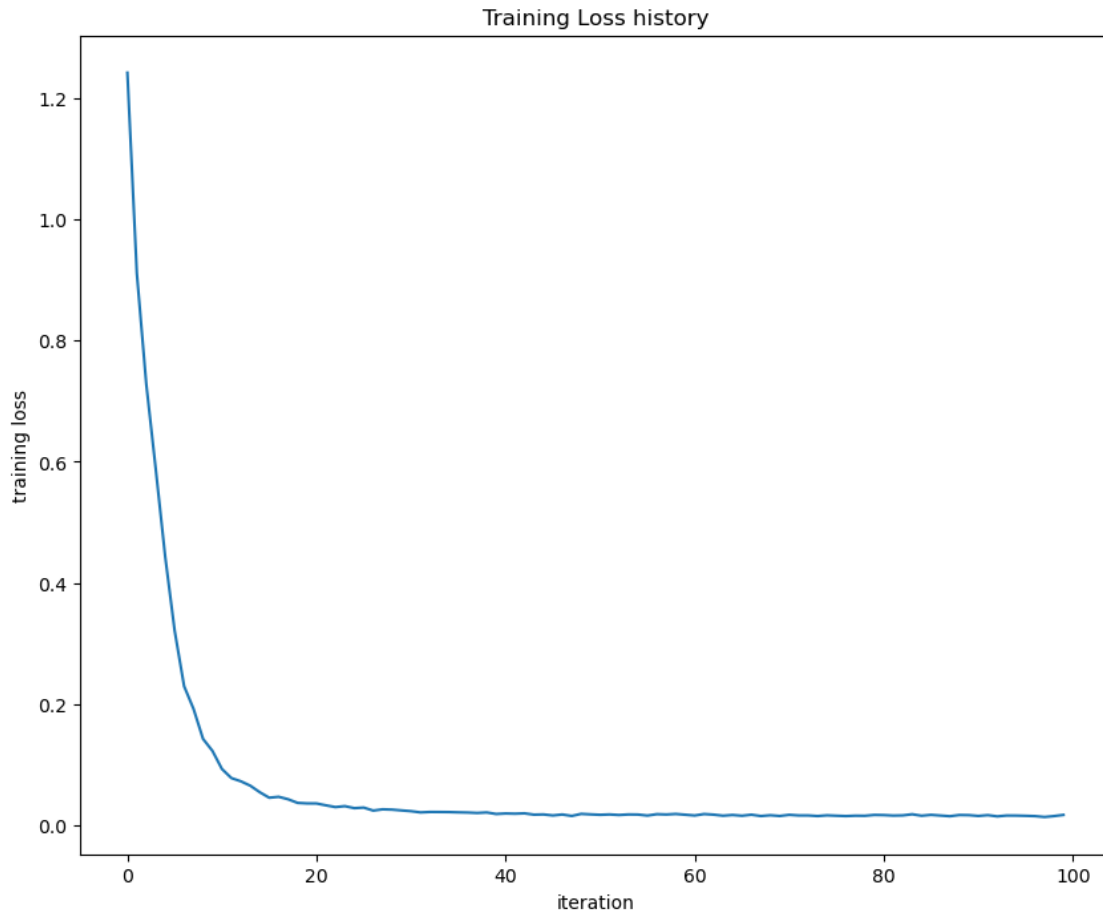
# plot the loss history
plt.plot(stats['loss_history'])
plt.xlabel('iteration')
plt.ylabel('training loss')
plt.title('Training Loss history')
plt.show()

```

```

Final training loss: 0.017143643532923733

```



## 6 Load the data

Now that you have implemented a two-layer network that passes gradient checks and works on toy data, it's time to load up MNIST data so we can use it to train a classifier on a real dataset.

```
[ ]: from mnist_data import load_mnist

def get_mnist_data(num_training=55000, num_validation=5000, num_test=10000):
    """
    Load the MNIST dataset from disk and perform preprocessing to prepare
    it for the two-layer neural net classifier.
    """
    # Load the raw MNIST data
    X_train, y_train, X_val, y_val, X_test, y_test = load_mnist.
    ↪read_data_sets('mnist_data')

    # Normalize the data: subtract the mean image
```

```

mean_image = np.mean(X_train, axis=0)
X_train = X_train - mean_image
X_val = X_val - mean_image
X_test = X_test - mean_image

# Reshape data to rows
X_train = X_train.reshape(num_training, -1)
X_val = X_val.reshape(num_validation, -1)
X_test = X_test.reshape(num_test, -1)

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_mnist_data()
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)

```

```

Extracting mnist_data/train-images-idx3-ubyte.gz
Extracting mnist_data/train-labels-idx1-ubyte.gz
Extracting mnist_data/t10k-images-idx3-ubyte.gz
Extracting mnist_data/t10k-labels-idx1-ubyte.gz
Train data shape: (55000, 784)
Train labels shape: (55000,)
Validation data shape: (5000, 784)
Validation labels shape: (5000,)
Test data shape: (10000, 784)
Test labels shape: (10000,)

```

```

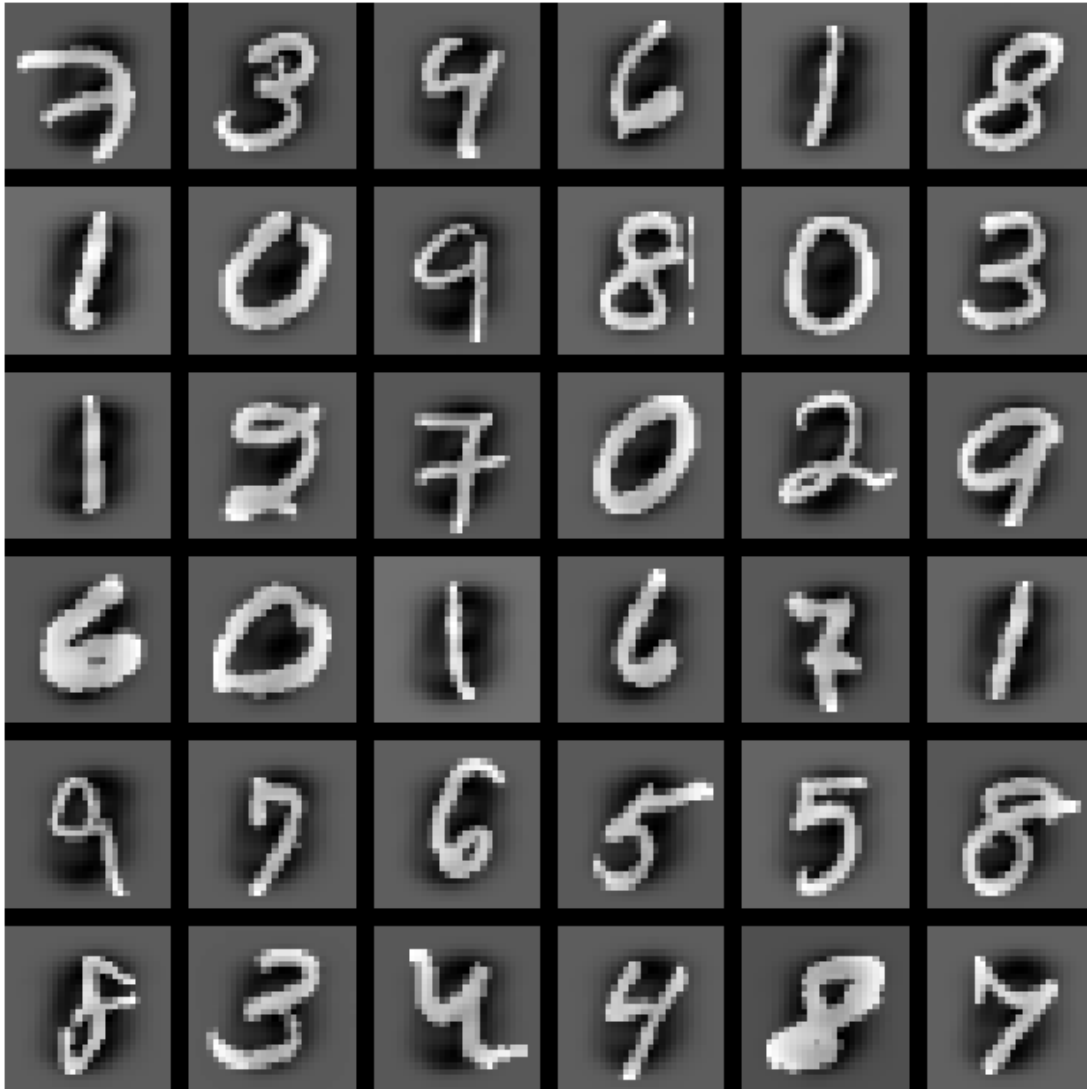
[ ]: from mnist_data.vis_utils import visualize_grid

# Visualize mnist data

def show_mnist_image(data):
    data = data.reshape(-1, 28, 28, 1)
    plt.imshow(visualize_grid(data, padding=3).astype('uint8').squeeze(axis=2))
    plt.gca().axis('off')
    plt.show()

show_mnist_image(X_train[:36])

```



## 7 Train a network

To train our network we will use SGD with momentum. We will use fixed learning rate to train this model.

```
[ ]: input_size = 28 * 28 * 1
      hidden_size = 50
      num_classes = 10
      net = TwoLayerNet(input_size, hidden_size, num_classes)

      # Train the network
      stats = net.train(X_train, y_train, X_val, y_val,
```

```

        num_iters=2000, batch_size=200,
        learning_rate=1e-4, learning_rate_decay=1,
        reg=0, verbose=True)

# Predict on the validation set
val_acc = (net.predict(X_val) == y_val).mean()
print('Validation accuracy: ', val_acc)

```

```

iteration 0 / 2000: loss 2.302571
iteration 100 / 2000: loss 2.302191
iteration 200 / 2000: loss 2.300980
iteration 300 / 2000: loss 2.294519
iteration 400 / 2000: loss 2.248588
iteration 500 / 2000: loss 2.089670
iteration 600 / 2000: loss 1.771894
iteration 700 / 2000: loss 1.313522
iteration 800 / 2000: loss 1.153549
iteration 900 / 2000: loss 0.918654
iteration 1000 / 2000: loss 0.755496
iteration 1100 / 2000: loss 0.705605
iteration 1200 / 2000: loss 0.688104
iteration 1300 / 2000: loss 0.711299
iteration 1400 / 2000: loss 0.528406
iteration 1500 / 2000: loss 0.491089
iteration 1600 / 2000: loss 0.487798
iteration 1700 / 2000: loss 0.455021
iteration 1800 / 2000: loss 0.401196
iteration 1900 / 2000: loss 0.324637
Validation accuracy: 0.8976

```

## 8 Learning Rate Decay

In the previous run, we used the same learning rate during the whole training process. This fix-sized learning rate disregards the benefit of larger learning rate at the beginning of the training, and it might suffer from overshooting around the minima.

Add learning rate decay to the train function, run the model again with larger starting learning rate and learning rate decay, then compare the losses.

```

[ ]: net = TwoLayerNet(input_size, hidden_size, num_classes)
      stats_LRDecay = net.train(X_train, y_train, X_val, y_val,
                               num_iters=2000, batch_size=200,
                               learning_rate=5e-4, learning_rate_decay=0.95,
                               reg=0, verbose=True)

      # Predict on the validation set
      val_acc_LRDecay = (net.predict(X_val) == y_val).mean()
      print('Validation accuracy: ', val_acc_LRDecay)

```

```

iteration 0 / 2000: loss 2.302582
iteration 100 / 2000: loss 2.163480
iteration 200 / 2000: loss 0.991625
iteration 300 / 2000: loss 0.511010
iteration 400 / 2000: loss 0.288836
iteration 500 / 2000: loss 0.381050
iteration 600 / 2000: loss 0.363377
iteration 700 / 2000: loss 0.301115
iteration 800 / 2000: loss 0.268232
iteration 900 / 2000: loss 0.226179
iteration 1000 / 2000: loss 0.245112
iteration 1100 / 2000: loss 0.255568
iteration 1200 / 2000: loss 0.256864
iteration 1300 / 2000: loss 0.260666
iteration 1400 / 2000: loss 0.282674
iteration 1500 / 2000: loss 0.203182
iteration 1600 / 2000: loss 0.221189
iteration 1700 / 2000: loss 0.200646
iteration 1800 / 2000: loss 0.297953
iteration 1900 / 2000: loss 0.289433
Validation accuracy: 0.9422

```

## 9 Debug the training

With the default parameters we provided above, you should get a validation accuracy of about 0.94 on the validation set. This isn't very good for MNIST data which has reports of up to 0.99 accuracy.

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.

(You can think of the first layer weights as a projection  $W^T X$ . This is very similar to how we project our training data using PCA projection in our previous homework. Just like how we visualize the eigenfaces. We can also visualize the weights of the neural network in the same manner.)

Below, we will also show you lossed between two models we trained above. Do you notice the difference between the two?

```

[ ]: # Plot the loss function and train / validation accuracies
plt.subplot(2, 1, 1)
plt.plot(stats['loss_history'], label='with_fixed_LR')
plt.plot(stats_LRDecay['loss_history'], label='with_LR_decay')
plt.title('Loss history')
plt.legend()
plt.xlabel('Iteration')

```

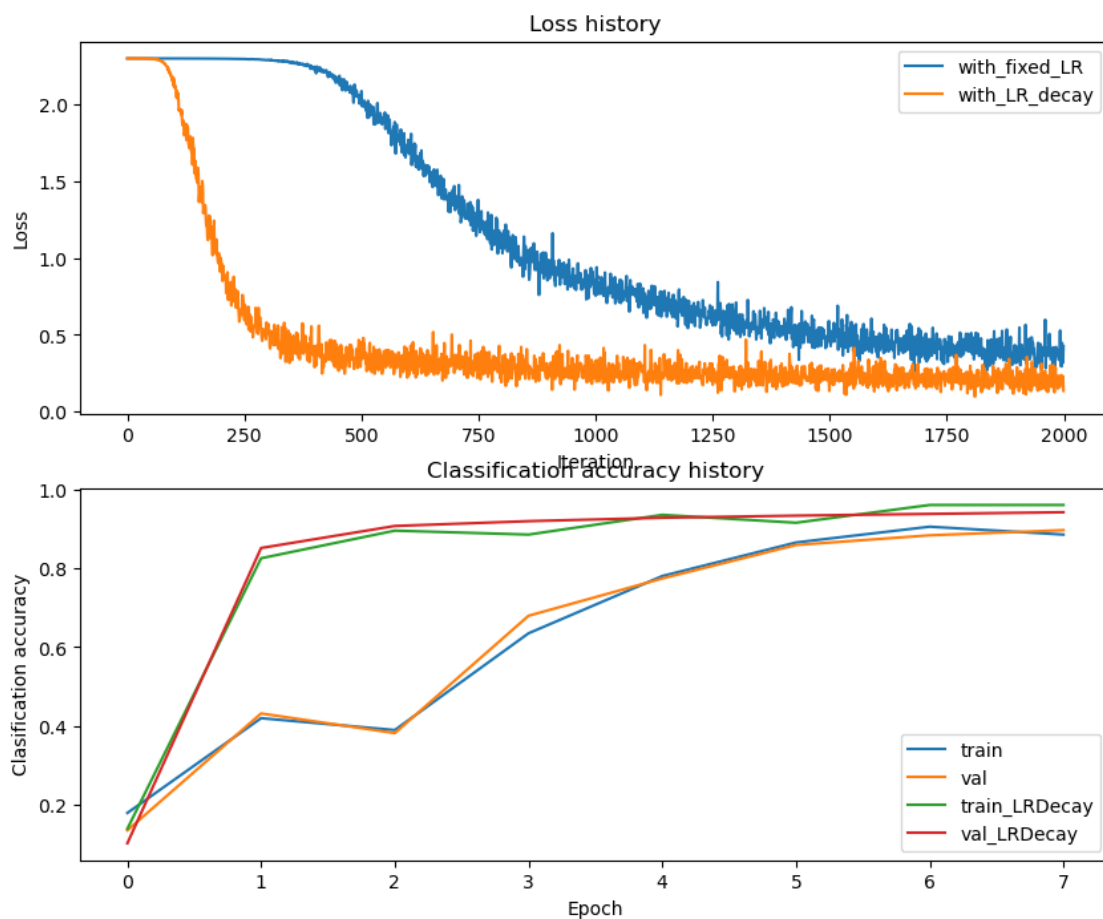


```

plt.ylabel('Loss')

plt.subplot(2, 1, 2)
plt.plot(stats['train_acc_history'], label='train')
plt.plot(stats['val_acc_history'], label='val')
plt.plot(stats_LRDecay['train_acc_history'], label='train_LRDecay')
plt.plot(stats_LRDecay['val_acc_history'], label='val_LRDecay')
plt.title('Classification accuracy history')
plt.legend()
plt.xlabel('Epoch')
plt.ylabel('Clasification accuracy')
plt.show()

```



```

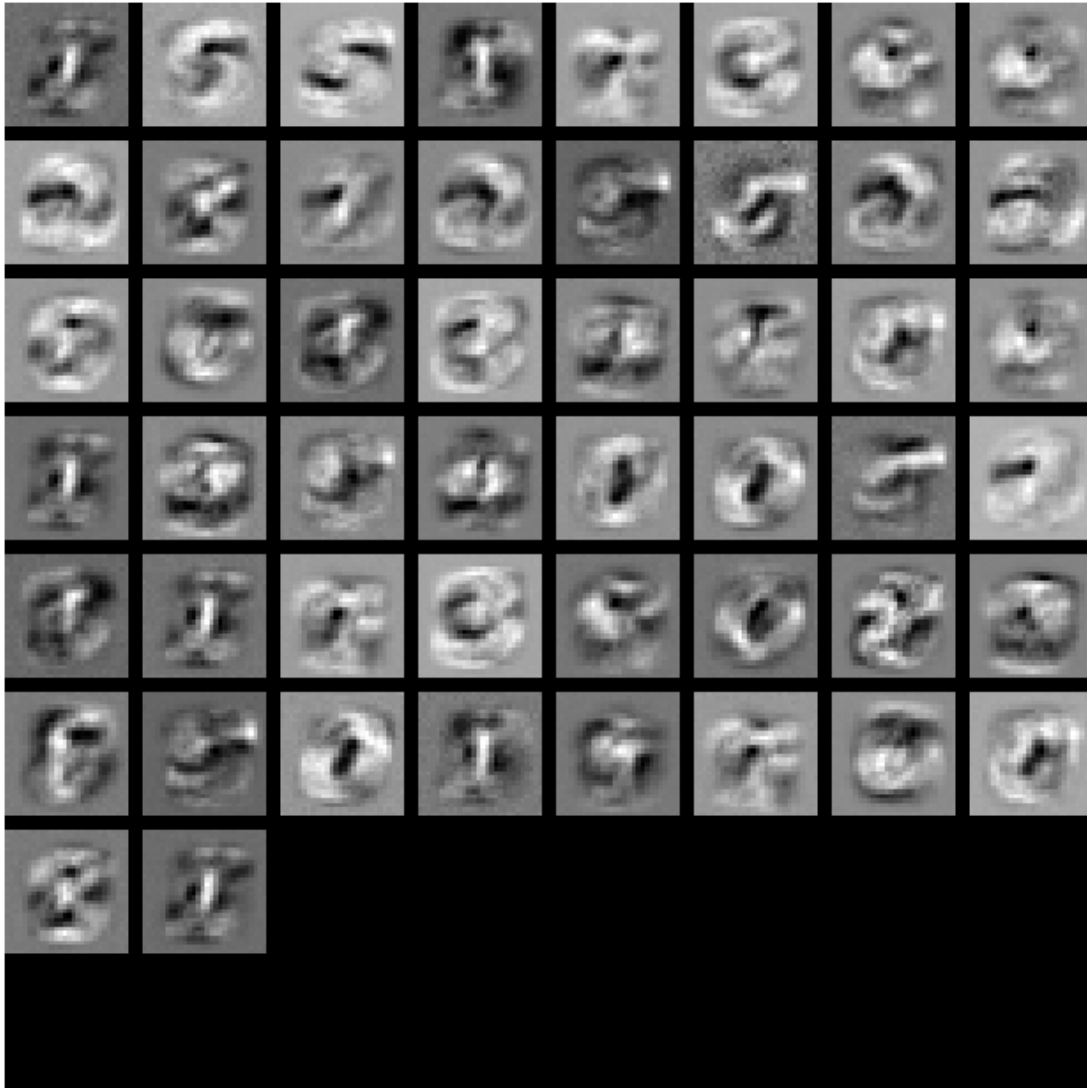
[ ]: # Visualize the weights of the network

def show_net_weights(net):
    W1 = net.params['W1']
    W1 = W1.reshape(28, 28, 1, -1).transpose(3, 0, 1, 2)

```

```
plt.imshow(visualize_grid(W1, padding=3).astype('uint8').squeeze(axis=2))
plt.gca().axis('off')
plt.show()
```

```
show_net_weights(net)
```



## 10 Tune your hyperparameters

**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, number 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 97.4% on the validation set.

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

```
[ ]: best_net = None # store the best model into this

#####
# TODO#8: Tune hyperparameters using the validation set. Store your best
# trained model in best_net.
#
# 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 the previous exercises.
#####

best_val = -1

# Define hyperparameters to search over
hidden_sizes = [50, 100, 128, 200]
learning_rates = [1e-3, 1e-2, 5e-4]
reg_strengths = [1e-3, 5e-4, 5e-6]

# Grid search over hyperparameters
for hidden_size in hidden_sizes:
    for learning_rate in learning_rates:
        for reg in reg_strengths:
            # Create a new neural network
            net = TwoLayerNet(input_size, hidden_size, num_classes)
```

```

# Train the network
stats = net.train(X_train, y_train, X_val, y_val,
                  num_iters=3000, batch_size=200,
                  learning_rate=learning_rate,
↪learning_rate_decay=0.95,
                  reg=reg, verbose=False)

# Predict on the validation set
val_acc = (net.predict(X_val) == y_val).mean()
print(f'hidden_size: {hidden_size}, learning_rate: {learning_rate},
↪reg: {reg}, val_acc: {val_acc}')

# Keep track of the best model
if val_acc > best_val:
    best_val = val_acc
    best_net = net

# Print the best validation accuracy
print('Best validation accuracy: %f' % best_val)

# Evaluate the best model on the test set
test_acc = (best_net.predict(X_test) == y_test).mean()
print('Test accuracy: %f' % test_acc)

#####
#                               END OF TODO#8                               ↪
↪#
#####

```

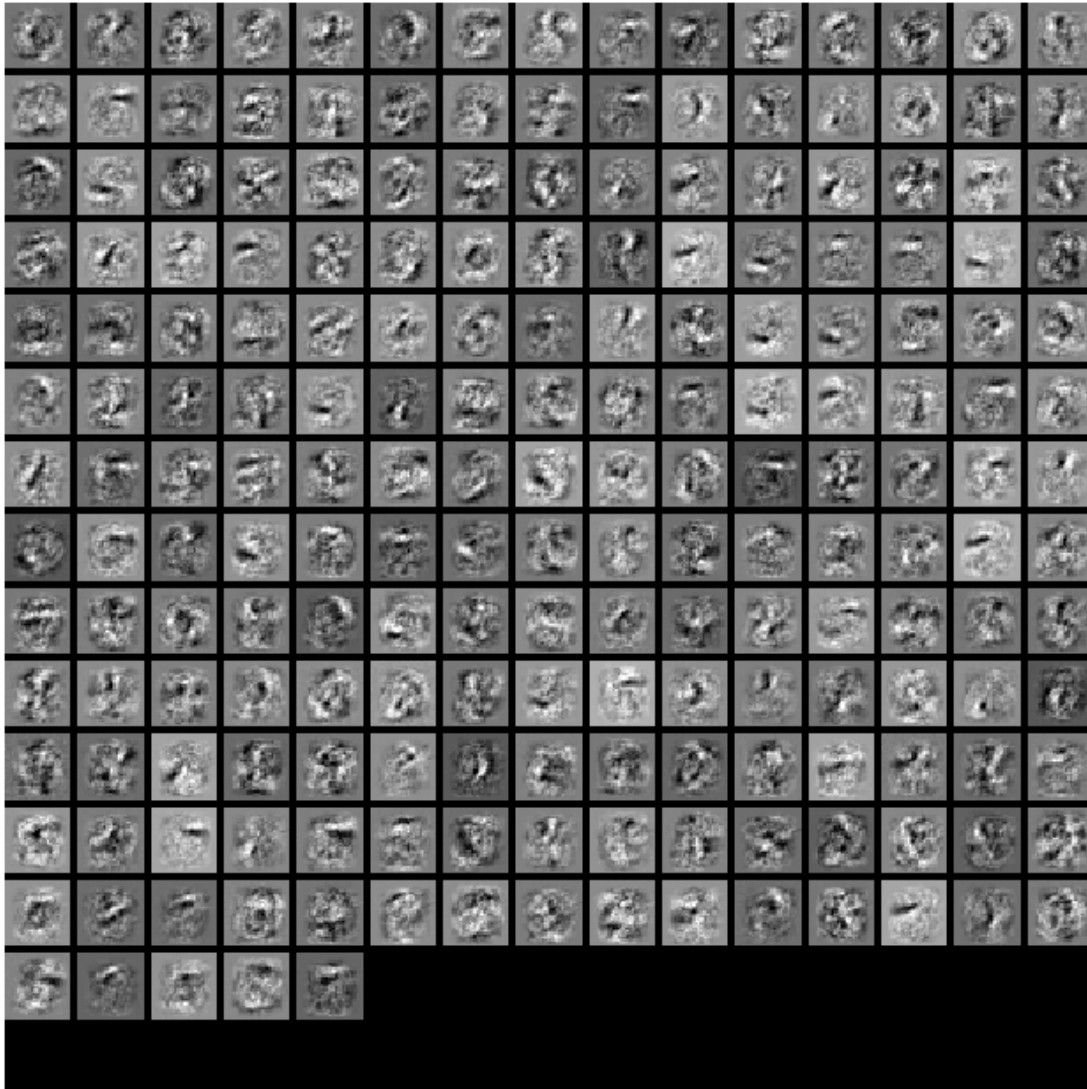
```

hidden_size: 50, learning_rate: 0.001, reg: 0.001, val_acc: 0.9658
hidden_size: 50, learning_rate: 0.001, reg: 0.0005, val_acc: 0.9628
hidden_size: 50, learning_rate: 0.001, reg: 5e-06, val_acc: 0.9652
hidden_size: 50, learning_rate: 0.01, reg: 0.001, val_acc: 0.974
hidden_size: 50, learning_rate: 0.01, reg: 0.0005, val_acc: 0.9742
hidden_size: 50, learning_rate: 0.01, reg: 5e-06, val_acc: 0.9738
hidden_size: 50, learning_rate: 0.0005, reg: 0.001, val_acc: 0.9564
hidden_size: 50, learning_rate: 0.0005, reg: 0.0005, val_acc: 0.9536
hidden_size: 50, learning_rate: 0.0005, reg: 5e-06, val_acc: 0.9534
hidden_size: 100, learning_rate: 0.001, reg: 0.001, val_acc: 0.9682
hidden_size: 100, learning_rate: 0.001, reg: 0.0005, val_acc: 0.9676
hidden_size: 100, learning_rate: 0.001, reg: 5e-06, val_acc: 0.968
hidden_size: 100, learning_rate: 0.01, reg: 0.001, val_acc: 0.98
hidden_size: 100, learning_rate: 0.01, reg: 0.0005, val_acc: 0.9792
hidden_size: 100, learning_rate: 0.01, reg: 5e-06, val_acc: 0.9794
hidden_size: 100, learning_rate: 0.0005, reg: 0.001, val_acc: 0.9548
hidden_size: 100, learning_rate: 0.0005, reg: 0.0005, val_acc: 0.9528

```

```
hidden_size: 100, learning_rate: 0.0005, reg: 5e-06, val_acc: 0.9552
hidden_size: 128, learning_rate: 0.001, reg: 0.001, val_acc: 0.9692
hidden_size: 128, learning_rate: 0.001, reg: 0.0005, val_acc: 0.9664
hidden_size: 128, learning_rate: 0.001, reg: 5e-06, val_acc: 0.97
hidden_size: 128, learning_rate: 0.01, reg: 0.001, val_acc: 0.9826
hidden_size: 128, learning_rate: 0.01, reg: 0.0005, val_acc: 0.9808
hidden_size: 128, learning_rate: 0.01, reg: 5e-06, val_acc: 0.9818
hidden_size: 128, learning_rate: 0.0005, reg: 0.001, val_acc: 0.9554
hidden_size: 128, learning_rate: 0.0005, reg: 0.0005, val_acc: 0.9554
hidden_size: 128, learning_rate: 0.0005, reg: 5e-06, val_acc: 0.9554
hidden_size: 200, learning_rate: 0.001, reg: 0.001, val_acc: 0.97
hidden_size: 200, learning_rate: 0.001, reg: 0.0005, val_acc: 0.9694
hidden_size: 200, learning_rate: 0.001, reg: 5e-06, val_acc: 0.9674
hidden_size: 200, learning_rate: 0.01, reg: 0.001, val_acc: 0.9814
hidden_size: 200, learning_rate: 0.01, reg: 0.0005, val_acc: 0.9816
hidden_size: 200, learning_rate: 0.01, reg: 5e-06, val_acc: 0.9828
hidden_size: 200, learning_rate: 0.0005, reg: 0.001, val_acc: 0.9568
hidden_size: 200, learning_rate: 0.0005, reg: 0.0005, val_acc: 0.9566
hidden_size: 200, learning_rate: 0.0005, reg: 5e-06, val_acc: 0.9566
Best validation accuracy: 0.982800
Test accuracy: 0.981200
```

```
[ ]: # visualize the weights of the best network
      show_net_weights(best_net)
```



## 11 Run on the test set

When you are done experimenting, you should evaluate your final trained network on the test set; you should get above 96.3%.

```
[ ]: test_acc = (best_net.predict(X_test) == y_test).mean()
      print('Test accuracy: ', test_acc)
```

Test accuracy: 0.9812