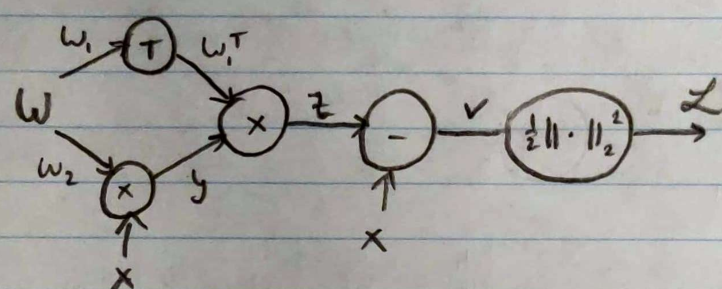
ECE 147 HW 3

Andrea Borette

1) a. W^T is meant to be a kind of inverse operation to W. While W reduces x to a lower number of dimensions, W^T takes the m-dimensional "compressed" data and tries to restore it as well as possible. Therefore, we want the result of $W^T W x$ to recreate the original data x as well as possible, so we to try to minimize the norm of the difference multiplied by $\frac{1}{2}$ as a regularization coefficient.

b.



c. All of the derivatives for $\mathcal{L}$ which get backpropagated to W need to be summed, per the law of total derivatives, to get to the real $\frac{\partial \mathcal{L}}{\partial W}$.

d. Let us use backpropagation.

Equations: $\mathcal{L} = \frac{1}{2}\|v\|_2^2$ $\qquad v = z - x$

$z = W_1^T y$ $\qquad y = W_2 x$

Derivatives: $\frac{\partial \mathcal{L}}{\partial v} = \frac{1}{2}\frac{\partial}{\partial v}(v^T v) = v$ $\qquad \frac{\partial v}{\partial z} = I$

$\frac{\partial z}{\partial y} = W_1$ $\qquad \frac{\partial z}{\partial W_1^T} \approx y^T$ per tensor trick from discussion

$\frac{\partial y}{\partial W_2} \approx x^T$

Backpropagation: $\frac{\partial \mathcal{L}}{\partial z} = \frac{\partial v}{\partial z}\frac{\partial \mathcal{L}}{\partial v} = Iv = v$

$\frac{\partial \mathcal{L}}{\partial W_1^T} = \frac{\partial z}{\partial W_1^T} \leftrightarrow \frac{\partial \mathcal{L}}{\partial z} = v y^T$ per discussion trick

$\frac{\partial \mathcal{L}}{\partial W_1} = \left(\frac{\partial \mathcal{L}}{\partial W_1^T}\right)^T = y v^T$

$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial z}{\partial y}\frac{\partial \mathcal{L}}{\partial z} = W_1 v = W v$

$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial y}{\partial W_2} \leftrightarrow \frac{\partial \mathcal{L}}{\partial y} = W v x^T$

Then $\nabla_W \mathcal{L} = \frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial W_1} + \frac{\partial \mathcal{L}}{\partial W_2} = y v^T + W v x^T = Wx(z-x)^T + W(z-x)$

$\nabla_W \mathcal{L} = Wx(W^T W x - x)^T + W(W^T W x - x)x^T$

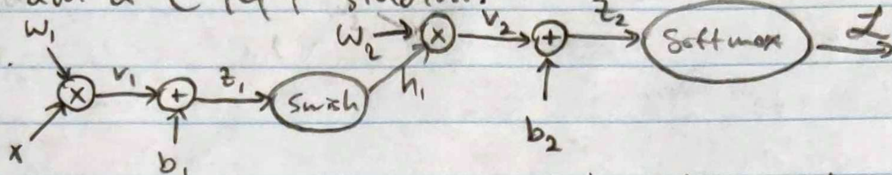
$$\nabla_W \mathcal{L} = Wx((W^TW - I)x)^T + W(W^TW - I)xx^T$$
$$= W(xx^T(W^TW - I) + (W^TW - I)xx^T)$$

Since $(xx^T(W^TW - I))^T = (W^TW - I)xx^T$,

$$\nabla_W \mathcal{L} = 2W(W^TW - I)xx^T$$

2) I am a C147 student.

3) a.



(C) $z_2 = v_2 + b_2$

(C) $v_2 = W_2 h_1$

(H) $h_1 = \text{Swish}(z_1)$

(H) $z_1 = v_1 + b_1$

(H) $v_1 = W_1 x$

(C) $\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial z_2}{\partial b_2}\frac{\partial \mathcal{L}}{\partial z_2} = \frac{\partial \mathcal{L}}{\partial z_2}$

(C) $\frac{\partial \mathcal{L}}{\partial v_2} = \frac{\partial z_2}{\partial v_2}\frac{\partial \mathcal{L}}{\partial z_2} = I\frac{\partial \mathcal{L}}{\partial z_2} = \frac{\partial \mathcal{L}}{\partial z_2}$

(H) $\frac{\partial \mathcal{L}}{\partial h_1} = \frac{\partial v_2}{\partial h_1}\frac{\partial \mathcal{L}}{\partial v_2} = W_2^T\frac{\partial \mathcal{L}}{\partial z_2}$

(C×H) $\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial v_2}{\partial W_2}\frac{\partial \mathcal{L}}{\partial v_2} = \underset{(C\times 1)}{\frac{\partial \mathcal{L}}{\partial z_2}}\,\underset{(1\times H)}{h_1^T}$ ← per the discussion trick

W_1 $H\times D$ $\quad$ W_2 $C\times H$

$$\frac{\partial h_{1i}}{\partial z_j} = \frac{\partial}{\partial z_i}\left(\frac{z_i}{1+e^{-z_i}}\right)\delta_{ij} = \frac{1+e^{-z_i}+z_ie^{-z_i}}{(1+e^{-z_i})^2}\delta_{ij}$$

$$\Rightarrow \frac{\partial h_1}{\partial z_1} = \begin{bmatrix} \frac{1+e^{-z_{11}}+z_{11}e^{-z_{11}}}{(1+e^{-z_{11}})^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1+e^{-z_{1H}}+z_{1H}e^{-z_{1H}}}{(1+e^{-z_{1H}})^2} \end{bmatrix} \quad (*)$$

(H) $\frac{\partial \mathcal{L}}{\partial z_1} = \frac{\partial h_1}{\partial z_1}\frac{\partial \mathcal{L}}{\partial h_1} = \underset{H\times H}{\frac{\partial h_1}{\partial z_1}}\,\underset{H\times C}{W_2^T}\,\underset{C\times 1}{\frac{\partial \mathcal{L}}{\partial z_2}}$

(H) $\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial z_1}{\partial b_1}\frac{\partial \mathcal{L}}{\partial z_1} = I\frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2} = \frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2}$

(H) $\frac{\partial \mathcal{L}}{\partial v_1} = \frac{\partial z_1}{\partial v_1}\frac{\partial \mathcal{L}}{\partial z_1} = I\frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2} = \frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2}$

(H×D) $\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial v_1}{\partial W_1}\frac{\partial \mathcal{L}}{\partial v_1} = \underbrace{\frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2}}_{H\times 1}\underset{1\times D}{x^T}$ ← discussion trick

Then

b) $\nabla_{W_2}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial z_2}h_1^T$

$\nabla_{b_2}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial z_2}$

c) $\nabla_{W_1}\mathcal{L} = \frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2}x^T$

$\nabla_{b_1}\mathcal{L} = \frac{\partial h_1}{\partial z_1}W_2^T\frac{\partial \mathcal{L}}{\partial z_2}$

where $\frac{\partial h_1}{\partial z_1}$ is defined as above in (*)

two_layer_nn

February 10, 2024

0.1 This is the 2-layer neural network notebook for ECE C147/C247 Homework #3

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

Please print out the notebook entirely when completed.

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

[2]:
```
import random
import numpy as np
from utils.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))))
```

0.2 Toy example

Before loading CIFAR-10, there will be a toy example to test your implementation of the forward and backward pass. Make sure to read the description of TwoLayerNet class in neural_net.py file , understand the architecture and initializations

[5]:
```
from nndl.neural_net import TwoLayerNet
```

[6]:
```
# 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()
```

0.2.1 Compute forward pass scores

[30]:

```
## Implement the forward pass of the neural network.
## See the loss() method in TwoLayerNet class for the same

# 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
```

0.2.2 Forward pass loss

[81]:
```python
loss, _ = net.loss(X, y, reg=0.05)
correct_loss = 1.071696123862817

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

```
Loss: 1.071696123862817
Difference between your loss and correct loss:
0.0
```

0.2.3 Backward pass

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

[178]:
```python
from utils.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, grads[param_name])))
#     print(param_grad_num)
```

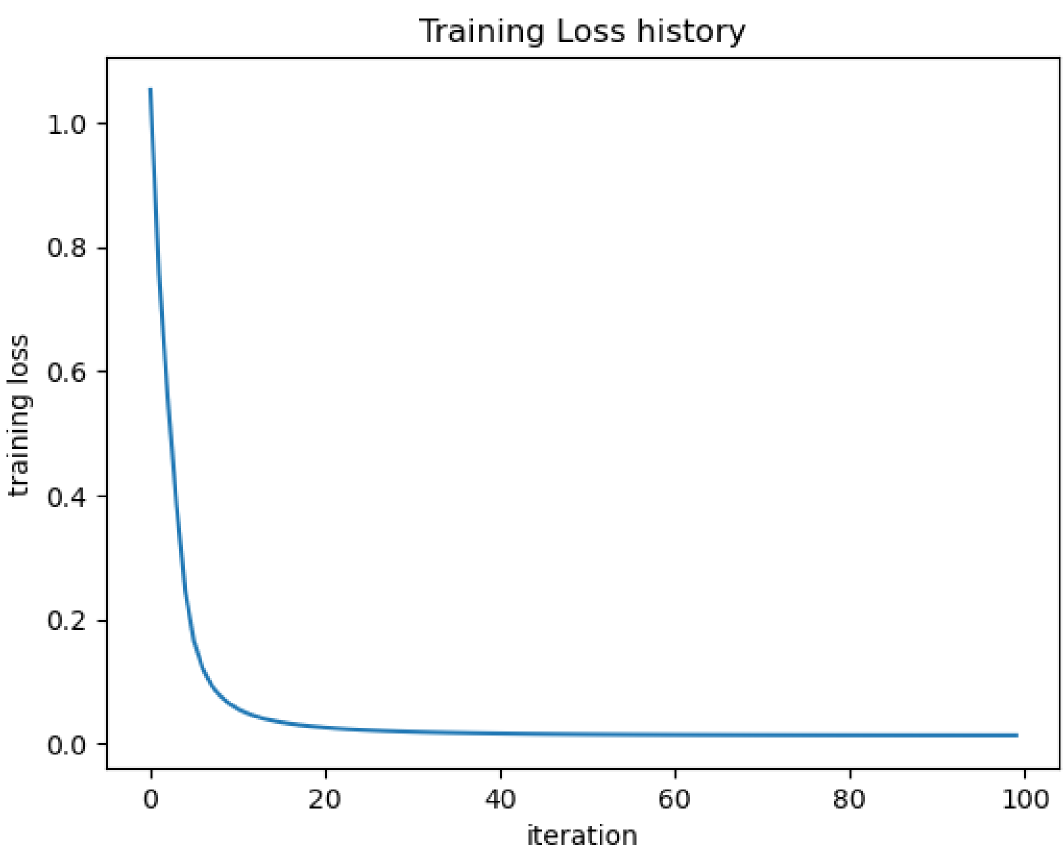
```
b2 max relative error: 1.8392106647421603e-10
W2 max relative error: 3.4254767397498007e-10
b1 max relative error: 3.1726804786908923e-09
W1 max relative error: 1.2832908996874818e-09
```

0.2.4 Training the network

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

[182]:
```
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:  0.013338037215396807
```



0.3 Classify CIFAR-10

Do classification on the CIFAR-10 dataset.

[184]:
```
from utils.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.
    """
    # Load the raw CIFAR-10 data
    cifar10_dir = '/home/andrea/git/UCLA/UCLA_ECE147/cifar-10-batches-py' #␣
↪remember to use correct path
    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,)
```

0.3.1 Running SGD

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

[189]:
```
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 2.3027781109127994
iteration 100 / 1000: loss 2.3023051866249236
iteration 200 / 1000: loss 2.2971832381421646
iteration 300 / 1000: loss 2.2672203812398948
iteration 400 / 1000: loss 2.1834964063436906
iteration 500 / 1000: loss 2.162807507433956
iteration 600 / 1000: loss 2.077024585250963
iteration 700 / 1000: loss 1.9925966641437924
iteration 800 / 1000: loss 1.9871874501072693
iteration 900 / 1000: loss 1.9944860279554801
Validation accuracy:  0.28
```

0.4 Questions:

The training accuracy isn't great.

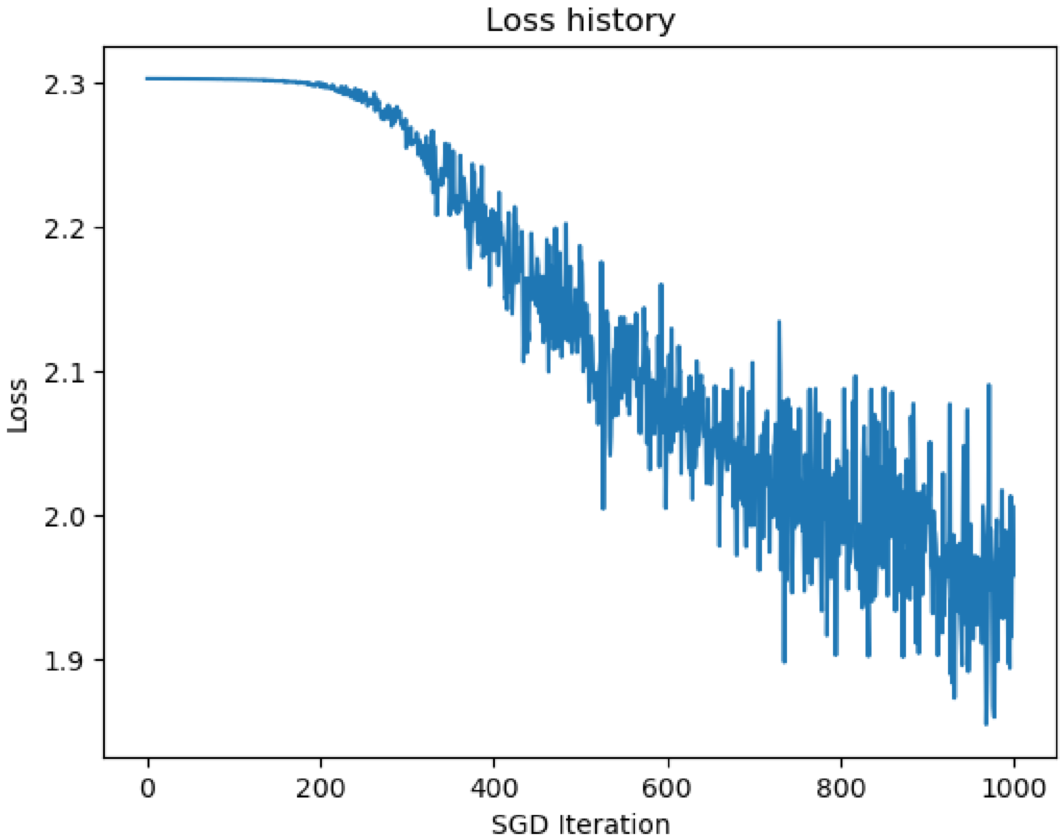
(1) What are some of the reasons why this is the case? Take the following cell to do some analyses and then report your answers in the cell following the one below.

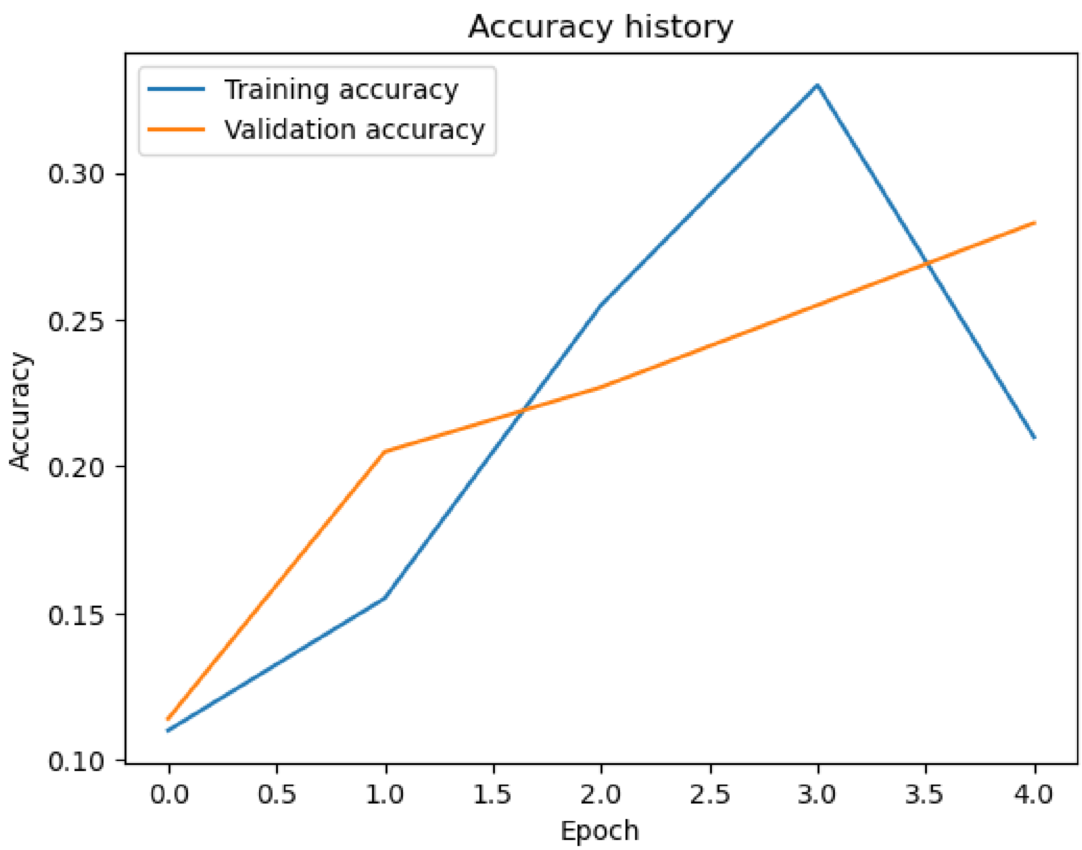
(2) How should you fix the problems you identified in (1)?

[190]:
```
stats['train_acc_history']
```

[190]:
```
[0.11, 0.155, 0.255, 0.33, 0.21]
```

[196]:
```
# ================================================================ #
# 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.title("Loss history")
plt.plot(np.arange(len(stats['loss_history'])), stats['loss_history'])
plt.xlabel("SGD Iteration")
plt.ylabel("Loss")
plt.show()
plt.title("Accuracy history")
plt.plot(np.arange(len(stats['train_acc_history'])),␣
 ↪stats['train_acc_history'], label="Training accuracy")
plt.plot(np.arange(len(stats['val_acc_history'])), stats['val_acc_history'],␣
 ↪label="Validation accuracy")
plt.xlabel("Epoch")
plt.ylabel("Accuracy")
plt.legend()
plt.show()

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


Loss history
2.3
2.2
2.1
2.0
1.9
Loss
0
200
400
600
800
1000
SGD Iteration



0.5 Answers:

(1) Towards the beginning of the loss history, loss stays roughly constant, indicating that the initial learning rate is too low. Interestingly, between epohs 3 and 4, training accuracy decreases as validation accuracy increases, the opposite of overfitting. Further, while loss is noisy, it does decrease somewhat linearly. This would suggest that the model could benefit from more training.

(2) I would change the hyperparameters of the network.

0.6 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.

[233]:
```
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)!
# ================================================================ #

best_pred = 0.0
# net = None

def nn_loss(args):
    global best_net, best_pred
    learning_rate, learning_rate_decay, reg = args

    net = TwoLayerNet(input_size, hidden_size, num_classes)

    # Train the network
    stats = net.train(X_train, y_train, X_val, y_val,
            num_iters=2250, batch_size=500,
            learning_rate=learning_rate,␣
 ↪learning_rate_decay=learning_rate_decay,
            reg=reg, verbose=False)
    val_acc = (net.predict(X_val) == y_val).mean()
    print('Iteration accuracy: ', val_acc)
    if (val_acc > best_pred):
        best_net = net
        best_pred = val_acc
    return -val_acc

result = opt.minimize(nn_loss, [5e-4,0.95,0.2],
                      method='Nelder-Mead',
                      tol=1e-4,
                      options={"maxiter": 10},
                      bounds=[(0,1),(0,1),(0,1)])

# best_net = TwoLayerNet(input_size, hidden_size, num_classes)

# # Train the network
# stats = best_net.train(X_train, y_train, X_val, y_val,
#         num_iters=2250, batch_size=500,
#         learning_rate=result["x"][0], learning_rate_decay=result["x"][1],
#         reg=result["x"][2], verbose=False)

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

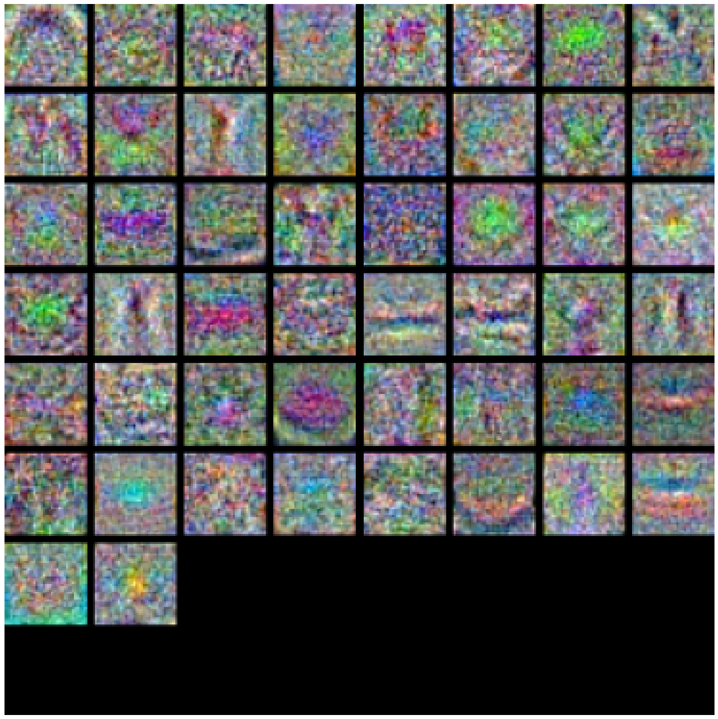
```
# ================================================================ #
val_acc = (best_net.predict(X_val) == y_val).mean()
print('Validation accuracy: ', val_acc)
```

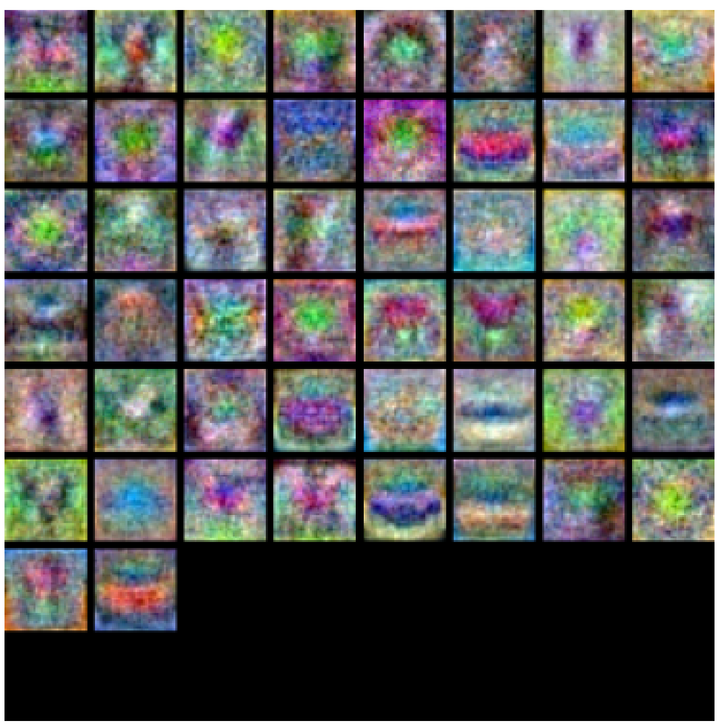
```
Iteration accuracy:  0.473
Iteration accuracy:  0.467
Iteration accuracy:  0.488
Iteration accuracy:  0.459
Iteration accuracy:  0.484
Iteration accuracy:  0.493
Iteration accuracy:  0.487
Iteration accuracy:  0.5
Iteration accuracy:  0.496
Iteration accuracy:  0.495
Iteration accuracy:  0.499
Iteration accuracy:  0.498
Iteration accuracy:  0.496
Iteration accuracy:  0.507
Iteration accuracy:  0.493
Iteration accuracy:  0.499
Iteration accuracy:  0.507
Validation accuracy:  0.507
```

[234]:
```
val_acc = (best_net.predict(X_val) == y_val).mean()
print('Validation accuracy: ', val_acc)
```

```
Validation accuracy:  0.507
```

[235]:
```
from utils.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)
```





0.7 Question:

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

0.8 Answer:

(1) The suboptimal weights, when represented as images, are a lot more blurry, whereas the more optimal ones create clearer and more concise “images”.

0.9 Evaluate on test set

[236]:
```
test_acc = (best_net.predict(X_test) == y_test).mean()
print('Test accuracy: ', test_acc)
```

```
Test accuracy:  0.488
```

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


class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network. The net has an input dimension of
  D, 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
      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.
    # ================================================================ #

    z1 = W1@(X.T) + b1[:,np.newaxis] # (H x D) x (D x N) = (H x N)
    h1 = (z1 >= 0)*z1 # ReLU
    z2 = W2@h1 + b2[:,np.newaxis]  # (C x H) x (H x N) = (C x N)
    scores = z2.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

    loss = np.log(np.exp(z2.T).sum(axis=1)).mean() - z2.T[np.arange(len(z2.T)), y].mean() + \
            reg*0.5*(np.linalg.norm(W1, ord='fro')**2 + np.linalg.norm(W2, ord='fro')**2)

    # ================================================================ #
    # 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.
    # ================================================================ #
```

```
    a = np.zeros_like(z2.T)
    a[np.arange(len(z2.T)), y]=1
    dL_dz2 = (np.exp(z2.T)/np.exp(z2.T).sum(axis=1)[:,np.newaxis] - a).T
    dL_db2 = dL_dz2.sum(axis=1)
    dL_dv2 = dL_dz2
    dL_dW2 = dL_dv2@h1.T
    dL_dh1 = W2.T@dL_dv2
    dL_dz1 = (h1 > 0)*dL_dh1
    dL_db1 = dL_dz1.sum(axis=1)
    dL_dv1 = dL_dz1
    dL_dW1 = dL_dv1@X

    dL_db2 /= N
    dL_dW2 /= N
    dL_db1 /= N
    dL_dW1 /= N

    dL_dW2 += reg * W2
    dL_dW1 += reg * W1

    grads["b2"] = dL_db2
    grads["W2"] = dL_dW2
    grads["b1"] = dL_db1
    grads["W1"] = dL_dW1

    # ================================================================ #
    # 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.
    - 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.
      # ================================================================ #
      batch_size = min(batch_size, num_train)
      idxs = np.random.choice(np.arange(num_train), size=batch_size, replace=False)
      X_batch = X[idxs]
      y_batch = y[idxs]

      # ================================================================ #
      # 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,
      '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

    # ================================================================ #
    # YOUR CODE HERE:
    #   Predict the class given the input data.
    # ================================================================ #
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    N, D = X.shape

    z1 = W1@(X.T) + b1[:,np.newaxis] # (H x D) x (D x N) = (H x N)
    h1 = (z1 >= 0)*z1 # ReLU
    z2 = W2@h1 + b2[:,np.newaxis]  # (C x H) x (H x N) = (C x N)
    y_pred = np.argmax(z2.T, axis=1)

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

    return y_pred
```

FC_nets

February 10, 2024

1 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).

1.1 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.

```python
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
```

[160]:
```
## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.fc_net import *
from utils.data_utils import get_CIFAR10_data
from utils.gradient_check import eval_numerical_gradient,␣
↪eval_numerical_gradient_array
from utils.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))))
```

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

[161]:
```
# Load the (preprocessed) CIFAR10 data.
# you may find an error here, this is may be because you forgot to use correct␣
↪path in get_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,)
```

1.2 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.

1.2.1 Affine layer forward pass

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

[162]:
```
# 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
```

1.2.2 Affine layer backward pass

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

[163]:
```
# 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: 7.458618411912428e-10
dw error: 1.2082959647153856e-10
db error: 1.071187009206661e-11
```

1.3 Activation layers

In this section you'll implement the ReLU activation.

1.3.1 ReLU forward pass

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

[164]:
```
# 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
```

1.3.2 ReLU backward pass

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

[165]:
```
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: 3.2755923187363565e-12
```

1.4 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`.

1.4.1 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.

[166]:
```
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, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_relu_forward(x, w,␣
 ↪b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_relu_forward(x, w,␣
 ↪b)[0], b, dout)

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: 2.867898912179718e-10
dw error: 2.379121993829847e-10
db error: 1.0809161516023713e-11
```

1.5 Softmax loss

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

[167]:
```
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)



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

# Test softmax_loss function. Loss should be 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 softmax_loss:
loss: 2.302651889291118
dx error: 7.901272461786415e-09
```

1.6 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.

[168]:

```
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: 1.5215703686475096e-08
W2 relative error: 3.2068321167375225e-10
b1 relative error: 8.368195737354163e-09
b2 relative error: 4.3291360264321544e-10
Running numeric gradient check with reg = 0.7
W1 relative error: 2.527915175868136e-07
W2 relative error: 7.976652806155026e-08
b1 relative error: 1.5646801536371197e-08
b2 relative error: 7.759095355706557e-10
```

1.7 Solver

We will now use the utils Solver class to train these networks. Familiarize yourself with the API in `utils/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 50%.

[169]:
```
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 50%.  We won't have you optimize this further
#   since you did it in the previous notebook.
#
# ================================================================ #

model = TwoLayerNet(hidden_dims=200)
model.params['W1'] = model.params['W1'].T
model.params['W2'] = model.params['W2'].T
solver = Solver(model, data,
                update_rule='sgd',
                optim_config={
```

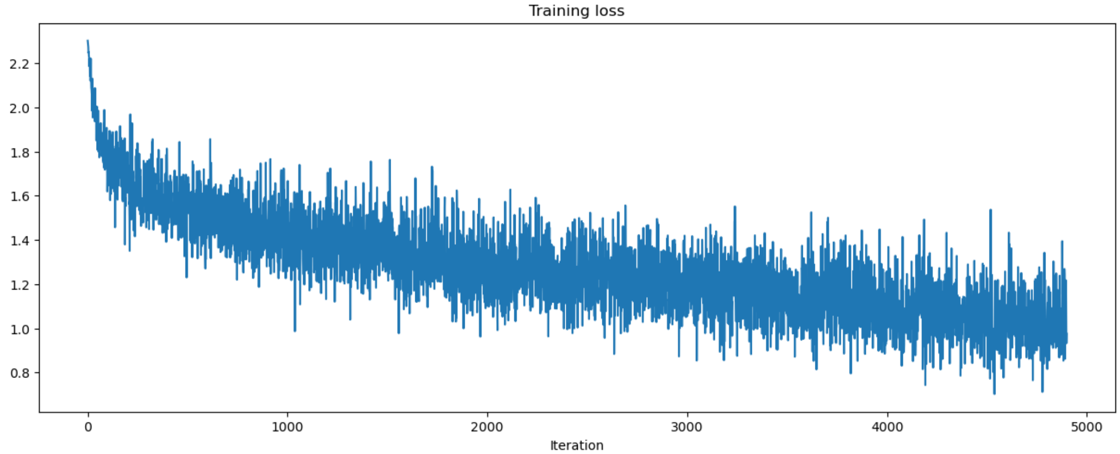
```python
                  'learning_rate': 1e-3,
                },
                lr_decay=0.95,
                num_epochs=10, batch_size=100,
                print_every=100)
solver.train()

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

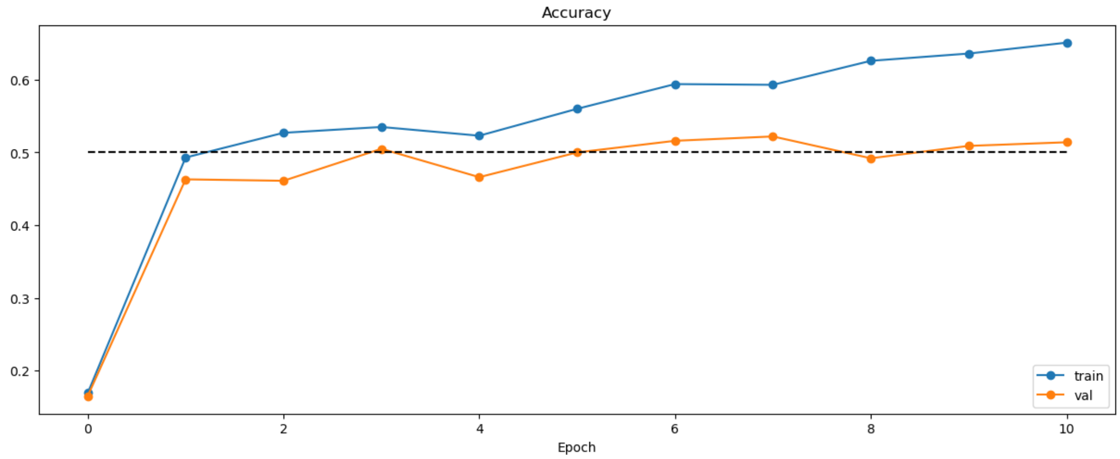
```
(Iteration 1 / 4900) loss: 2.302311
(Epoch 0 / 10) train acc: 0.169000; val_acc: 0.165000
(Iteration 101 / 4900) loss: 1.714482
(Iteration 201 / 4900) loss: 1.705519
(Iteration 301 / 4900) loss: 1.722993
(Iteration 401 / 4900) loss: 1.567324
(Epoch 1 / 10) train acc: 0.493000; val_acc: 0.463000
(Iteration 501 / 4900) loss: 1.569744
(Iteration 601 / 4900) loss: 1.544536
(Iteration 701 / 4900) loss: 1.488613
(Iteration 801 / 4900) loss: 1.407575
(Iteration 901 / 4900) loss: 1.471260
(Epoch 2 / 10) train acc: 0.527000; val_acc: 0.461000
(Iteration 1001 / 4900) loss: 1.353576
(Iteration 1101 / 4900) loss: 1.452678
(Iteration 1201 / 4900) loss: 1.303694
(Iteration 1301 / 4900) loss: 1.417480
(Iteration 1401 / 4900) loss: 1.325125
(Epoch 3 / 10) train acc: 0.535000; val_acc: 0.505000
(Iteration 1501 / 4900) loss: 1.314353
(Iteration 1601 / 4900) loss: 1.391857
(Iteration 1701 / 4900) loss: 1.251908
(Iteration 1801 / 4900) loss: 1.347470
(Iteration 1901 / 4900) loss: 1.205544
(Epoch 4 / 10) train acc: 0.523000; val_acc: 0.466000
(Iteration 2001 / 4900) loss: 1.212428
(Iteration 2101 / 4900) loss: 1.308649
(Iteration 2201 / 4900) loss: 1.180710
(Iteration 2301 / 4900) loss: 1.276170
(Iteration 2401 / 4900) loss: 0.995311
(Epoch 5 / 10) train acc: 0.560000; val_acc: 0.500000
(Iteration 2501 / 4900) loss: 1.240898
(Iteration 2601 / 4900) loss: 1.342720
(Iteration 2701 / 4900) loss: 1.097030
(Iteration 2801 / 4900) loss: 1.372613
(Iteration 2901 / 4900) loss: 1.361078
```

```
(Epoch 6 / 10) train acc: 0.594000; val_acc: 0.516000
(Iteration 3001 / 4900) loss: 1.283214
(Iteration 3101 / 4900) loss: 1.188728
(Iteration 3201 / 4900) loss: 1.273334
(Iteration 3301 / 4900) loss: 1.284245
(Iteration 3401 / 4900) loss: 1.123954
(Epoch 7 / 10) train acc: 0.593000; val_acc: 0.522000
(Iteration 3501 / 4900) loss: 1.204882
(Iteration 3601 / 4900) loss: 0.987374
(Iteration 3701 / 4900) loss: 0.978199
(Iteration 3801 / 4900) loss: 1.152009
(Iteration 3901 / 4900) loss: 0.993330
(Epoch 8 / 10) train acc: 0.626000; val_acc: 0.492000
(Iteration 4001 / 4900) loss: 1.136052
(Iteration 4101 / 4900) loss: 1.245813
(Iteration 4201 / 4900) loss: 1.007951
(Iteration 4301 / 4900) loss: 1.122997
(Iteration 4401 / 4900) loss: 1.229099
(Epoch 9 / 10) train acc: 0.636000; val_acc: 0.509000
(Iteration 4501 / 4900) loss: 1.000854
(Iteration 4601 / 4900) loss: 0.991933
(Iteration 4701 / 4900) loss: 1.093868
(Iteration 4801 / 4900) loss: 1.017199
(Epoch 10 / 10) train acc: 0.651000; val_acc: 0.514000
```

[170]:

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

plt.subplot(2, 1, 1)
plt.title('Training loss')
plt.plot(solver.loss_history, '-')
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()
```





1.8 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 HW #4.

[171]:
```
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)
```

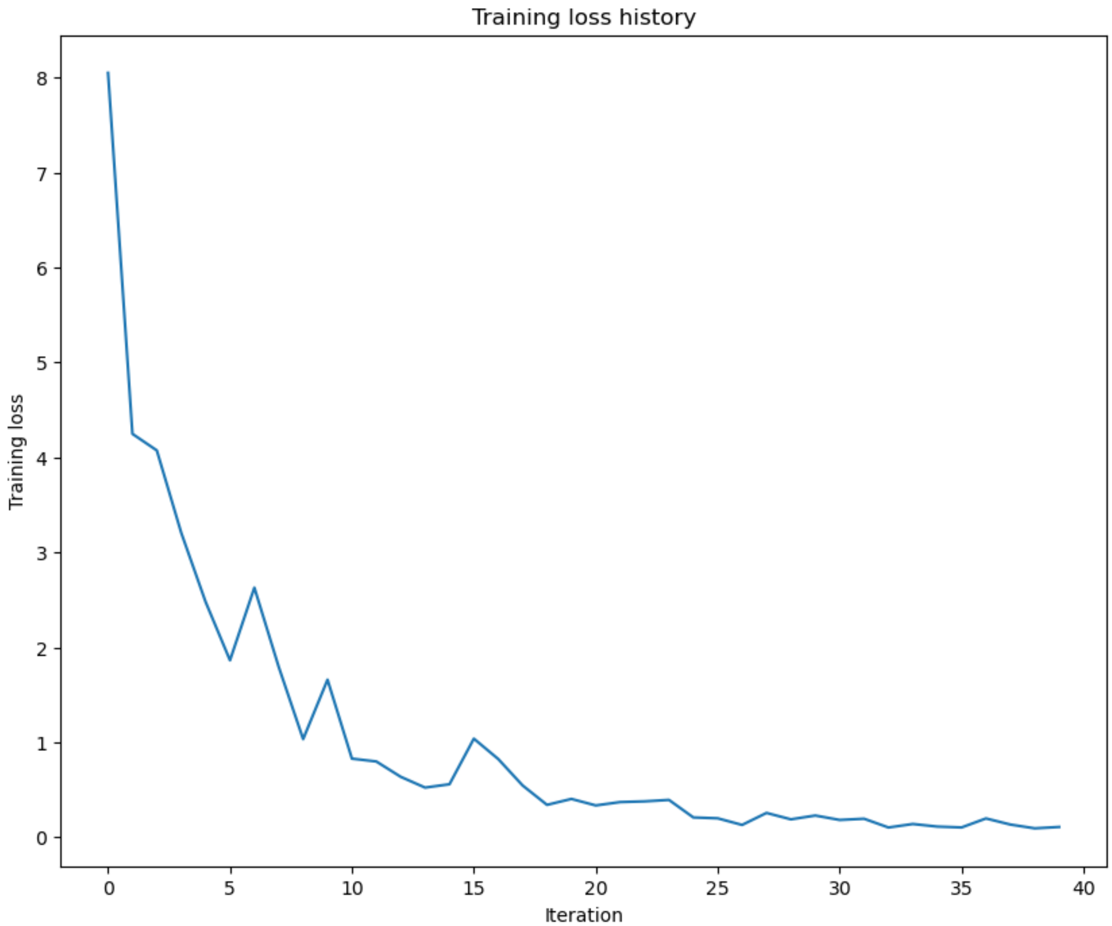
```
    print('Initial loss: {}'.format(loss))

    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: 2.30016839575148
W1 relative error: 2.0065648682203341e-07
W2 relative error: 2.1328229056219816e-05
W3 relative error: 2.2975730024776007e-06
b1 relative error: 1.117913018511571e-06
b2 relative error: 1.339219241775373e-09
b3 relative error: 1.528152980898322e-10
Running check with reg = 3.14
Initial loss: 7.101640034859168
W1 relative error: 1.642451932225041e-08
W2 relative error: 2.7828468127574597e-08
W3 relative error: 1.8514594191979535e-08
b1 relative error: 3.6439425585560934e-07
b2 relative error: 1.7028875924686097e-06
b3 relative error: 1.0645662363166267e-10
```

[172]:
```
# 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 dataset.
# Your training accuracy should be 1.0 to receive full credit on this part.
weight_scale = 3e-2
learning_rate = 5e-4

model = FullyConnectedNet([100, 100],
              weight_scale=weight_scale, dtype=np.float64)
solver = Solver(model, small_data,
```

```
                print_every=10, num_epochs=20, batch_size=25,
                update_rule='sgd',
                optim_config={
                  'learning_rate': learning_rate,
                }
         )
solver.train()

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

```
(Iteration 1 / 40) loss: 8.048919
(Epoch 0 / 20) train acc: 0.040000; val_acc: 0.100000
(Epoch 1 / 20) train acc: 0.180000; val_acc: 0.105000
(Epoch 2 / 20) train acc: 0.280000; val_acc: 0.124000
(Epoch 3 / 20) train acc: 0.460000; val_acc: 0.111000
(Epoch 4 / 20) train acc: 0.600000; val_acc: 0.132000
(Epoch 5 / 20) train acc: 0.640000; val_acc: 0.128000
(Iteration 11 / 40) loss: 0.827931
(Epoch 6 / 20) train acc: 0.820000; val_acc: 0.133000
(Epoch 7 / 20) train acc: 0.840000; val_acc: 0.120000
(Epoch 8 / 20) train acc: 0.860000; val_acc: 0.135000
(Epoch 9 / 20) train acc: 0.940000; val_acc: 0.127000
(Epoch 10 / 20) train acc: 0.940000; val_acc: 0.125000
(Iteration 21 / 40) loss: 0.335242
(Epoch 11 / 20) train acc: 0.940000; val_acc: 0.122000
(Epoch 12 / 20) train acc: 0.980000; val_acc: 0.134000
(Epoch 13 / 20) train acc: 0.980000; val_acc: 0.132000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.133000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.140000
(Iteration 31 / 40) loss: 0.182661
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.137000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.132000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.136000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.132000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.132000
```


Training loss history
Training loss
Iteration
0
1
2
3
4
5
6
7
8
0
5
10
15
20
25
30
35
40

```python
import numpy as np

from .layers import *
from .layer_utils import *


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['W1'] = weight_scale * np.random.randn(hidden_dims, input_dim)
    self.params['b1'] = np.zeros(hidden_dims)
    self.params['W2'] = weight_scale * np.random.randn(num_classes, hidden_dims)
    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.
    """
    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.
    # ================================================================ #
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    reg = self.reg
    h1, cache_h1 = affine_relu_forward(X, W1, b1)
    z2, cache_z2 = affine_forward(h1, W2, b2)
    scores = z2

    # ================================================================ #
    # 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.
    #
    #   And be sure to use the layers you prior implemented.
    # ================================================================ #
    loss, dL = softmax_loss(scores, y)
    dh1, dw2, db2 = affine_backward(dL, cache_z2)
    dx, dw1, db1 = affine_relu_backward(dh1, cache_h1)

    reg_loss = reg*0.5*(np.linalg.norm(W1, ord='fro')**2 + np.linalg.norm(W2, ord='fro')**2)
    loss += reg_loss
    dw1 += reg*W1
    dw2 += reg*W2

    grads['W1'] = dw1
    grads['b1'] = db1
```

```
    grads['W2'] = dw2
    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.
    # ================================================================ #
```

```
    for i, dims in enumerate(hidden_dims):
        if self.use_batchnorm:
            prev_size = (input_dim if i == 0 else hidden_dims[i-1])
            self.params[f'gamma_{i+1}'] = np.ones(prev_size)
            self.params[f'beta_{i+1}'] = np.zeros(prev_size)
        if i == 0:
            self.params['W1'] = weight_scale * np.random.randn(hidden_dims[i], input_dim)
            self.params['b1'] = np.zeros(hidden_dims[i])
        else:
            self.params[f'W{i+1}'] = weight_scale * np.random.randn(hidden_dims[i], hidden_dim
s[i-1])
            self.params[f'b{i+1}'] = np.zeros(hidden_dims[i])
    L = len(hidden_dims)+1
    self.params[f'W{L}'] = weight_scale * np.random.randn(num_classes, hidden_dims[len(hidden_
dims)-1])
    self.params[f'b{L}'] = np.zeros(num_classes)

    # ================================================================ #
    # 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.
    """
    """
    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.
    """
    X = X.astype(self.dtype)
    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".
    # ================================================================== #
    e = 1e-8
    reg = self.reg
    cache = []
    state = np.reshape(X, (len(X),-1))
    N = len(X)
    for i in range(1, self.num_layers):
        assert len(state) == N
        # Assuming state: (N x D)
        state, cache_ = affine_forward(state, self.params[f'W{i}'].T, self.params[f'b{i}'])
        cache.append(cache_)
        assert len(state) == N
        # Assuming state: (N x M)
        if self.use_batchnorm:
            mu = state.mean(axis=0)[:,np.newaxis]
            var = state.var(axis=0)[:,np.newaxis]
            x_hat = (state - mu)/np.sqrt(var+e)
            cache.append((mu, var, x_hat, state))
            state = self.params[f'gamma_{i}']*x_hat + self.params[f'beta_{i}']
            assert len(state) == N

        state, cache_ = relu_forward(state)
        cache.append(cache_)
        assert len(state) == N

        if self.use_dropout:
            if mode == 'test':
                state *= (1-self.dropout_param['p'])
            else:
                M = np.random.rand((state.shape[1])) < self.dropout_param['p']
                state *= M
                cache.append(M)
            assert len(state) == N


    L = self.num_layers
    state, cache_ = affine_forward(state, self.params[f'W{L}'].T, self.params[f'b{L}'])
    cache.append(cache_)
    scores = state
```

```
    # ================================================================ #
    # 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.
    # ================================================================ #

    loss, dL = softmax_loss(scores, y)
    dL, dw, db = affine_backward(dL, cache.pop())
    grads[f'W{L}'] = dw.T + reg*self.params[f'W{L}']
    grads[f'b{L}'] = db
    assert grads[f'W{L}'].shape == self.params[f'W{L}'].shape, f"{grads[f'W{L}'].shape=} | {se
lf.params[f'W{L}'].shape=}"
    assert grads[f'b{L}'].shape == self.params[f'b{L}'].shape, f"{grads[f'b{L}'].shape=} | {se
lf.params[f'b{L}'].shape=}"
    reg_loss = reg*0.5*(np.linalg.norm(self.params[f'W{L}'], ord='fro')**2)

    for i in range(self.num_layers-1, 0, -1):
        if self.use_dropout:
            if mode == 'test':
                dL *= (1-self.dropout_param['p'])
            else:
                M = cache.pop()
                dL *= M

        dL = relu_backward(dL, cache.pop())

        # Assuming dL: (N x D) - Each row is a feature, each column an observation. Sum along
observations
        if self.use_batchnorm:
            mu, var, x_hat, x = cache.pop()
            dbeta = dL.sum(axis=0)
            dgamma = (dL*x_hat).sum(x=0)
            dx_hat = dL*self.params[f'gamma_{i}']
            dmu = (-1/(var+e))*dx_hat.sum(axis=0)
            dvar = (-0.5/((var+e)**(3.0/2.0))) * ((x - mu)*dx_hat).sum(axis=0)
            dL = 1/((var+e)**(1.0/2.0))*dx_hat + 2*((x - mu)/len(m))*dvar + (1.0/len(x))*dmu
            grads[f'beta_{i}'] = dbeta
            grads[f'gamma_{i}'] = dgamma
            assert grads[f'gamma_{i}'].shape == self.params[f'gamma_{i}'].shape,\
                   f"{grads[f'gamma_{i}'].shape=} | {self.params[f'gamma_{i}'].shape=}"
            assert grads[f'beta_{i}'].shape == self.params[f'beta_{i}'].shape,\
                   f"{grads[f'beta_{i}'].shape=} | {self.params[f'beta_{i}'].shape=}"

        dL, dw, db = affine_backward(dL, cache.pop())
        grads[f'W{i}'] = dw.T + reg*self.params[f'W{i}']
        grads[f'b{i}'] = db
        assert grads[f'W{i}'].shape == self.params[f'W{i}'].shape, f"{grads[f'W{i}'].shape=} |
 {self.params[f'W{i}'].shape=}"
        assert grads[f'b{i}'].shape == self.params[f'b{i}'].shape, f"{grads[f'b{i}'].shape=} |
 {self.params[f'b{i}'].shape=}"
        reg_loss += reg*0.5*(np.linalg.norm(self.params[f'W{i}'], ord='fro')**2)

    loss += reg_loss
```

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

```
import numpy as np
import pdb




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.
    # ================================================================ #

    X = np.reshape(x, (len(x),-1))
    out = w.T@X.T + b[:,np.newaxis]
    out = out.T

    # ================================================================ #
    # 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:
    - 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.
    # ================================================================ #

    # dout is N x M
    # dx should be N x d1 x ... x dk; it relates to dout through multiplication with w, which
is D x M
    # dw should be D x M; it relates to dout through multiplication with x, which is N x D aft
er reshaping
    # db should be M; it is just the sum over dout examples
    X = np.reshape(x, (len(x),-1))

    db = dout.T.sum(axis=1)
    dw = dout.T@X
    dw = dw.T
    dx = w@dout.T
    dx = np.reshape(dx.T, x.shape)

    # ================================================================ #
    # 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 = x*(x > 0)
    # ================================================================ #
    # 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

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

    # ================================================================ #
```

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

    # ReLU directs linearly to those > 0
    dx = (x > 0)*dout

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

    return 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
```