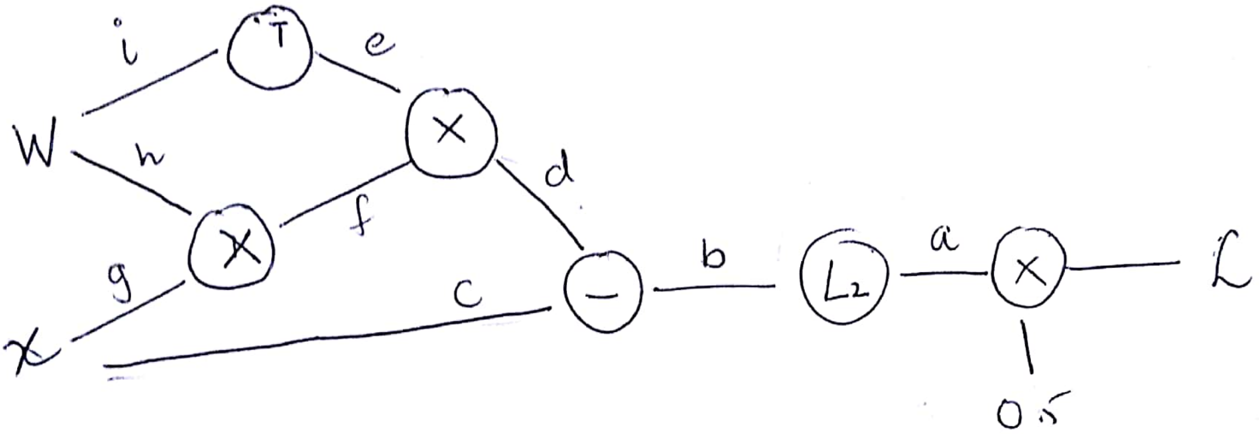
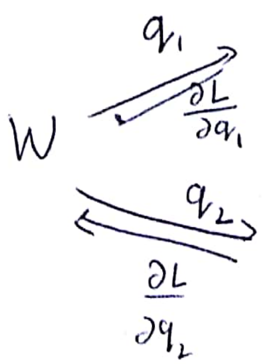
HW3

(1) (a) W^T gets the m-dimensional vector Wx back to $\mathbb{R}^n$. ~~Then~~ ~~is~~ The norm of this vector should be as close as x for most of the information to be preserved. Thus the norm of $(W^TWx - x)$ must be minimized to get the best m-dimensional vector that approximates the n-dimensional vector.

(b)



(c) We account for these 2 paths by adding the gradients from each of the path.



$$\frac{\partial L(q_1, q_2)}{\partial W} = \underbrace{\frac{\partial q_1}{\partial W}}_{1}\frac{\partial L}{\partial q_1} + \underbrace{\frac{\partial q_2}{\partial W}}_{1}\frac{\partial L}{\partial q_2}$$

$$= \frac{\partial L}{\partial q_1} + \frac{\partial L}{\partial q_2}$$

(9) In the computational graph.

$$L = 0.5a$$

$$\frac{\partial L}{\partial a} = 0.5 \qquad (1)$$

$$a = b^T b$$

$$\frac{\partial a}{\partial b} = 2b \qquad \frac{\partial L}{\partial b} = 0.5 \times 2b = b$$

$$b = d - c$$

$$\frac{\partial b}{\partial d} = 1 \qquad \therefore \frac{\partial L}{\partial d} = 1 \cdot b = b$$

$\cancel{ef = d}$ $d = ef$

$$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial d} f^T = b f^T$$

$$\frac{\partial L}{\partial f} = e^T \frac{\partial L}{\partial d} = e^T b$$

$$f = hg$$

$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial f} \cdot g^T = e^T b g^T = (W^T)^T (W^T W x - x) x^T$$

$$= W (W^T W x - x) x^T$$

$$e = i^T$$

$$\frac{\partial L}{\partial i} = \left(\frac{\partial L}{\partial e}\right)^T$$

$$= (b f^T)^T$$

$$= f\, b^T$$

$$= Wx\,(W^T W x - x)^T$$

*

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial h} + \frac{\partial L}{\partial i}$$

$$= \boxed{W\,(W^T W x - x)x^T + Wx\,(W^T W x - x)^T}$$

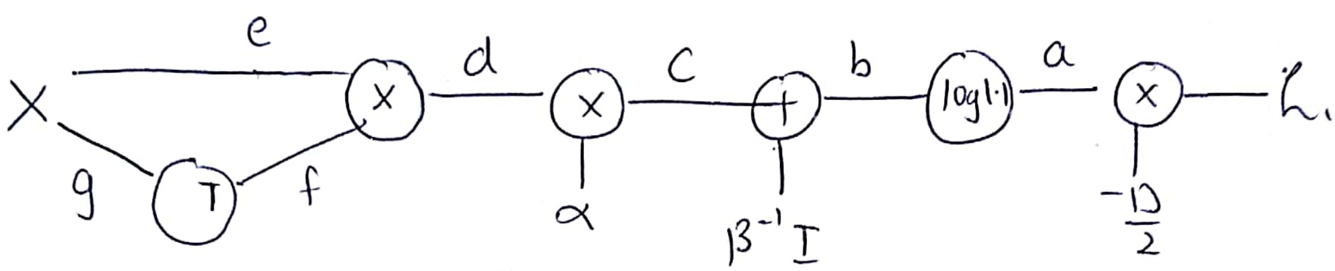
(2) $$L = -C - \frac{D}{2}\log|K| - \frac{1}{2}\operatorname{tr}(K^{-1}YY^T)$$

$$K = \alpha XX^T + \beta^{-1} I$$

$$L_1 = -\frac{D}{2}\log|\alpha XX^T + \beta^{-1} I|$$

$$L_2 = -\frac{1}{2}\operatorname{tr}\left((\alpha XX^T + \beta^{-1} I)^{-1} YY^T\right)$$

(a)



(b) $$L_1 = -\frac{D}{2} a$$

$$\frac{\partial L_1}{\partial a} = -\frac{D}{2}$$

$\cancel{\frac{\partial a}{\partial b}}$ $a = \log|b|$

$$\frac{\partial a}{\partial b} = \frac{\partial \log|b|}{\partial b} = b^{-T} \quad \therefore \frac{\partial L_1}{\partial b} = -b^{-T}\frac{D}{2}$$

$$b = c + \beta^{-1} I$$

$$\frac{\partial b}{\partial c} = 1 \quad \therefore \frac{\partial L_1}{\partial c} = -b^{-T}\frac{D}{2}$$

$$c = \alpha d$$

$$\frac{\partial c}{\partial d} = \alpha \qquad \therefore \frac{\partial L_1}{\partial d} = -\alpha\, b^{-T} \frac{D}{2}$$

$$d = e f$$

$$\frac{\partial L_1}{\partial e} = \frac{\partial L_1}{\partial d} f^T = -\alpha\, b^{-T} \frac{D}{2} f^T$$

$$\frac{\partial L_1}{\partial f} = e^T \frac{\partial L_1}{\partial d} = -e^T \alpha\, b^{-T} \frac{D}{2}$$

$$\partial f = g^T$$

$$\cancel{\frac{\partial g}{\partial f} = \frac{\partial f}{\partial g}}$$

$$\frac{\partial L_1}{\partial g} = \left(\frac{\partial L}{\partial f}\right)^T = -\alpha \frac{D}{2} \left(e^T b^{-T}\right)^T$$

$$= -\alpha \frac{D}{2} \left(b^{-1} e\right)$$

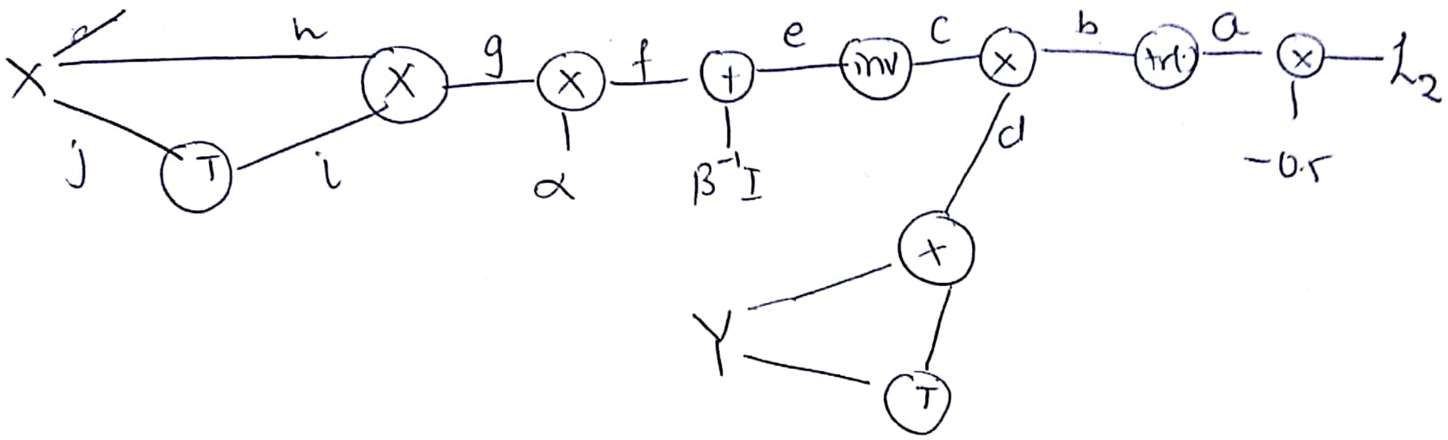
$$\therefore \frac{\partial L_1}{\partial X} = \frac{\partial L_1}{\partial e} + \frac{\partial L_1}{\partial g}$$

$$= -\alpha \frac{D}{2} \left(b^{-T} f^T + b^{-1} e \right)$$

$$= -\alpha \frac{D}{2} \left(\left(\alpha X X^T + \beta^{-1} I\right)^{-T} X + \left(\alpha X X^T + \beta^{-1} I\right)^{-1} X \right)$$

$$\boxed{\therefore \frac{\partial L_1}{\partial X} = -\alpha D\left(\alpha X X^T + \beta^{-1} I\right)^{-1} X}$$

(c)



(d) $L_2 = -0.5a$

$$\frac{\partial L_2}{\partial a} = -0.5$$

$$a = tr(b)$$

$$\frac{\partial a}{\partial b} = \frac{\partial\, tr(b)}{\partial b} = I \quad \therefore \frac{\partial L_2}{\partial b} = -0.5 I$$

$$c = e^{-1}$$

$$\frac{\partial c}{\partial e} = -e^{-T} \frac{\partial c}{\partial e^{-1}} e^{-T} = -e^{-T} e^{-T}$$

$$\therefore \frac{\partial L_2}{\partial e} = -e^{-T} e^{-T} \frac{\partial L_2}{\partial c}$$

$$b = cd$$

$$\frac{\partial b}{\partial c} = d$$

$$\therefore \frac{\partial L_2}{\partial c} = d \frac{\partial L_2}{\partial b} = \frac{-d}{2}$$

$$\frac{\partial L_2}{\partial e} = -e^{-T} e^{-T} \frac{\partial L_2}{\partial c} = \frac{+e^{-T} e^{-T} d}{2}$$

$$e = f + \beta^{-1} I$$

$$\frac{\partial e}{\partial f} = 1 \quad \therefore \quad \frac{\partial L_2}{\partial f} = \cancel{\frac{\partial b_2}{\partial}} \frac{\partial e}{\partial f} \frac{\partial L_2}{\partial e} = \frac{e^{-T} e^{-T} d}{2}$$

$$f = \alpha g$$

$$\frac{\partial f}{\partial g} = \alpha \quad \therefore \quad \frac{\partial L_2}{\partial g} = \frac{\partial f}{\partial g} \frac{\partial L_2}{\partial f} = \alpha \frac{e^{-T} e^{-T} d}{2}$$

$$g = hi$$

$$\frac{\partial L_2}{\partial h} = \frac{\partial L_2}{\partial g} i^T = \frac{\alpha}{2} e^{-T} e^{-T} d \, i^T$$

$$\frac{\partial L_2}{\partial i} = h^T \frac{\partial L_2}{\partial g} = \frac{\alpha}{2} h^T e^{-T} e^{-T} d$$

$$i = j^T$$

$$\frac{\partial L_2}{\partial j} = \left(\frac{\partial L_2}{\partial i}\right)^T$$

$$= \left(\frac{\alpha}{2} h^T e^{-T} e^{-T} d\right)^T$$

$$= \frac{\alpha}{2} d^T e^{-1} e^{-1} h.$$

$$\frac{\partial L_2}{\partial X} = \frac{\partial L_2}{\partial h} + \frac{\partial L_2}{\partial j}$$

$$= \frac{\alpha}{2} e^{-T} e^{-T} d\, i^T + \frac{\alpha}{2} d^T e^{-1} e^{-1} h$$

$$= \frac{\alpha}{2}\Big[\left((\alpha XX^T + \beta^{-1}I)^{-T}(\alpha XX^T + \beta^{-1}I)^{-T} YY^T X\right)$$

$$+ \left(YY^T (\alpha XX^T + \beta^{-1}I)^{-2} X\right)\Big]$$

$$= \frac{\alpha}{2}\left[(\alpha XX^T + \beta^{-1}I)^{-2} YY^T + YY^T(\alpha XX^T + \beta^{-1}I)^{-2}\right] X$$

$$\therefore \boxed{\frac{\partial L_2}{\partial X} = \frac{\alpha}{2}\left[(\alpha XX^T + \beta^{-1}I)^{-2} YY^T + YY^T(\alpha XX^T + \beta^{-1}I)^{-2}\right] X}$$

(e) $\frac{\partial L}{\partial X}$

$$\mathcal{L} = -C + \mathcal{L}_1 + \mathcal{L}_2$$

$$\frac{\partial \mathcal{L}}{\partial X} = -\underbrace{\frac{\partial C}{\partial X}}_{0} + \frac{\partial L_1}{\partial X} + \frac{\partial L_2}{\partial X}$$

$$= \frac{\partial L_1}{\partial X} + \frac{\partial L_2}{\partial X}$$

$$= \boxed{-\alpha D(\alpha X X^T + \beta^{-1} I)^{-1} X + \frac{\alpha}{2}\left[(\alpha X X^T + \beta^{-1} I)^{-2} Y Y^T + Y Y^T (\alpha X X^T + \beta^{-1} I)^{-2}\right] X}$$

```python
import numpy as np
import matplotlib.pyplot as plt

"""
This code was originally written for CS 231n at Stanford University
(cs231n.stanford.edu).  It has been modified in various areas for use in the
ECE 239AS class at UCLA.  This includes the descriptions of what code to
implement as well as some slight potential changes in variable names to be
consistent with class nomenclature.  We thank Justin Johnson & Serena Yeung for
permission to use this code.  To see the original version, please visit
cs231n.stanford.edu.
"""

class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network. The net has an input dimension of
  N, a hidden layer dimension of H, and performs classification over C classes.
  We train the network with a softmax loss function and L2 regularization on the
  weight matrices. The network uses a ReLU nonlinearity after the first fully
  connected layer.

  In other words, the network has the following architecture:

  input - fully connected layer - ReLU - fully connected layer - softmax

  The outputs of the second fully-connected layer are the scores for each class.
  """

  def __init__(self, input_size, hidden_size, output_size, std=1e-4):
    """
    Initialize the model. Weights are initialized to small random values and
    biases are initialized to zero. Weights and biases are stored in the
    variable self.params, which is a dictionary with the following keys:

    W1: First layer weights; has shape (H, D)
    b1: First layer biases; has shape (H,)
    W2: Second layer weights; has shape (C, H)
    b2: Second layer biases; has shape (C,)

    Inputs:
    - input_size: The dimension D of the input data.
    - hidden_size: The number of neurons H in the hidden layer.
    - output_size: The number of classes C.
    """
    self.params = {}
    self.params['W1'] = std * np.random.randn(hidden_size, input_size)
    self.params['b1'] = np.zeros(hidden_size)
    self.params['W2'] = std * np.random.randn(output_size, hidden_size)
    self.params['b2'] = np.zeros(output_size)


  def loss(self, X, y=None, reg=0.0):
    """
    Compute the loss and gradients for a two layer fully connected neural
    network.

    Inputs:
    - X: Input data of shape (N, D). Each X[i] is a training sample.
    - y: Vector of training labels. y[i] is the label for X[i], and each y[i] is
      an integer in the range 0 <= y[i] < C. This parameter is optional; if it
      is not passed then we only return scores, and if it is passed then we
```

```python
      instead return the loss and gradients.
    - reg: Regularization strength.

    Returns:
    If y is None, return a matrix scores of shape (N, C) where scores[i, c] is
    the score for class c on input X[i].

    If y is not None, instead return a tuple of:
    - loss: Loss (data loss and regularization loss) for this batch of training
      samples.
    - grads: Dictionary mapping parameter names to gradients of those parameters
      with respect to the loss function; has the same keys as self.params.
    """
    # Unpack variables from the params dictionary
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    N, D = X.shape

    # Compute the forward pass
    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the output scores of the neural network.  The result
    #   should be (N, C). As stated in the description for this class,
    #   there should not be a ReLU layer after the second FC layer.
    #   The output of the second FC layer is the output scores. Do not
    #   use a for loop in your implementation.
    # ================================================================ #
    a1 = W1.dot(X.T) + b1.reshape((b1.shape[0],1))
    h1 = np.maximum(0,a1)
    scores = W2.dot(h1) + b2.reshape((b2.shape[0],1))
    scores = scores.T


    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #


    # If the targets are not given then jump out, we're done
    if y is None:
      return scores

    # Compute the loss
    loss = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the loss of the neural network.  This includes the
    #   softmax loss and the L2 regularization for W1 and W2. Store the
    #   total loss in the variable loss.  Multiply the regularization
    #   loss by 0.5 (in addition to the factor reg).
    # ================================================================ #

    # scores is num_examples by num_classes
    a = scores.T
    ind = np.arange(y.shape[0])
    ay = a[y,ind]
    lgk = -1*np.amax(a,axis=0)
    zexp = np.exp(a+lgk)
```

```python
        zsum = np.sum(zexp,axis=0)
        loss_arr = np.log(zsum) - ay -lgk
        loss = np.mean(loss_arr)
        l2_loss = 0.5*reg*(np.sum(np.square(W1)) + np.sum(np.square(W2)))
        loss += l2_loss

        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

        grads = {}

        # ================================================================ #
        # YOUR CODE HERE:
        #   Implement the backward pass.  Compute the derivatives of the
        #   weights and the biases.  Store the results in the grads
        #   dictionary.  e.g., grads['W1'] should store the gradient for
        #   W1, and be of the same size as W1.
        # ================================================================ #

        db2 = np.divide(zexp,zsum)
        db2[y,ind] -= 1
        dW2 = db2.dot(h1.T)
        DW2 = dW2/N
        Db2 = db2.sum(axis=1)/N
        dh1 = (W2.T).dot(db2)
        da1 = np.multiply(np.sign(np.maximum(0,a1)),dh1)
        db1 = da1
        Db1 = db1.sum(axis=1)/N
        dW1 = da1.dot(X)
        DW1 = dW1/N
        grads['W1']=DW1 + reg*W1
        grads['b1'] = Db1
        grads['W2'] = DW2 + reg*W2
        grads['b2'] = Db2

        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

        return loss, grads

    def train(self, X, y, X_val, y_val,
              learning_rate=1e-3, learning_rate_decay=0.95,
              reg=1e-5, num_iters=100,
              batch_size=200, verbose=False):
        """
        Train this neural network using stochastic gradient descent.

        Inputs:
        - X: A numpy array of shape (N, D) giving training data.
        - y: A numpy array f shape (N,) giving training labels; y[i] = c means that
          X[i] has label c, where 0 <= c < C.
        - X_val: A numpy array of shape (N_val, D) giving validation data.
        - y_val: A numpy array of shape (N_val,) giving validation labels.
        - learning_rate: Scalar giving learning rate for optimization.
        - learning_rate_decay: Scalar giving factor used to decay the learning rate
          after each epoch.
        - reg: Scalar giving regularization strength.
        - num_iters: Number of steps to take when optimizing.
        - batch_size: Number of training examples to use per step.
```

```python
    - verbose: boolean; if true print progress during optimization.
    """
    num_train = X.shape[0]
    iterations_per_epoch = max(num_train / batch_size, 1)

    # Use SGD to optimize the parameters in self.model
    loss_history = []
    train_acc_history = []
    val_acc_history = []

    for it in np.arange(num_iters):
      X_batch = None
      y_batch = None

      # ================================================================ #
      # YOUR CODE HERE:
      #   Create a minibatch by sampling batch_size samples randomly.
      # ================================================================ #
      ind = np.random.choice(num_train,batch_size)
      X_batch = X[ind]
      y_batch = y[ind]

      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

       # Compute loss and gradients using the current minibatch
      loss, grads = self.loss(X_batch, y=y_batch, reg=reg)
      loss_history.append(loss)

      # ================================================================ #
      # YOUR CODE HERE:
      #   Perform a gradient descent step using the minibatch to update
      #   all parameters (i.e., W1, W2, b1, and b2).
      # ================================================================ #
      self.params['W1'] -= learning_rate*grads['W1']
      self.params['W2'] -= learning_rate*grads['W2']
      self.params['b1'] -= learning_rate*grads['b1']
      self.params['b2'] -= learning_rate*grads['b2']


      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

      if verbose and it % 100 == 0:
        print('iteration {} / {}: loss {}'.format(it, num_iters, loss))

      # Every epoch, check train and val accuracy and decay learning rate.
      if it % iterations_per_epoch == 0:
        # Check accuracy
        train_acc = (self.predict(X_batch) == y_batch).mean()
        val_acc = (self.predict(X_val) == y_val).mean()
        train_acc_history.append(train_acc)
        val_acc_history.append(val_acc)

        # Decay learning rate
        learning_rate *= learning_rate_decay

    return {
      'loss_history': loss_history,
```

```python
        'train_acc_history': train_acc_history,
        'val_acc_history': val_acc_history,
      }

  def predict(self, X):
    """
    Use the trained weights of this two-layer network to predict labels for
    data points. For each data point we predict scores for each of the C
    classes, and assign each data point to the class with the highest score.

    Inputs:
    - X: A numpy array of shape (N, D) giving N D-dimensional data points to
      classify.

    Returns:
    - y_pred: A numpy array of shape (N,) giving predicted labels for each of
      the elements of X. For all i, y_pred[i] = c means that X[i] is predicted
      to have class c, where 0 <= c < C.
    """
    y_pred = None
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    N, D = X.shape

    # ================================================================ #
    # YOUR CODE HERE:
    #   Predict the class given the input data.
    # ================================================================ #
    a1 = W1.dot(X.T) + b1.reshape((b1.shape[0],1))
    h1 = np.maximum(0,a1)
    scores = W2.dot(h1) + b2.reshape((b2.shape[0],1))
    y_pred = np.argmax(scores,axis=0)


    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return y_pred
```

This is the 2-layer neural network workbook for ECE 239AS Assignment #3

Please follow the notebook linearly to implement a two layer neural network.

Please print out the workbook entirely when completed.

We thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class (cs231n.stanford.edu). These are the functions in the cs231n folders and code in the jupyer notebook to preprocess and show the images. The classifiers used are based off of code prepared for CS 231n as well.

The goal of this workbook is to give you experience with training a two layer neural network.

In [1]:

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

%matplotlib inline
%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))))
```

Toy example

Before loading CIFAR-10, there will be a toy example to test your implementation of the forward and backward pass

In [2]:

```
from nndl.neural_net import TwoLayerNet
```

In [3]:

```
# 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()
```

Compute forward pass scores

In [4]:

```
## Implement the forward pass of the neural network.

# Note, there is a statement if y is None: return scores, which is why
# the following call will calculate the scores.
scores = net.loss(X)
print('Your scores:')
print(scores)
print()
print('correct scores:')
correct_scores = np.asarray([
    [-1.07260209,  0.05083871, -0.87253915],
    [-2.02778743, -0.10832494, -1.52641362],
    [-0.74225908,  0.15259725, -0.39578548],
    [-0.38172726,  0.10835902, -0.17328274],
    [-0.64417314, -0.18886813, -0.41106892]])
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:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
 [-0.64417314 -0.18886813 -0.41106892]]

correct scores:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
 [-0.64417314 -0.18886813 -0.41106892]]

Difference between your scores and correct scores:
3.381231222787662e-08
```

Forward pass loss

In [5]:

```
loss, _ = net.loss(X, y, reg=0.05)
correct_loss = 1.071696123862817

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

```
Difference between your loss and correct loss:
0.0
```

In [6]:

```
print(loss)
```

```
1.071696123862817
```

Backward pass

Implements the backwards pass of the neural network. Check your gradients with the gradient check utilities provided.

In [7]:

```
from cs231n.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('{} max relative error: {}'.format(param_name, rel_error(param_grad_num, grad
s[param_name])))
```

```
W1 max relative error: 1.2832788797639875e-09
b1 max relative error: 1.7679559573203866e-09
W2 max relative error: 2.9632221903873815e-10
b2 max relative error: 1.8391536011791491e-10
```

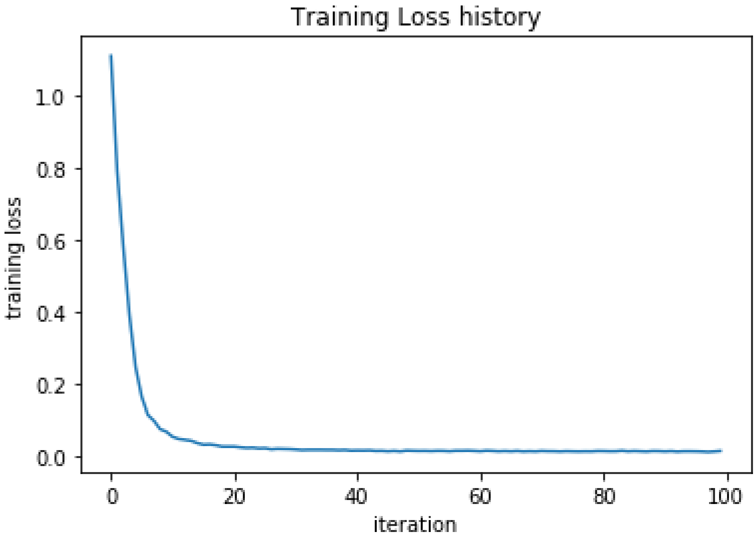
Training the network

Implement neural_net.train() to train the network via stochastic gradient descent, much like the softmax and SVM.

In [8]:

```
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:  0.014497864587765804
```



Classify CIFAR-10

Do classification on the CIFAR-10 dataset.

In [9]:

```
from cs231n.data_utils import load_CIFAR10

def get_CIFAR10_data(num_training=49000, num_validation=1000, num_test=1000):
    """
    Load the CIFAR-10 dataset from disk and perform preprocessing to prepare
    it for the two-layer neural net classifier. These are the same steps as
    we used for the SVM, but condensed to a single function.
    """
    # Load the raw CIFAR-10 data
    cifar10_dir = 'cifar-10-batches-py'
    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]

    # Normalize the data: subtract the mean image
    mean_image = np.mean(X_train, axis=0)
    X_train -= mean_image
    X_val -= mean_image
    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


# Invoke the above function to get our data.
X_train, y_train, X_val, y_val, X_test, y_test = get_CIFAR10_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)
```

```
Train data shape:  (49000, 3072)
Train labels shape:  (49000,)
Validation data shape:  (1000, 3072)
Validation labels shape:  (1000,)
Test data shape:  (1000, 3072)
Test labels shape:  (1000,)
```

Running SGD

If your implementation is correct, you should see a validation accuracy of around 28-29%.

In [10]:

```
input_size = 32 * 32 * 3
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=1000, batch_size=200,
            learning_rate=1e-4, learning_rate_decay=0.95,
            reg=0.25, verbose=True)

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

# Save this net as the variable subopt_net for later comparison.
subopt_net = net
```

```
iteration 0 / 1000: loss 2.302757518613176
iteration 100 / 1000: loss 2.302120159207236
iteration 200 / 1000: loss 2.2956136007408703
iteration 300 / 1000: loss 2.2518259043164135
iteration 400 / 1000: loss 2.188995235046776
iteration 500 / 1000: loss 2.1162527791897747
iteration 600 / 1000: loss 2.064670827698217
iteration 700 / 1000: loss 1.9901688623083942
iteration 800 / 1000: loss 2.002827640124685
iteration 900 / 1000: loss 1.9465176817856495
Validation accuracy:  0.283
```

Questions:

The training accuracy isn't great.

(1) What are some of the reasons why this is the case? Take the following cell to do some analyses and then report your answers in the cell following the one below.

(2) How should you fix the problems you identified in (1)?

In [11]:

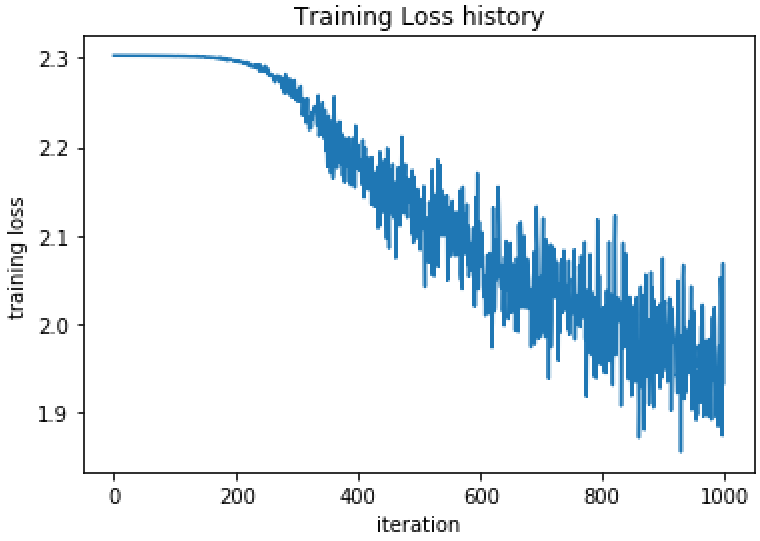
```
stats['train_acc_history']
```

Out[11]:

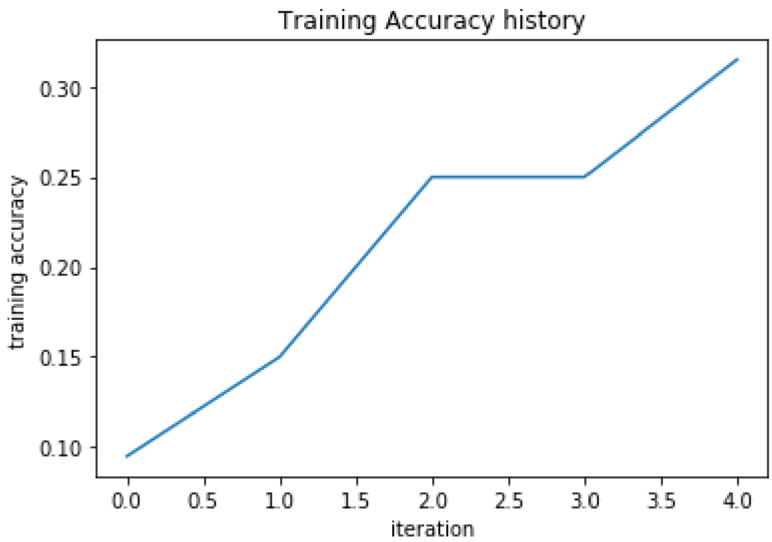
```
[0.095, 0.15, 0.25, 0.25, 0.315]
```

In [12]:

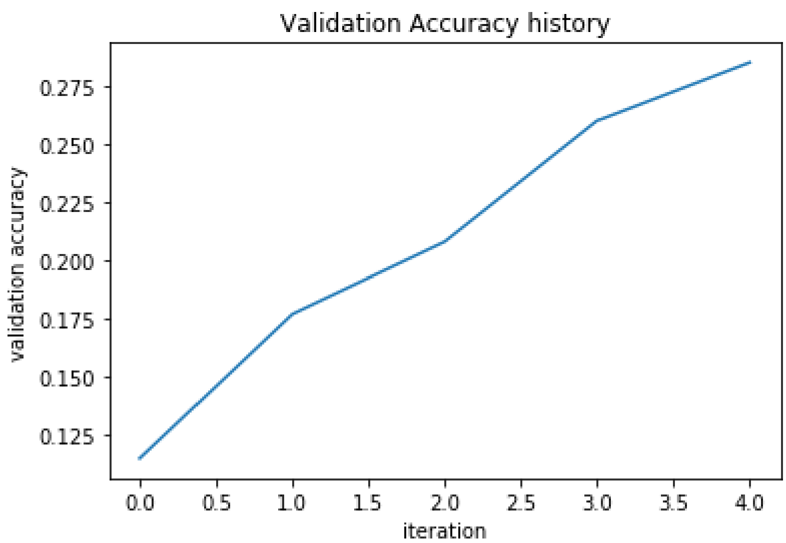
```
# ================================================================ #
# YOUR CODE HERE:
#   Do some debugging to gain some insight into why the optimization
#   isn't great.
# ================================================================ #

# Plot the loss function and train / validation accuracies

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

plt.plot(stats['train_acc_history'])
plt.xlabel('iteration')
plt.ylabel('training accuracy')
plt.title('Training Accuracy history')
plt.show()

plt.plot(stats['val_acc_history'])
plt.xlabel('iteration')
plt.ylabel('validation accuracy')
plt.title('Validation Accuracy history')
plt.show()
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```



Training Loss history
training loss
2.3
2.2
2.1
2.0
1.9
0
200
400
600
800
1000
iteration



Training Accuracy history
training accuracy
0.30
0.25
0.20
0.15
0.10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
iteration



Validation Accuracy history
validation accuracy
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
iteration

Answers:

(1) As can be seen in the plots they have not yet saturated. This is because the number of iterations is too less or the learning rate is too small.

(2) The way to fix this problem is by increasing the number of iterations and/or increase the learning rate.

Optimize the neural network

Use the following part of the Jupyter notebook to optimize your hyperparameters on the validation set. Store your nets as best_net.

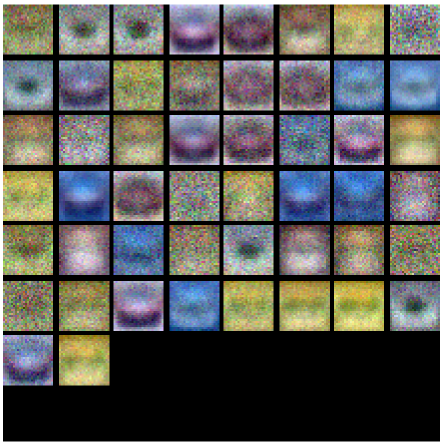
In [16]:

```
import numpy as np
best_net = None # store the best model into this

# ================================================================ #
# YOUR CODE HERE:
#   Optimize over your hyperparameters to arrive at the best neural
#   network.  You should be able to get over 50% validation accuracy.
#   For this part of the notebook, we will give credit based on the
#   accuracy you get.  Your score on this question will be multiplied by:
#      min(floor((X - 28%)) / %22, 1)
#   where if you get 50% or higher validation accuracy, you get full
#   points.
#
#   Note, you need to use the same network structure (keep hidden_size = 50)!
# ================================================================ #
input_size = 32 * 32 * 3
hidden_size = 50
num_classes = 10
iters = 5000 #fix iteration number
learning_rates = [1e-5,5e-5,1e-4,5e-4,1e-3]
lrn = np.shape(learning_rates)[0]
val_accs = np.zeros((lrn,1))
for i in range(lrn):
    net = TwoLayerNet(input_size, hidden_size, num_classes)
    stats = net.train(X_train, y_train, X_val, y_val,
                      num_iters=iters, batch_size=200,
                      learning_rate=learning_rates[i], learning_rate_decay=0.95,
                      reg=0.25, verbose=False)
    # Predict on the validation set
    val_accs[i] = (net.predict(X_val) == y_val).mean()
    print('Validation accuracy: {} for learning rate: {}'.format(val_accs[i],learning_r
ates[i]))

ii = np.argmax(val_accs)
best_lr = learning_rates[ii]
print('Best learning rate: {}'.format(best_lr))

net = TwoLayerNet(input_size, hidden_size, num_classes)
stats = net.train(X_train, y_train, X_val, y_val,
                  num_iters=iters, batch_size=200,
                  learning_rate=best_lr, learning_rate_decay=0.95,
                  reg=0.25, verbose=False)

# Predict on the validation set
val_acc = (net.predict(X_val) == y_val).mean()
print('Learning rate: {}, Validation accuracy: {}'.format(best_lr,val_acc))

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
best_net = net
```

```
Validation accuracy: [0.168] for learning rate: 1e-05
Validation accuracy: [0.362] for learning rate: 5e-05
Validation accuracy: [0.436] for learning rate: 0.0001
Validation accuracy: [0.497] for learning rate: 0.0005
Validation accuracy: [0.516] for learning rate: 0.001
Best learning rate: 0.001
Learning rate: 0.001, Validation accuracy: 0.505
```

In [17]:

```
from cs231n.vis_utils import visualize_grid

# Visualize the weights of the network

def show_net_weights(net):
    W1 = net.params['W1']
    W1 = W1.T.reshape(32, 32, 3, -1).transpose(3, 0, 1, 2)
    plt.imshow(visualize_grid(W1, padding=3).astype('uint8'))
    plt.gca().axis('off')
    plt.show()

show_net_weights(subopt_net)
show_net_weights(best_net)
```



Question:

(1) What differences do you see in the weights between the suboptimal net and the best net you arrived at?

Answer:

(1) Each weight of the both the nets captures different features of the image. The weights of best net are more distinct with wide variations as compared to suboptimal net, thus capturing more features and hence a higher accuracy

Evaluate on test set

In [15]:

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

```
Test accuracy:  0.502
```

```python
import numpy as np
import pdb

"""
This code was originally written for CS 231n at Stanford University
(cs231n.stanford.edu).  It has been modified in various areas for use in the
ECE 239AS class at UCLA.  This includes the descriptions of what code to
implement as well as some slight potential changes in variable names to be
consistent with class nomenclature.  We thank Justin Johnson & Serena Yeung for
permission to use this code.  To see the original version, please visit
cs231n.stanford.edu.
"""


def affine_forward(x, w, b):
    """
  Computes the forward pass for an affine (fully-connected) layer.

  The input x has shape (N, d_1, ..., d_k) and contains a minibatch of N
  examples, where each example x[i] has shape (d_1, ..., d_k). We will
  reshape each input into a vector of dimension D = d_1 * ... * d_k, and
  then transform it to an output vector of dimension M.

  Inputs:
  - x: A numpy array containing input data, of shape (N, d_1, ..., d_k)
  - w: A numpy array of weights, of shape (D, M)
  - b: A numpy array of biases, of shape (M,)

  Returns a tuple of:
  - out: output, of shape (N, M)
  - cache: (x, w, b)
  """

  # ================================================================ #
  # YOUR CODE HERE:
  #   Calculate the output of the forward pass.  Notice the dimensions
  #   of w are D x M, which is the transpose of what we did in earlier
  #   assignments.
  # ================================================================ #

    N = x.shape[0]
    x_flat = x.reshape((N,-1))
    out = x_flat.dot(w) + b


  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #

    cache = (x, w, b)
    return out, cache


def affine_backward(dout, cache):
    """
  Computes the backward pass for an affine layer.

  Inputs:
  - dout: Upstream derivative, of shape (N, M)
  - cache: Tuple of:
```

```python
      - x: Input data, of shape (N, d_1, ... d_k)
      - w: Weights, of shape (D, M)

    Returns a tuple of:
    - dx: Gradient with respect to x, of shape (N, d1, ..., d_k)
    - dw: Gradient with respect to w, of shape (D, M)
    - db: Gradient with respect to b, of shape (M,)
    """
      x, w, b = cache
      #dx, dw, db = None, None, None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the gradients for the backward pass.
    # ================================================================ #

      N = x.shape[0]
      x_flat = x.reshape((N,-1))
      dw = (x_flat.T).dot(dout)
      dx = dout.dot(w.T)
      dx = dx.reshape(x.shape)
      db = np.sum(dout,axis=0)

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

      return dx, dw, db

  def relu_forward(x):
    """
    Computes the forward pass for a layer of rectified linear units (ReLUs).

    Input:
    - x: Inputs, of any shape

    Returns a tuple of:
    - out: Output, of the same shape as x
    - cache: x
    """
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the ReLU forward pass.
    # ================================================================ #

    out = np.maximum(0,x)
    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    cache = x
    return out, cache


  def relu_backward(dout, cache):
      """
    Computes the backward pass for a layer of rectified linear units (ReLUs).

    Input:
    - dout: Upstream derivatives, of any shape
    - cache: Input x, of same shape as dout
```

```python
  Returns:
  - dx: Gradient with respect to x
  """
    x = cache

  # =============================================================== #
  # YOUR CODE HERE:
  #   Implement the ReLU backward pass
  # =============================================================== #

    Ind = np.sign(np.maximum(0,x))
    dx = np.multiply(Ind,dout)

  # =============================================================== #
  # END YOUR CODE HERE
  # =============================================================== #
    return dx

def svm_loss(x, y):
  """
  Computes the loss and gradient using for multiclass SVM classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

  Returns a tuple of:
  - loss: Scalar giving the loss
  - dx: Gradient of the loss with respect to x
  """
  N = x.shape[0]
  correct_class_scores = x[np.arange(N), y]
  margins = np.maximum(0, x - correct_class_scores[:, np.newaxis] + 1.0)
  margins[np.arange(N), y] = 0
  loss = np.sum(margins) / N
  num_pos = np.sum(margins > 0, axis=1)
  dx = np.zeros_like(x)
  dx[margins > 0] = 1
  dx[np.arange(N), y] -= num_pos
  dx /= N
  return loss, dx


def softmax_loss(x, y):
  """
  Computes the loss and gradient for softmax classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

  Returns a tuple of:
  - loss: Scalar giving the loss
  - dx: Gradient of the loss with respect to x
  """
```

```
probs = np.exp(x - np.max(x, axis=1, keepdims=True))
probs /= np.sum(probs, axis=1, keepdims=True)
N = x.shape[0]
loss = -np.sum(np.log(probs[np.arange(N), y])) / N
dx = probs.copy()
dx[np.arange(N), y] -= 1
dx /= N
return loss, dx
```

```python
import numpy as np

from .layers import *
from .layer_utils import *

"""
This code was originally written for CS 231n at Stanford University
(cs231n.stanford.edu).  It has been modified in various areas for use in the
ECE 239AS class at UCLA.  This includes the descriptions of what code to
implement as well as some slight potential changes in variable names to be
consistent with class nomenclature.  We thank Justin Johnson & Serena Yeung for
permission to use this code.  To see the original version, please visit
cs231n.stanford.edu.
"""

class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network with ReLU nonlinearity and
  softmax loss that uses a modular layer design. We assume an input dimension
  of D, a hidden dimension of H, and perform classification over C classes.

  The architecure should be affine - relu - affine - softmax.

  Note that this class does not implement gradient descent; instead, it
  will interact with a separate Solver object that is responsible for running
  optimization.

  The learnable parameters of the model are stored in the dictionary
  self.params that maps parameter names to numpy arrays.
  """

  def __init__(self, input_dim=3*32*32, hidden_dims=100, num_classes=10,
               dropout=0, weight_scale=1e-3, reg=0.0):
    """
    Initialize a new network.

    Inputs:
    - input_dim: An integer giving the size of the input
    - hidden_dims: An integer giving the size of the hidden layer
    - num_classes: An integer giving the number of classes to classify
    - dropout: Scalar between 0 and 1 giving dropout strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - reg: Scalar giving L2 regularization strength.
    """
    self.params = {}
    self.reg = reg

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize W1, W2, b1, and b2.  Store these as self.params['W1'],
    #   self.params['W2'], self.params['b1'] and self.params['b2']. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    #   The dimensions of W1 should be (input_dim, hidden_dim) and the
    #   dimensions of W2 should be (hidden_dims, num_classes)
    # ================================================================ #

    self.params = {}
    self.params['W1'] = weight_scale * np.random.randn(input_dim, hidden_dims)
    self.params['b1'] = np.zeros(hidden_dims)
```

```python
        self.params['W2'] = weight_scale * np.random.randn(hidden_dims, num_classes)
        self.params['b2'] = np.zeros(num_classes)

        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

    def loss(self, X, y=None):
        """
        Compute loss and gradient for a minibatch of data.

        Inputs:
        - X: Array of input data of shape (N, d_1, ..., d_k)
        - y: Array of labels, of shape (N,). y[i] gives the label for X[i].

        Returns:
        If y is None, then run a test-time forward pass of the model and return:
        - scores: Array of shape (N, C) giving classification scores, where
          scores[i, c] is the classification score for X[i] and class c.

        If y is not None, then run a training-time forward and backward pass and
        return a tuple of:
        - loss: Scalar value giving the loss
        - grads: Dictionary with the same keys as self.params, mapping parameter
          names to gradients of the loss with respect to those parameters.
        """
        W1, b1 = self.params['W1'], self.params['b1']
        W2, b2 = self.params['W2'], self.params['b2']
        N = X.shape[0]
        #scores = None

        # ================================================================ #
        # YOUR CODE HERE:
        #   Implement the forward pass of the two-layer neural network. Store
        #   the class scores as the variable 'scores'.  Be sure to use the layers
        #   you prior implemented.
        # ================================================================ #
        a1,_ = affine_forward(X,W1,b1)
        h1,_ = relu_forward(a1)
        scores,_ = affine_forward(h1,W2,b2)

        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

        # If y is None then we are in test mode so just return scores
        if y is None:
            return scores

        loss, grads = 0, {}
        # ================================================================ #
        # YOUR CODE HERE:
        #   Implement the backward pass of the two-layer neural net.  Store
        #   the loss as the variable 'loss' and store the gradients in the
        #   'grads' dictionary.  For the grads dictionary, grads['W1'] holds
        #   the gradient for W1, grads['b1'] holds the gradient for b1, etc.
        #   i.e., grads[k] holds the gradient for self.params[k].
        #
        #   Add L2 regularization, where there is an added cost 0.5*self.reg*W^2
        #   for each W.  Be sure to include the 0.5 multiplying factor to
        #   match our implementation.
```

```python
        #
        #   And be sure to use the layers you prior implemented.
        # ================================================================ #

        loss, dsc = softmax_loss(scores,y)
        loss += 0.5*self.reg*(np.sum(np.square(W1)) + np.sum(np.square(W2)))
        dh1, dw2, db2 = affine_backward(dsc,(h1,W2,b2))
        da1 = relu_backward(dh1,a1)
        dx, dw1, db1 = affine_backward(da1,(X,W1,b1))
        grads['W1']=dw1 + self.reg*W1
        grads['b1'] = db1
        grads['W2'] = dw2 + self.reg*W2
        grads['b2'] = db2


        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

        return loss, grads


class FullyConnectedNet(object):
    """
    A fully-connected neural network with an arbitrary number of hidden layers,
    ReLU nonlinearities, and a softmax loss function. This will also implement
    dropout and batch normalization as options. For a network with L layers,
    the architecture will be

    {affine - [batch norm] - relu - [dropout]} x (L - 1) - affine - softmax

    where batch normalization and dropout are optional, and the {...} block is
    repeated L - 1 times.

    Similar to the TwoLayerNet above, learnable parameters are stored in the
    self.params dictionary and will be learned using the Solver class.
    """

    def __init__(self, hidden_dims, input_dim=3*32*32, num_classes=10,
                 dropout=0, use_batchnorm=False, reg=0.0,
                 weight_scale=1e-2, dtype=np.float32, seed=None):
        """
        Initialize a new FullyConnectedNet.

        Inputs:
        - hidden_dims: A list of integers giving the size of each hidden layer.
        - input_dim: An integer giving the size of the input.
        - num_classes: An integer giving the number of classes to classify.
        - dropout: Scalar between 0 and 1 giving dropout strength. If dropout=0 then
          the network should not use dropout at all.
        - use_batchnorm: Whether or not the network should use batch normalization.
        - reg: Scalar giving L2 regularization strength.
        - weight_scale: Scalar giving the standard deviation for random
          initialization of the weights.
        - dtype: A numpy datatype object; all computations will be performed using
          this datatype. float32 is faster but less accurate, so you should use
          float64 for numeric gradient checking.
        - seed: If not None, then pass this random seed to the dropout layers. This
          will make the dropout layers deteriminstic so we can gradient check the
          model.
        """
```

```
    self.use_batchnorm = use_batchnorm
    self.use_dropout = dropout > 0
    self.reg = reg
    self.num_layers = 1 + len(hidden_dims)
    self.dtype = dtype
    self.params = {}

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize all parameters of the network in the self.params dictionary.
    #   The weights and biases of layer 1 are W1 and b1; and in general the
    #   weights and biases of layer i are Wi and bi. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    # ================================================================ #
    self.ly = len(hidden_dims) + 1
    layers = hidden_dims
    layers.insert(self.ly,num_classes)
    layers.insert(0,input_dim)

    for i in range(self.ly):
        j = i+1
        wi = 'W'+ str(j)
        bi = 'b'+ str(j)
        self.params[wi] = weight_scale * np.random.randn(layers[i], layers[i+1])
        self.params[bi] = np.zeros(layers[i+1])

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # When using dropout we need to pass a dropout_param dictionary to each
    # dropout layer so that the layer knows the dropout probability and the mode
    # (train / test). You can pass the same dropout_param to each dropout layer.
    self.dropout_param = {}
    if self.use_dropout:
      self.dropout_param = {'mode': 'train', 'p': dropout}
      if seed is not None:
        self.dropout_param['seed'] = seed

    # With batch normalization we need to keep track of running means and
    # variances, so we need to pass a special bn_param object to each batch
    # normalization layer. You should pass self.bn_params[0] to the forward pass
    # of the first batch normalization layer, self.bn_params[1] to the forward
    # pass of the second batch normalization layer, etc.
    self.bn_params = []
    if self.use_batchnorm:
      self.bn_params = [{'mode': 'train'} for i in np.arange(self.num_layers - 1)]

    # Cast all parameters to the correct datatype
    for k, v in self.params.items():
      self.params[k] = v.astype(dtype)


  def loss(self, X, y=None):
    """
    Compute loss and gradient for the fully-connected net.

    Input / output: Same as TwoLayerNet above.
    """
    X = X.astype(self.dtype)
```

```python
mode = 'test' if y is None else 'train'

# Set train/test mode for batchnorm params and dropout param since they
# behave differently during training and testing.
if self.dropout_param is not None:
  self.dropout_param['mode'] = mode
if self.use_batchnorm:
  for bn_param in self.bn_params:
    bn_param[mode] = mode

scores = None
# ================================================================ #
# YOUR CODE HERE:
#   Implement the forward pass of the FC net and store the output
#   scores as the variable "scores".
# ================================================================ #
infer = {}
inp = X
for i in range(self.ly):
    wi = self.params['W'+str(i+1)]
    bi = self.params['b'+str(i+1)]
    infer['a'+str(i+1)],_ = affine_forward(inp,wi,bi)
    if i==(self.ly-1):
        scores = infer['a'+str(i+1)]
    else:
        infer['h'+str(i+1)],_ = relu_forward(infer['a'+str(i+1)])
        inp = infer['h'+str(i+1)]


# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #

# If test mode return early
if mode == 'test':
  return scores

loss, grads = 0.0, {}
# ================================================================ #
# YOUR CODE HERE:
#   Implement the backwards pass of the FC net and store the gradients
#   in the grads dict, so that grads[k] is the gradient of self.params[k]
#   Be sure your L2 regularization includes a 0.5 factor.
# ================================================================ #

idx = np.arange(self.ly)
idx = np.flip(idx)
back = {}
loss, back['dsmx'] = softmax_loss(scores,y)
up_grad = back['dsmx']
for i in idx:
    wi = self.params['W'+str(i+1)]
    bi = self.params['b'+str(i+1)]
    loss += 0.5*self.reg*(np.sum(np.square(wi)))

    if i==0:
        back['dx'], back['dw1'], back['db1'] = affine_backward(up_grad,(X,wi,bi))
    else:
        back['dh'+str(i)], back['dw'+str(i+1)], back['db'+str(i+1)]
        = affine_backward(up_grad,(infer['h'+str(i)],wi,bi))
        back['da'+str(i)] = relu_backward(back['dh'+str(i)],infer['a'+str(i)])
```

```
            up_grad = back['da'+str(i)]
        grads['W'+str(i+1)]= back['dw'+str(i+1)] + self.reg*wi
        grads['b'+str(i+1)] = back['db'+str(i+1)]

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #
    return loss, grads
```

Fully connected networks

In the previous notebook, you implemented a simple two-layer neural network class. However, this class is not modular. If you wanted to change the number of layers, you would need to write a new loss and gradient function. If you wanted to optimize the network with different optimizers, you'd need to write new training functions. If you wanted to incorporate regularizations, you'd have to modify the loss and gradient function.

Instead of having to modify functions each time, for the rest of the class, we'll work in a more modular framework where we define forward and backward layers that calculate losses and gradients respectively. Since the forward and backward layers share intermediate values that are useful for calculating both the loss and the gradient, we'll also have these function return "caches" which store useful intermediate values.

The goal is that through this modular design, we can build different sized neural networks for various applications.

In this HW #3, we'll define the basic architecture, and in HW #4, we'll build on this framework to implement different optimizers and regularizations (like BatchNorm and Dropout).

CS231n has built a solid API for building these modular frameworks and training them, and we will use their very well implemented framework as opposed to "reinventing the wheel." This includes using their Solver, various utility functions, and their layer structure. This also includes nndl.fc_net, nndl.layers, and nndl.layer_utils. As in prior assignments, we thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class (cs231n.stanford.edu).

Modular layers

This notebook will build modular layers in the following manner. First, there will be a forward pass for a given layer with inputs (`x`) and return the output of that layer (`out`) as well as cached variables (`cache`) that will be used to calculate the gradient in the backward pass.

```
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 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
```

In [1]:

```
## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.fc_net import *
from cs231n.data_utils import get_CIFAR10_data
from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradient_arra
y
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))))
```

In [2]:

```
# Load the (preprocessed) CIFAR10 data.

data = get_CIFAR10_data()
for k in data.keys():
  print('{}: {} '.format(k, data[k].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,)
```

Linear layers

In this section, we'll implement the forward and backward pass for the linear layers.

The linear layer forward pass is the function `affine_forward` in `nndl/layers.py` and the backward pass is `affine_backward`.

After you have implemented these, test your implementation by running the cell below.

Affine layer forward pass

Implement `affine_forward` and then test your code by running the following cell.

In [3]:

```
# Test the affine_forward function

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

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

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

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

# Compare your output with ours. The error should be around 1e-9.
print('Testing affine_forward function:')
print('difference: {}'.format(rel_error(out, correct_out)))
```

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

Affine layer backward pass

Implement `affine_backward` and then test your code by running the following cell.

In [4]:

```
# Test the affine_backward function

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 1e-10
print('Testing affine_backward function:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
print('dw error: {}'.format(rel_error(dw_num, dw)))
print('db error: {}'.format(rel_error(db_num, db)))
```

```
Testing affine_backward function:
dx error: 1.1235598675182531e-10
dw error: 1.3661153421230798e-10
db error: 1.4156286109039033e-11
```

Activation layers

In this section you'll implement the ReLU activation.

ReLU forward pass

Implement the `relu_forward` function in `nndl/layers.py` and then test your code by running the following cell.

In [5]:

```
# Test the relu_forward function

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

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

# Compare your output with ours. The error should be around 1e-8
print('Testing relu_forward function:')
print('difference: {}'.format(rel_error(out, correct_out)))
```

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

ReLU backward pass

Implement the `relu_backward` function in `nndl/layers.py` and then test your code by running the following cell.

In [6]:

```
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 around 1e-12
print('Testing relu_backward function:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
```

```
Testing relu_backward function:
dx error: 3.2755994729317884e-12
```

Combining the affine and ReLU layers

Often times, an affine layer will be followed by a ReLU layer. So let's make one that puts them together. Layers that are combined are stored in `nndl/layer_utils.py`.

Affine-ReLU layers

We've implemented `affine_relu_forward()` and `affine_relu_backward` in `nndl/layer_utils.py`. Take a look at them to make sure you understand what's going on. Then run the following cell to ensure its implemented correctly.

In [7]:

```
from nndl.layer_utils import affine_relu_forward, affine_relu_backward

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

print('Testing affine_relu_forward and affine_relu_backward:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
print('dw error: {}'.format(rel_error(dw_num, dw)))
print('db error: {}'.format(rel_error(db_num, db)))
```

```
Testing affine_relu_forward and affine_relu_backward:
dx error: 1.9507299665612664e-10
dw error: 3.95303746788456e-10
db error: 3.275601996442511e-12
```

Softmax and SVM losses

You've already implemented these, so we have written these in `layers.py`. The following code will ensure they are working correctly.

In [8]:

```
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)

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

# Test svm_loss function. Loss should be around 9 and dx error should be 1e-9
print('Testing svm_loss:')
print('loss: {}'.format(loss))
print('dx error: {}'.format(rel_error(dx_num, dx)))

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

# Test softmax_loss function. Loss should be 2.3 and dx error should be 1e-8
print('\nTesting softmax_loss:')
print('loss: {}'.format(loss))
print('dx error: {}'.format(rel_error(dx_num, dx)))
```

```
Testing svm_loss:
loss: 9.000036384628125
dx error: 8.182894472887002e-10

Testing softmax_loss:
loss: 2.302589184644751
dx error: 8.625121336595623e-09
```

Implementation of a two-layer NN

In `nndl/fc_net.py`, implement the class `TwoLayerNet` which uses the layers you made here. When you have finished, the following cell will test your implementation.

In [9]:

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

std = 1e-2
model = TwoLayerNet(input_dim=D, hidden_dims=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'

for reg in [0.0, 0.7]:
  print('Running numeric gradient check with reg = {}'.format(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('{} relative error: {}'.format(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: 1.8336562786695002e-08
W2 relative error: 3.201560569143183e-10
b1 relative error: 9.828315204644842e-09
b2 relative error: 4.329134954569865e-10
Running numeric gradient check with reg = 0.7
W1 relative error: 2.5279152310200606e-07
W2 relative error: 2.8508510893102143e-08
b1 relative error: 1.564679947504764e-08
b2 relative error: 9.089617896905665e-10
```

Solver

We will now use the cs231n Solver class to train these networks. Familiarize yourself with the API in `cs231n/solver.py`. After you have done so, declare an instance of a TwoLayerNet with 200 units and then train it with the Solver. Choose parameters so that your validation accuracy is at least 40%.

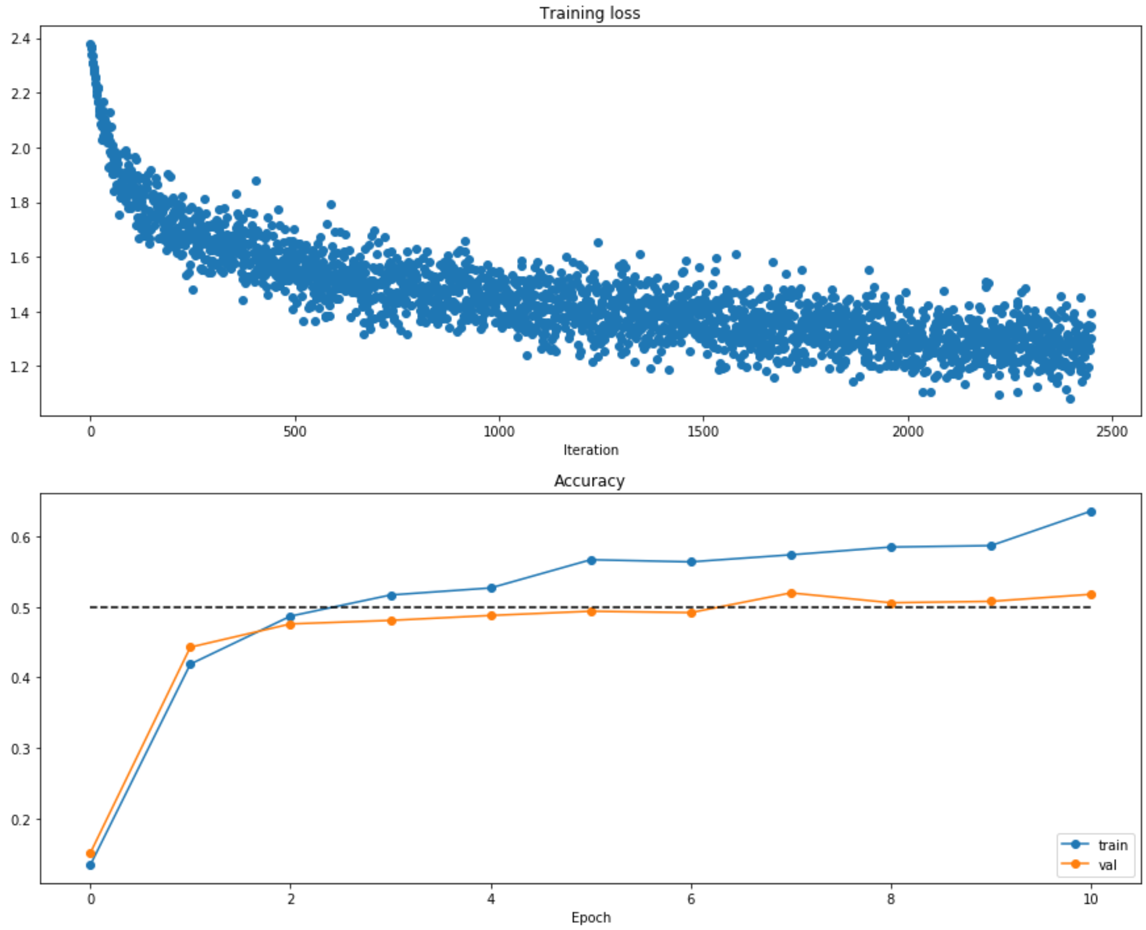
In [10]:

```
#model = TwoLayerNet()
#solver = None

# ================================================================ #
# YOUR CODE HERE:
#   Declare an instance of a TwoLayerNet and then train
#   it with the Solver. Choose hyperparameters so that your validation
#   accuracy is at least 40%.  We won't have you optimize this further
#   since you did it in the previous notebook.
#
# ================================================================ #
data = get_CIFAR10_data()
X_train  = data['X_train']
N = X_train.shape[0]
D = 3*32*32
H = 200
C = 10
reg = 0.25

std = 1e-3
model = TwoLayerNet(input_dim=D, hidden_dims=H, num_classes=C, weight_scale=std, reg=re
g)

solver = Solver(model, data,
                update_rule='sgd',
                optim_config={
                  'learning_rate': 1e-3,
                },
                lr_decay=0.95,
                num_epochs=10, batch_size=200,
                print_every=100)
solver.train()

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```

```
(Iteration 1 / 2450) loss: 2.379872
(Epoch 0 / 10) train acc: 0.134000; val_acc: 0.151000
(Iteration 101 / 2450) loss: 1.778104
(Iteration 201 / 2450) loss: 1.729202
(Epoch 1 / 10) train acc: 0.419000; val_acc: 0.443000
(Iteration 301 / 2450) loss: 1.712557
(Iteration 401 / 2450) loss: 1.541243
(Epoch 2 / 10) train acc: 0.487000; val_acc: 0.476000
(Iteration 501 / 2450) loss: 1.561295
(Iteration 601 / 2450) loss: 1.494405
(Iteration 701 / 2450) loss: 1.502597
(Epoch 3 / 10) train acc: 0.517000; val_acc: 0.481000
(Iteration 801 / 2450) loss: 1.580222
(Iteration 901 / 2450) loss: 1.521317
(Epoch 4 / 10) train acc: 0.527000; val_acc: 0.488000
(Iteration 1001 / 2450) loss: 1.450027
(Iteration 1101 / 2450) loss: 1.360991
(Iteration 1201 / 2450) loss: 1.372445
(Epoch 5 / 10) train acc: 0.567000; val_acc: 0.494000
(Iteration 1301 / 2450) loss: 1.299600
(Iteration 1401 / 2450) loss: 1.383735
(Epoch 6 / 10) train acc: 0.564000; val_acc: 0.492000
(Iteration 1501 / 2450) loss: 1.400473
(Iteration 1601 / 2450) loss: 1.435647
(Iteration 1701 / 2450) loss: 1.267419
(Epoch 7 / 10) train acc: 0.574000; val_acc: 0.520000
(Iteration 1801 / 2450) loss: 1.312978
(Iteration 1901 / 2450) loss: 1.250044
(Epoch 8 / 10) train acc: 0.585000; val_acc: 0.506000
(Iteration 2001 / 2450) loss: 1.284060
(Iteration 2101 / 2450) loss: 1.364746
(Iteration 2201 / 2450) loss: 1.193909
(Epoch 9 / 10) train acc: 0.587000; val_acc: 0.508000
(Iteration 2301 / 2450) loss: 1.261326
(Iteration 2401 / 2450) loss: 1.202334
(Epoch 10 / 10) train acc: 0.636000; val_acc: 0.518000
```

In [11]:

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

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

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



Multilayer Neural Network

Now, we implement a multi-layer neural network.

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

Implement the initialization, the forward pass, and the backward pass. There will be lines for batchnorm and dropout layers and caches; ignore these all for now. That'll be in assignment #4.

In [12]:

```
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 = {}'.format(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: {}'.format(loss))

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False, h=1e-5)
    print('{} relative error: {}'.format(name, rel_error(grad_num, grads[name])))
```

```
Running check with reg = 0
Initial loss: 2.298960306609043
W1 relative error: 2.383398886246872e-05
W2 relative error: 5.370193171152765e-07
W3 relative error: 1.2800140015272063e-07
b1 relative error: 1.1061819897808475e-08
b2 relative error: 2.198098230478872e-09
b3 relative error: 1.022969417433634e-10
Running check with reg = 3.14
Initial loss: 7.026526536127609
W1 relative error: 4.1734673357718115e-09
W2 relative error: 9.379783094326438e-07
W3 relative error: 4.522284318549473e-09
b1 relative error: 1.742118971765098e-07
b2 relative error: 8.151767684967441e-09
b3 relative error: 1.895286861349364e-10
```

In [13]:

```
# Use the three layer neural network to overfit a small dataset.

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'],
}


#### !!!!!!
# Play around with the weight_scale and learning_rate so that you can overfit a small d
ataset.
# Your training accuracy should be 1.0 to receive full credit on this part.
weight_scale = 5e-2
learning_rate = 1e-3

model = FullyConnectedNet([100, 100],
              weight_scale=weight_scale, dtype=np.float64)
solver = Solver(model, small_data,
                print_every=10, num_epochs=20, batch_size=25,
                update_rule='sgd',
                optim_config={
                  'learning_rate': learning_rate,
                }
         )
solver.train()

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

```
(Iteration 1 / 40) loss: 30.791211
(Epoch 0 / 20) train acc: 0.220000; val_acc: 0.142000
(Epoch 1 / 20) train acc: 0.200000; val_acc: 0.126000
(Epoch 2 / 20) train acc: 0.540000; val_acc: 0.146000
(Epoch 3 / 20) train acc: 0.660000; val_acc: 0.157000
(Epoch 4 / 20) train acc: 0.720000; val_acc: 0.168000
(Epoch 5 / 20) train acc: 0.800000; val_acc: 0.153000
(Iteration 11 / 40) loss: 1.773943
(Epoch 6 / 20) train acc: 0.880000; val_acc: 0.180000
(Epoch 7 / 20) train acc: 0.880000; val_acc: 0.159000
(Epoch 8 / 20) train acc: 0.980000; val_acc: 0.169000
(Epoch 9 / 20) train acc: 1.000000; val_acc: 0.169000
(Epoch 10 / 20) train acc: 1.000000; val_acc: 0.169000
(Iteration 21 / 40) loss: 0.000461
(Epoch 11 / 20) train acc: 1.000000; val_acc: 0.169000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.169000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.169000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.168000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.166000
(Iteration 31 / 40) loss: 0.000224
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.166000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.166000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.166000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.166000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.166000
```