

Deep Learning PS 1

1 Gradient Descent

1) Find ω^* of the following optimization:

$$\underset{\omega}{\operatorname{argmin}} f(\omega^{(t)}) + \langle \omega - \omega^{(t)}, \nabla f(\omega^{(t)}) \rangle + \frac{\lambda}{2} \|\omega - \omega^{(t)}\|^2$$

In order to find ω^* we must compute the first derivative of the above function and set it equal to 0 since we know it is a convex function.

$$\begin{aligned} \frac{d}{d\omega} [f(\omega^{(t)}) + \langle \nabla f(\omega^{(t)}) \omega - \nabla f(\omega^{(t)}) \omega^{(t)} + \frac{\lambda}{2} \|\omega - \omega^{(t)}\|^2] \\ = \frac{d}{d\omega} [\nabla f(\omega^{(t)}) \omega + \frac{\lambda}{2} \|\omega - \omega^{(t)}\|^2] \end{aligned}$$

$$\nabla f(\omega^{(t)}) + \lambda(\omega - \omega^{(t)}) = h(\omega)$$

$$h(\omega) = 0$$

$$0 = \nabla f(\omega^{(t)}) + \lambda(\omega - \omega^{(t)})$$

$$-\nabla f(\omega^{(t)}) = \lambda(\omega - \omega^{(t)})$$

$$\frac{-\nabla f(\omega^{(t)})}{\lambda} = \omega - \omega^{(t)}$$

$$\boxed{\omega^* = \omega^{(t)} + \frac{\nabla f(\omega^{(t)})}{\lambda}}$$

From the gradient descent update rule we get that $\omega^{t+1} = \omega^* - \eta \nabla f(\omega^*)$, hence

$$\eta = 1/\lambda$$

seminar 3 Oct 2019

1.29 exercise 2

2) Show that $\sum_{t=1}^T \langle w^t - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2$

$$\Rightarrow 0 \leq \frac{\|w^*\|^2}{2n} - \sum_{t=1}^T \langle w^t - w^*, v_t \rangle + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2$$

(inequality principle)
 $\|w^t - w^*\| \leq \|w^t\| + \|w^*\|$

$$\Rightarrow 0 \leq \frac{\|w^*\|^2}{n} - 2 \sum_{t=1}^T \langle w^t - w^*, v_t \rangle + n \sum_{t=1}^T \|v_t\|^2$$

strengthens terms on the right side of above eqn
restain

$$\Rightarrow 0 \leq \|w^*\|^2 - 2n \sum_{t=1}^T \langle w^t - w^*, v_t \rangle + n^2 \sum_{t=1}^T \|v_t\|^2$$

is zero on some t or T so $\exists t_1, t_2$ b/w

$$\Rightarrow 0 \leq \|w^*\|^2 + n \sum_{t=1}^T [-2 \langle w^t - w^*, v_t \rangle + n \|v_t\|^2]$$

$$\Rightarrow 0 \leq \|w^*\|^2 + \sum_{t=1}^T -2n \langle w^t - w^*, v_t \rangle + n^2 \sum_{t=1}^T \|v_t\|^2 + \|w^t - w^*\|^2$$

$\left[\sum_{t=1}^T \langle w^t - w^*, v_t \rangle \right] = - \sum_{t=1}^T \langle w^t - w^*, v_t \rangle$

At this point we have completed the square and are able to simplify the inequality the following way:

$$0 \leq \|w^*\|^2 + \sum_{t=1}^T \|w^t - w^* - nv_t\|^2 - \sum_{t=1}^T \|w^t - w^*\|^2$$

It is important to note that the given recurrence $w^{t+1} = w^t - nv_t$ w/ $w^1 = 0$ can be expressed as $w^{t+1} = -n \sum_{k=1}^t v_k$. Using this we can further simplify:

$$0 \leq \|w^*\|^2 + \sum_{t=1}^T \left\| -n \sum_{k=1}^t v_k - w^* - nv_t \right\|^2 - \sum_{t=1}^T \|w^t - w^*\|^2$$

$$0 \leq \|w^*\|^2 + \sum_{t=1}^T \|w^{t+1} - w^*\|^2 - \sum_{t=1}^T \|w^t - w^*\|^2$$

$$0 \leq \|w^*\|^2 + \sum_{t=1}^T \|w^{t+1} - w^*\|^2 - \|w^t - w^*\|^2$$

$$0 \leq \|w^*\|^2 + \|w^{t+1} - w^*\|^2 - \|w^t - w^*\|^2$$

$$0 \leq \|w^{t+1} - w^*\|^2 \therefore \text{we have shown that}$$

$$\sum_{t=1}^T \langle w^t - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2.$$

3) Show that for $\bar{w} = \frac{1}{T} \sum_{t=1}^T w_t$, $f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$$

Consider the fact that: $f'(w)$

$$f(w) \geq f(w_t) + \langle w - w_t, \nabla f(w_t) \rangle \quad \forall w$$

which is of form $\epsilon_t(w) \geq 0$

Given this we can arrive to the following statement:

$$f(w_t) - f(w^*) \leq \langle w_t - w^*, \nabla f(w_t) \rangle$$

$$\frac{1}{T} \sum_{t=1}^T [f(w_t) - f(w^*)] \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$$

$$\frac{1}{T} \sum_{t=1}^T f(w_t) - \frac{1}{T} \sum_{t=1}^T f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$$

This is possible by Jensen's Inequality: $\frac{1}{T} \sum_{t=1}^T f(w_t) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$

Inequality b.s. $f\left(\frac{1}{T} \sum_{t=1}^T w_t\right) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle$

$$\therefore f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle.$$

Given upper bound $B + \rho$ for $\|w^*\| + \|\nabla f(w^*)\|$ & $n = (B^2/\rho^2 T)^{1/2}$ we can show GD convergence at a rate of $O(1/\sqrt{T})$ as follows:

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w_t - w^*, \nabla f(w_t) \rangle \leq \frac{1}{T} \left[\frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|w_t\|^2 \right]$$

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \left[\frac{B^2}{2(B^2/\rho^2 T)^{1/2}} + \frac{(B^2/\rho^2 T)^{1/2}}{2} T \rho^2 \right]$$

$$\leq \frac{1}{T} \left[\frac{B^2}{2(B^2/\rho^2 T)^{1/2}} \right]$$

$$\leq \frac{1}{T} \left[\frac{B^2}{(B^2/\rho^2 T)^{1/2}} \right] \approx \frac{1}{T} [\sqrt{T}] \therefore f(\bar{w}) - f(w^*) \leq 1/\sqrt{T}$$

and the convergence rate is $O(1/\sqrt{T})$.

4) Counter Example: SGD not fall well

First lets show that $f(w)$ has a minimum at 0:

$$f'(w) = \frac{e^w - 1}{e^w + 1} \text{ int. 2nd}$$

~~$f'(w)$~~ is equal to 0 when

~~$x=0$~~ . (int. 2nd)

this means that no matter if we choose $f_1(w)$ or $f_2(w)$, the derivative at $x=0$ should be 0 in order for the overall loss function to ~~decrease~~ not decrease since we are already at the minimum.

(~~100%~~) Let's suppose $f_2(w)$ is chosen.

$$f_2'(w) = \frac{-1}{e^w + 1} \text{ the value of } f_2'(w) \text{ at}$$

~~$x=0$~~ is $\frac{-1}{1+1} = \frac{-1}{2}$. This means that the

value of $f_2'(w) \neq 0$ & thus we have

shown that SGD is not guaranteed to

decrease the over all loss function at

every iteration in a case like this

one. Since it will perform an update

using the value of $f_2'(0)$.

the responses of w won't be

updated in (θ_1) to stop a

loss of $f_2(\theta_1)$ but $f_2(\theta_1) - f_2(0)$

$= e^{0.5} - 1 > 0$

For $f_2(\theta_1) =$

$f_2(\theta_1) = \frac{e^{0.5} - 1}{e^{0.5} + 1}$

$f_2(\theta_1) < 0$ but $f_2(\theta_1) > f_2(0)$

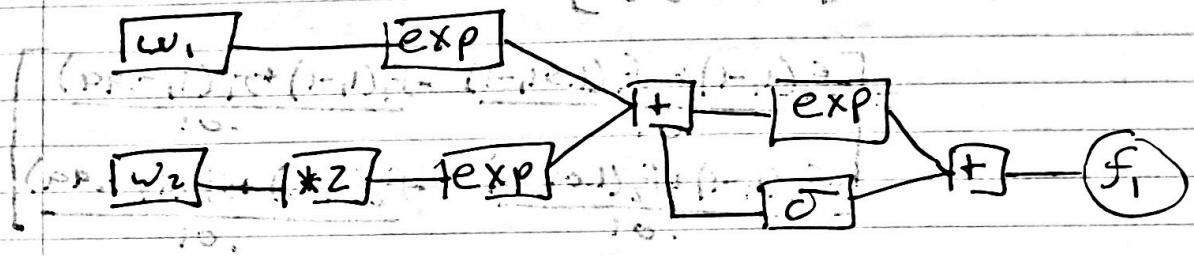
2 Automatic Differentiation

5a)

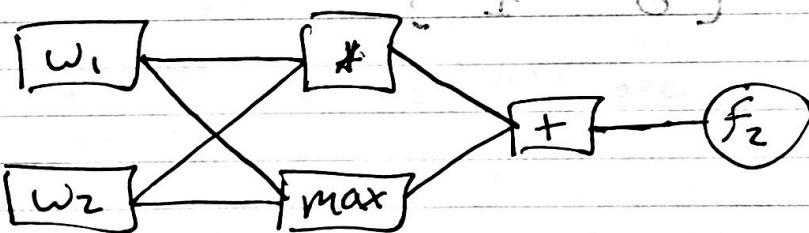
$$f_1$$

$$10. = w_1 \cdot w_2$$

$$\begin{bmatrix} 3.6 & 0.6 \\ 0.6 & 3.6 \end{bmatrix} = \frac{3.6}{0.6}$$



f_2



$$f(\vec{\omega}) = \langle f_1(\vec{\omega}), f_2(\vec{\omega}) \rangle$$

$$f((1, -1)) = \langle f_1((1, -1)), f_2((1, -1)) \rangle$$

$$f_1(1, -1) = e^{e^1 + e^{-2}} + \sigma(e^1 + e^{-2}) = 18.296$$

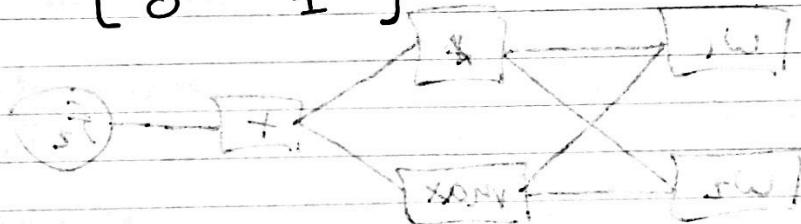
$$f_2(1, -1) = 1(-1) + \max(1, -1) = 0$$

$$\tilde{f}((1, -1)) = \langle 18.296, 0 \rangle$$

b)

$$\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix} \text{ for } \Delta w = .01$$

$$= \begin{bmatrix} \frac{-f_1(1, -1) + f_1(1.01, -1)}{.01} & \frac{-f_1(1, -1) + f_1(1, -0.99)}{.01} \\ \frac{-f_2(1, -1) + f_2(1.01, -1)}{.01} & \frac{-f_2(1, -1) + f_2(1, -0.99)}{.01} \end{bmatrix}$$
$$= \begin{bmatrix} 48.2 & 4.8 \\ 0 & 1 \end{bmatrix}$$



$$\langle (5.1)^2, (7.2)^2, 1.0 \rangle = (5.1)^2$$

$$\langle (5.1 \cdot 7.2), (1.1 \cdot 7.2), 1.0 \rangle = (1.1 \cdot 7.2)$$

$$\Delta PS.81 = (5.1^2) \cdot 7.2 + (7.2^2) \cdot 1.1 + (1.1 \cdot 7.2)$$

$$0 = (1.1 \cdot 7.2) \cdot 5.1 + (1.1 \cdot 7.2) = (1.1 \cdot 7.2)^2$$

$$\langle 0, \Delta PS.81 \rangle = (1.1 \cdot 7.2)^2$$

$$c) \frac{\partial f_2}{\partial a} = w_1, w_2 \quad b = \max(w_1, w_2) \quad f_2 = a + b$$

~~(b) f_2 based on $a+d=b$~~

$$\frac{\partial a}{\partial w_1} = w_2 \quad \frac{\partial b}{\partial w_1} = \begin{cases} 1, & w_1 > w_2 \\ 0, & \text{otherwise} \end{cases}$$

$$\left(\frac{\partial f_2}{\partial w_1} \right) = \frac{\partial a}{\partial w_1} + \frac{\partial b}{\partial w_1} = w_2 + \begin{cases} 1, & w_1 > w_2 \\ 0, & \text{otherwise} \end{cases}$$

w₂

$$\frac{\partial a}{\partial w_2} = w_1 \quad \frac{\partial b}{\partial w_2} = \begin{cases} 1, & w_1 < w_2 \\ 0, & \text{otherwise} \end{cases}$$

$$\left(\frac{\partial f_2}{\partial w_2} \right) = \frac{\partial a}{\partial w_2} + \frac{\partial b}{\partial w_2} = w_1 + \begin{cases} 1, & w_1 < w_2 \\ 0, & \text{otherwise} \end{cases}$$

f₁

$$a = 2w_2, \quad b = e^{w_1}, \quad c = e^a, \quad d = b+c, \quad h = c^d, \quad i = \sigma(d)$$

$$f_1 = h+i$$

$$\left(\frac{\partial f_1}{\partial w_1} \right) = \dots$$

$$\frac{\partial a}{\partial w_1} = 0, \quad \frac{\partial b}{\partial w_1} = e^{w_1}, \quad \frac{\partial c}{\partial w_1} = 0, \quad \frac{\partial d}{\partial w_1} = e^{w_1}, \quad \frac{\partial h}{\partial w_1} = e^{w_1} \cdot e^d$$

$$\frac{\partial i}{\partial w_1} = (1 - \sigma(d)) \sigma(d) e^{w_1}, \quad \frac{\partial f_1}{\partial w_1} = e^{w_1} e^{e^{w_1} + e^{2w_2}} + (1 - \sigma(e^{w_1} + e^{2w_2})) \sigma(e^{w_1} + e^{2w_2}) e^{w_1}$$

$$\frac{\partial a}{\partial w_2} = 2, \quad \frac{\partial b}{\partial w_2} = 0, \quad \frac{\partial c}{\partial w_2} = 2e^a, \quad \frac{\partial d}{\partial w_2} = 2e^a, \quad \frac{\partial h}{\partial w_2} = 2e^a e^d$$

$$\frac{\partial i}{\partial w_2} = (1 - \sigma(d)) \sigma(d) 2e^a, \quad \frac{\partial f_1}{\partial w_2} = 2e^{2w_2} e^{e^{w_1} + e^{2w_2}} + (1 - \sigma(e^{w_1} + e^{2w_2})) \sigma(e^{w_1} + e^{2w_2}) 2e^{2w_2}$$

$$\frac{\partial f}{\partial \vec{w}} = \begin{bmatrix} 42.303 & 4.710 \\ 0 & 1 \end{bmatrix} \text{ at } \vec{w} = \begin{bmatrix} 1.150 & 2.0348 \\ 0 \end{bmatrix} = \begin{bmatrix} 2.6 \\ 6.6 \end{bmatrix}$$

c) $f_1 = \max(w_1, w_2)$ if $w_1 > w_2$
 $a = 2w_2, b = e^{w_1}, c = e^a, d = b+c, h = e^d, i = \sigma(d)$

$$f_1 = h+i$$

$$\frac{\partial f_1}{\partial h} = 1, \frac{\partial f_1}{\partial i} = 1, \frac{\partial f_1}{\partial d} = \frac{\partial f_1}{\partial h} \frac{\partial h}{\partial d} + \frac{\partial f_1}{\partial i} \frac{\partial i}{\partial d}$$

$$= e^d + (1 - \sigma(d)) \sigma'(d)$$

$$\frac{\partial f_1}{\partial c} = \frac{\partial f_1}{\partial d} \frac{\partial d}{\partial c} = e^d + (1 - \sigma(d)) \sigma'(d)$$

$$\frac{\partial f_1}{\partial b} = \frac{\partial f_1}{\partial d} \frac{\partial d}{\partial b} = e^d + (1 - \sigma(d)) \sigma'(d)$$

$$\frac{\partial f_1}{\partial a} = \frac{\partial f_1}{\partial c} \frac{\partial c}{\partial a} = [e^d + (1 - \sigma(d)) \sigma'(d)] e^{w_1}$$

$$\frac{\partial f_1}{\partial w_1} = \frac{\partial f_1}{\partial b} \frac{\partial b}{\partial w_1} = [e^d + (1 - \sigma(d)) \sigma'(d)] e^{w_1}$$

$$(b) \frac{\partial f_1}{\partial w_2} = \frac{\partial f_1}{\partial a} e^{w_1} e^{2w_2} + (1 - \sigma(e^{w_1} + e^{2w_2})) \sigma'(e^{w_1} + e^{2w_2}) e^{w_1}$$

$$\frac{\partial f_1}{\partial w_2} = \frac{\partial f_1}{\partial a} \frac{\partial a}{\partial w_2} = [e^d + (1 - \sigma(d)) \sigma'(d)] 2e^{w_2}$$

$$= [e^{2w_2} e^{w_1} + e^{2w_2} + (1 - \sigma(e^{w_1} + e^{2w_2})) \sigma'(e^{w_1} + e^{2w_2})] 2e^{w_2}$$

(c) $f_2 = \min(w_1, w_2)$ if $w_1 < w_2$
 $a = w_1, w_2, b = \max(w_1, w_2), f_2 = a+b$

$$\frac{\partial f_2}{\partial a} = 1, \frac{\partial f_2}{\partial b} = 1, \frac{\partial f_2}{\partial w_1} = \frac{\partial f_2}{\partial a} \frac{\partial a}{\partial w_1} + \frac{\partial f_2}{\partial b} \frac{\partial b}{\partial w_1}$$

$$= w_2 + \begin{cases} 1, & w_1 > w_2 \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial f_2}{\partial w_2} = \frac{\partial f_2}{\partial a} \frac{\partial a}{\partial w_2} + \frac{\partial f_2}{\partial b} \frac{\partial b}{\partial w_2} = w_1 + \begin{cases} 1, & w_2 > w_1 \\ 0, & \text{otherwise} \end{cases}$$

$$\vec{\frac{\partial f}{\partial \bar{w}}} = \begin{bmatrix} 47.303 & 4.710 \\ 0 & 1 \end{bmatrix} \text{ at } \bar{w} = (1, -1)$$

David Jose Fernandez

Paper Review

Weight Agnostic Neural Networks

The key contributions of this paper is that the author's introduce the concept of creating networks that are naturally suited to solve a problem. Instead of taking the usual approach of hand crafting an architecture and then painstakingly training the network and tuning the hyper parameters in order to squeeze out every bit of accuracy, the main focus of this paper was to place more focus on the architecture and less on finding the optimal weights. This idea is extremely interesting and powerful since a network that can have better than chance performance only based on its architecture is probably capable of performing extremely well on the task at hand once the weights are properly trained. This idea goes hand in hand with the notion of precocial behaviors introduced in the abstract and is very relatable to real life, if you are naturally good at something then your skill ceiling is much higher or you are able to learn much faster than others. One of the strengths of the paper is that they do a very good job of clearly explaining the process used to generate these weight agnostic networks by providing a clear psudo-genetic algorithm that can be used to generate new architectures from previous ones that had performed best. I also believe that their effort to put emphasis on prioritizing less complex networks is a plus to the paper since less complex networks are easier to train and potentially less of a black box. One thing that I would consider a weakness

For sure my biggest takeaway from reading this paper is that it is far more important to find a good architecture than to spend hours upon hours training a network using different hyper-parameters hoping to find the optimal combination that will provide the highest accuracy. The work done in this paper clearly shows that a good architecture is a lot more important than training the weights and finding the right hyper-parameters. Previously I thought that the most important part of creating a model that had good performance on a task was properly tuning the weights and finding that "global minimum" for the loss function, but this paper has changed my opinion and showed me that the work you do before training can be more important than the training itself.

softmax

February 12, 2020

1 Softmax Classifier

This exercise guides you through the process of classifying images using a Softmax classifier. As part of this you will:

- Implement a fully vectorized loss function for the Softmax classifier
- Calculate the analytical gradient using vectorized code
- Tune hyperparameters on a validation set
- Optimize the loss function with Stochastic Gradient Descent (SGD)
- Visualize the learned weights

```
[1]: # start-up code!
import random

import matplotlib.pyplot as plt
import numpy as np

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

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

```
[3]: from load_cifar10_tvt import load_cifar10_train_val

X_train, y_train, X_val, y_val, X_test, y_test = load_cifar10_train_val()
print("Train data shape: ", X_train.shape)
print("Train labels shape: ", y_train.shape)
print("Val data shape: ", X_val.shape)
print("Val labels shape: ", y_val.shape)
print("Test data shape: ", X_test.shape)
print("Test labels shape: ", y_test.shape)
```

```

Train, validation and testing sets have been created as
X_i and y_i where i=train,val,test
Train data shape: (3073, 49000)
Train labels shape: (49000,)
Val data shape: (3073, 1000)
Val labels shape: (1000,)
Test data shape: (3073, 1000)
Test labels shape: (1000,)

```

Code for this section is to be written in cs231n/classifiers/softmax.py

[4]: # Now, implement the vectorized version in softmax_loss_vectorized.

```

import time

from cs231n.classifiers.softmax import softmax_loss_vectorized

# gradient check.
from cs231n.gradient_check import grad_check_sparse

W = np.random.randn(10, 3073) * 0.0001

tic = time.time()
loss, grad = softmax_loss_vectorized(W, X_train, y_train, 0.00001)
toc = time.time()
print("vectorized loss: %e computed in %fs" % (loss, toc - tic))

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

f = lambda w: softmax_loss_vectorized(w, X_train, y_train, 0.0)[0]
grad_numerical = grad_check_sparse(f, W, grad, 10)

```

```

vectorized loss: 2.320558e+00 computed in 0.158436s
loss: 2.320558
sanity check: 2.302585
numerical: -1.655697 analytic: -1.655697, relative error: 1.546679e-08
numerical: 1.525072 analytic: 1.525072, relative error: 9.998561e-09
numerical: 0.037356 analytic: 0.037356, relative error: 1.300151e-07
numerical: -0.493021 analytic: -0.493021, relative error: 2.896674e-08
numerical: 0.630111 analytic: 0.630111, relative error: 2.458352e-08
numerical: 1.965181 analytic: 1.965181, relative error: 1.988519e-08
numerical: 0.906063 analytic: 0.906063, relative error: 1.552846e-08
numerical: 2.729657 analytic: 2.729656, relative error: 2.394028e-08
numerical: -0.188580 analytic: -0.188580, relative error: 3.347255e-08
numerical: -0.489200 analytic: -0.489200, relative error: 1.041207e-07

```

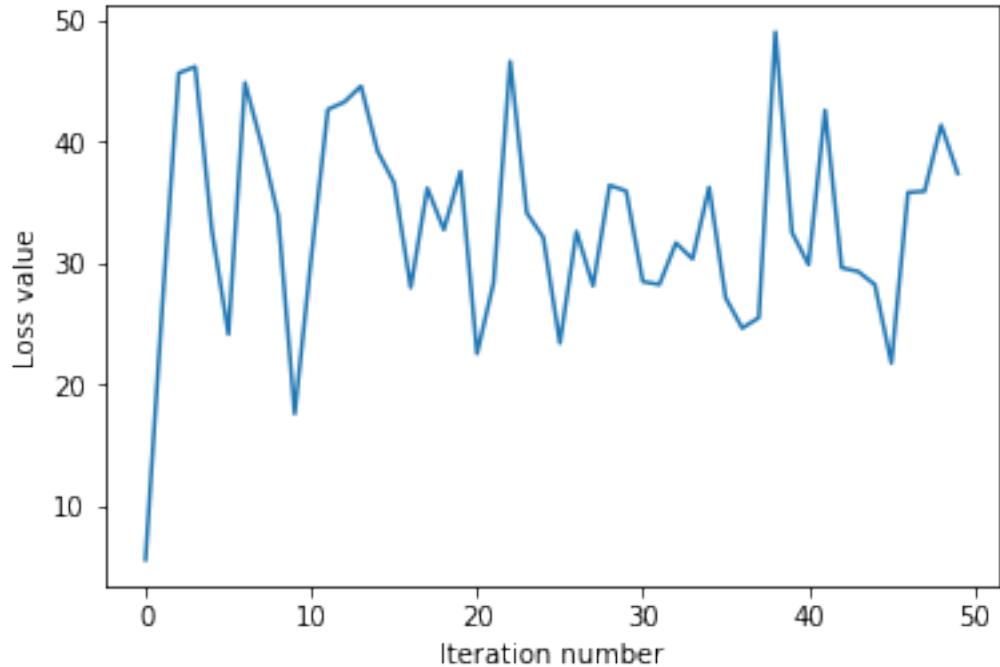
Code for this section is to be written in cs231n/classifiers/linear_classifier.py

```
[7]: # Now that efficient implementations to calculate loss function and gradient of softmax are ready,
# use it to train the classifier on the cifar-10 data
# Complete the `train` function in cs231n/classifiers/linear_classifier.py

from cs231n.classifiers.linear_classifier import Softmax

classifier = Softmax()
loss_hist = classifier.train(
    X_train,
    y_train,
    learning_rate=1e-4,
    reg=1e-6,
    num_iters=50,
    batch_size=500,
    verbose=False,
)
# Plot loss vs. iterations
plt.plot(loss_hist)
plt.xlabel("Iteration number")
plt.ylabel("Loss value")
```

[7]: Text(0, 0.5, 'Loss value')



```
[8]: # Complete the `predict` function in cs231n/classifiers/linear_classifier.py
# Evaluate on test set
y_test_pred = classifier.predict(X_test)
test_accuracy = np.mean(y_test == y_test_pred)
print("softmax on raw pixels final test set accuracy: %f" % (test_accuracy,))
```

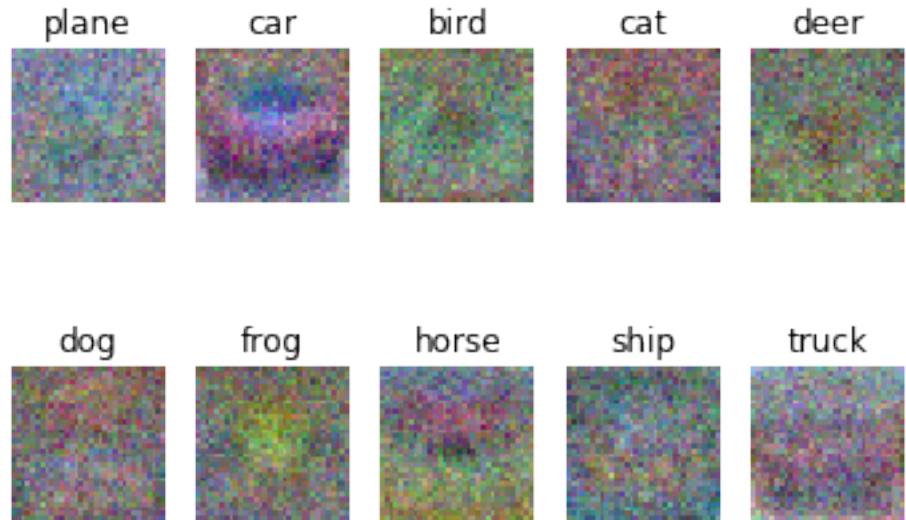
softmax on raw pixels final test set accuracy: 0.299000

```
[9]: # Visualize the learned weights for each class
w = classifier.W[:, :-1] # strip out the bias
w = w.reshape(10, 32, 32, 3)

w_min, w_max = np.min(w), np.max(w)

classes = [
    "plane",
    "car",
    "bird",
    "cat",
    "deer",
    "dog",
    "frog",
    "horse",
    "ship",
    "truck",
]
for i in range(10):
    plt.subplot(2, 5, i + 1)

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



two_layer_net

February 12, 2020

1 Implementing a Neural Network

In this exercise we will develop a neural network with fully-connected layers to perform classification, and test it out on the CIFAR-10 dataset.

[1]: # A bit of setup

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

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

# for auto-reloading 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 neural network parameters will be stored in a dictionary (`model` below), where the keys are the parameter names and the values are numpy arrays. Below, we initialize toy data and a toy model that we will use to verify your implementations.

[2]: # Create some toy data to check your implementations

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

def init_toy_model():
    model = {}
```

```

model['W1'] = np.linspace(-0.2, 0.6, num=input_size*hidden_size).
    ↪reshape(input_size, hidden_size)
model['b1'] = np.linspace(-0.3, 0.7, num=hidden_size)
model['W2'] = np.linspace(-0.4, 0.1, num=hidden_size*num_classes).
    ↪reshape(hidden_size, num_classes)
model['b2'] = np.linspace(-0.5, 0.9, num=num_classes)
return model

def init_toy_data():
    X = np.linspace(-0.2, 0.5, num=num_inputs*input_size).reshape(num_inputs, ↴
        input_size)
    y = np.array([0, 1, 2, 2, 1])
    return X, y

model = init_toy_model()
X, y = init_toy_data()

```

2 Forward pass: compute scores

Open the file cs231n/classifiers/neural_net.py and look at the function two_layer_net. This function is very similar to the loss functions you have written for the Softmax exercise in HW0: It takes the data and weights and computes the class scores, the loss, and the gradients on the parameters.

Implement the first part of the forward pass which uses the weights and biases to compute the scores for all inputs.

```
[3]: from cs231n.classifiers.neural_net import two_layer_net

scores = two_layer_net(X, model)
print(scores)
correct_scores = [[-0.5328368, 0.20031504, 0.93346689],
                  [-0.59412164, 0.15498488, 0.9040914 ],
                  [-0.67658362, 0.08978957, 0.85616275],
                  [-0.77092643, 0.01339997, 0.79772637],
                  [-0.89110401, -0.08754544, 0.71601312]]
```

```
# the difference should be very small. We get 3e-8
print('Difference between your scores and correct scores:')
print(np.sum(np.abs(scores - correct_scores)))
```

```
[[[-0.5328368  0.20031504  0.93346689]
  [-0.59412164  0.15498488  0.9040914 ]
  [-0.67658362  0.08978957  0.85616275]
  [-0.77092643  0.01339997  0.79772637]
  [-0.89110401 -0.08754544  0.71601312]]
Difference between your scores and correct scores:
3.848682278081994e-08
```

3 Forward pass: compute loss

In the same function, implement the second part that computes the data and regularization loss.

```
[4]: reg = 0.1
loss, _ = two_layer_net(X, model, y, reg)
correct_loss = 1.38191946092

# should be very small, we get 5e-12
print('Difference between your loss and correct loss:')
print(np.sum(np.abs(loss - correct_loss)))
```

Difference between your loss and correct loss:

4.6771475581408595e-12

4 Backward pass

Implement the rest of the function. This will compute the gradient of the loss with respect to the variables W_1 , b_1 , W_2 , and b_2 . Now that you (hopefully!) have a correctly implemented forward pass, you can debug your backward pass using a numeric gradient check:

```
[5]: 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 = two_layer_net(X, model, y, reg)

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

```
W2 max relative error: 1.314524e-09
b2 max relative error: 7.311059e-11
W1 max relative error: 4.426512e-09
b1 max relative error: 2.746125e-08
```

5 Train the network

To train the network we will use SGD with Momentum. Last assignment you implemented vanilla SGD. You will now implement the momentum update and the RMSProp update. Open the file

`classifier_trainer.py` and familiarize yourself with the `ClassifierTrainer` class. It performs optimization given an arbitrary cost function data, and model. By default it uses vanilla SGD, which we have already implemented for you. First, run the optimization below using Vanilla SGD:

```
[7]: from cs231n.classifier_trainer import ClassifierTrainer

model = init_toy_model()
trainer = ClassifierTrainer()
# call the trainer to optimize the loss
# Notice that we're using sample_batches=False, so we're performing GradientDescent (no sampled batches of data)
best_model, loss_history, _, _ = trainer.train(X, y, X, y,
                                                model, two_layer_net,
                                                reg=0.001,
                                                learning_rate=1e-4, momentum=0.0,
                                                learning_rate_decay=1,
                                                update='sgd', sample_batches=False,
                                                num_epochs=100,
                                                verbose=False)
print('Final loss with vanilla SGD: %f' % (loss_history[-1],))
```

```
starting iteration 0
starting iteration 10
starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
Final loss with vanilla SGD: 1.123347
```

Now fill in the **momentum update** in the first missing code block inside the `train` function, and run the same optimization as above but with the momentum update. You should see a much better result in the final obtained loss:

```
[8]: model = init_toy_model()
trainer = ClassifierTrainer()
# call the trainer to optimize the loss
# Notice that we're using sample_batches=False, so we're performing GradientDescent (no sampled batches of data)
best_model, loss_history, _, _ = trainer.train(X, y, X, y,
                                                model, two_layer_net,
                                                reg=0.001,
                                                learning_rate=1e-1, momentum=0.9,
                                                learning_rate_decay=1,
```

```

        update='momentum', □
→sample_batches=False,
                           num_epochs=100,
                           verbose=False)
correct_loss = 0.494394
print('Final loss with momentum SGD: %f. We get: %f' % (loss_history[-1], □
→correct_loss))

```

```

starting iteration 0
starting iteration 10
starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
Final loss with momentum SGD: 0.494394. We get: 0.494394

```

The **RMSProp** update step is given as follows:

```

cache = decay_rate * cache + (1 - decay_rate) * dx**2
x += - learning_rate * dx / np.sqrt(cache + 1e-8)

```

Here, `decay_rate` is a hyperparameter and typical values are [0.9, 0.99, 0.999].

Implement the **RMSProp** update rule inside the `train` function and rerun the optimization:

```

[9]: model = init_toy_model()
trainer = ClassifierTrainer()
# call the trainer to optimize the loss
# Notice that we're using sample_batches=False, so we're performing Gradient□
→Descent (no sampled batches of data)
best_model, loss_history, _, _ = trainer.train(X, y, X, y,
                                               model, two_layer_net,
                                               reg=0.001,
                                               learning_rate=1e-1, momentum=0.9, □
→learning_rate_decay=1,
                                               update='rmsprop', □
→sample_batches=False,
                                               num_epochs=100,
                                               verbose=False)
correct_loss = 0.439368
print('Final loss with RMSProp: %f. We get: %f' % (loss_history[-1], □
→correct_loss))

```

```

starting iteration 0
starting iteration 10

```

```

starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
Final loss with RMSProp: 0.439368. We get: 0.439368

```

6 Load the data

Now that you have implemented a two-layer network that passes gradient checks, it's time to load up our favorite CIFAR-10 data so we can use it to train a classifier.

```
[10]: 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.
    """
    # Load the raw CIFAR-10 data
    cifar10_dir = 'cs231n/datasets/cifar-10-batches-py'
    X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

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

```

7 Train a network

To train our network we will use SGD with momentum. In addition, we will adjust the learning rate with an exponential learning rate schedule as optimization proceeds; after each epoch, we will reduce the learning rate by multiplying it by a decay rate.

```
[11]: from cs231n.classifiers.neural_net import init_two_layer_model

model = init_two_layer_model(32*32*3, 50, 10) # input size, hidden size, number
                                                →of classes
trainer = ClassifierTrainer()
best_model, loss_history, train_acc, val_acc = trainer.train(X_train, y_train, u
                                                →X_val, y_val,
                                                model, two_layer_net,
                                                num_epochs=5, reg=1.0,
                                                momentum=0.9, learning_rate_decay
                                                →= 0.95,
                                                learning_rate=1e-5, verbose=True)
```

```

starting iteration 0
Finished epoch 0 / 5: cost 2.302593, train: 0.096000, val 0.096000, lr
1.000000e-05
starting iteration 10
starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50

```

```
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
starting iteration 100
starting iteration 110
starting iteration 120
starting iteration 130
starting iteration 140
starting iteration 150
starting iteration 160
starting iteration 170
starting iteration 180
starting iteration 190
starting iteration 200
starting iteration 210
starting iteration 220
starting iteration 230
starting iteration 240
starting iteration 250
starting iteration 260
starting iteration 270
starting iteration 280
starting iteration 290
starting iteration 300
starting iteration 310
starting iteration 320
starting iteration 330
starting iteration 340
starting iteration 350
starting iteration 360
starting iteration 370
starting iteration 380
starting iteration 390
starting iteration 400
starting iteration 410
starting iteration 420
starting iteration 430
starting iteration 440
starting iteration 450
starting iteration 460
starting iteration 470
starting iteration 480
Finished epoch 1 / 5: cost 2.281309, train: 0.155000, val 0.177000, lr
9.500000e-06
starting iteration 490
starting iteration 500
starting iteration 510
```

```
starting iteration 520
starting iteration 530
starting iteration 540
starting iteration 550
starting iteration 560
starting iteration 570
starting iteration 580
starting iteration 590
starting iteration 600
starting iteration 610
starting iteration 620
starting iteration 630
starting iteration 640
starting iteration 650
starting iteration 660
starting iteration 670
starting iteration 680
starting iteration 690
starting iteration 700
starting iteration 710
starting iteration 720
starting iteration 730
starting iteration 740
starting iteration 750
starting iteration 760
starting iteration 770
starting iteration 780
starting iteration 790
starting iteration 800
starting iteration 810
starting iteration 820
starting iteration 830
starting iteration 840
starting iteration 850
starting iteration 860
starting iteration 870
starting iteration 880
starting iteration 890
starting iteration 900
starting iteration 910
starting iteration 920
starting iteration 930
starting iteration 940
starting iteration 950
starting iteration 960
starting iteration 970
Finished epoch 2 / 5: cost 1.944457, train: 0.223000, val 0.247000, lr
9.025000e-06
```

starting iteration 980
starting iteration 990
starting iteration 1000
starting iteration 1010
starting iteration 1020
starting iteration 1030
starting iteration 1040
starting iteration 1050
starting iteration 1060
starting iteration 1070
starting iteration 1080
starting iteration 1090
starting iteration 1100
starting iteration 1110
starting iteration 1120
starting iteration 1130
starting iteration 1140
starting iteration 1150
starting iteration 1160
starting iteration 1170
starting iteration 1180
starting iteration 1190
starting iteration 1200
starting iteration 1210
starting iteration 1220
starting iteration 1230
starting iteration 1240
starting iteration 1250
starting iteration 1260
starting iteration 1270
starting iteration 1280
starting iteration 1290
starting iteration 1300
starting iteration 1310
starting iteration 1320
starting iteration 1330
starting iteration 1340
starting iteration 1350
starting iteration 1360
starting iteration 1370
starting iteration 1380
starting iteration 1390
starting iteration 1400
starting iteration 1410
starting iteration 1420
starting iteration 1430
starting iteration 1440
starting iteration 1450

```
starting iteration 1460
Finished epoch 3 / 5: cost 1.863379, train: 0.278000, val 0.284000, lr
8.573750e-06
starting iteration 1470
starting iteration 1480
starting iteration 1490
starting iteration 1500
starting iteration 1510
starting iteration 1520
starting iteration 1530
starting iteration 1540
starting iteration 1550
starting iteration 1560
starting iteration 1570
starting iteration 1580
starting iteration 1590
starting iteration 1600
starting iteration 1610
starting iteration 1620
starting iteration 1630
starting iteration 1640
starting iteration 1650
starting iteration 1660
starting iteration 1670
starting iteration 1680
starting iteration 1690
starting iteration 1700
starting iteration 1710
starting iteration 1720
starting iteration 1730
starting iteration 1740
starting iteration 1750
starting iteration 1760
starting iteration 1770
starting iteration 1780
starting iteration 1790
starting iteration 1800
starting iteration 1810
starting iteration 1820
starting iteration 1830
starting iteration 1840
starting iteration 1850
starting iteration 1860
starting iteration 1870
starting iteration 1880
starting iteration 1890
starting iteration 1900
starting iteration 1910
```

```
starting iteration 1920
starting iteration 1930
starting iteration 1940
starting iteration 1950
Finished epoch 4 / 5: cost 1.957268, train: 0.347000, val 0.342000, lr
8.145063e-06
starting iteration 1960
starting iteration 1970
starting iteration 1980
starting iteration 1990
starting iteration 2000
starting iteration 2010
starting iteration 2020
starting iteration 2030
starting iteration 2040
starting iteration 2050
starting iteration 2060
starting iteration 2070
starting iteration 2080
starting iteration 2090
starting iteration 2100
starting iteration 2110
starting iteration 2120
starting iteration 2130
starting iteration 2140
starting iteration 2150
starting iteration 2160
starting iteration 2170
starting iteration 2180
starting iteration 2190
starting iteration 2200
starting iteration 2210
starting iteration 2220
starting iteration 2230
starting iteration 2240
starting iteration 2250
starting iteration 2260
starting iteration 2270
starting iteration 2280
starting iteration 2290
starting iteration 2300
starting iteration 2310
starting iteration 2320
starting iteration 2330
starting iteration 2340
starting iteration 2350
starting iteration 2360
starting iteration 2370
```

```

starting iteration 2380
starting iteration 2390
starting iteration 2400
starting iteration 2410
starting iteration 2420
starting iteration 2430
starting iteration 2440
Finished epoch 5 / 5: cost 1.774308, train: 0.341000, val 0.360000, lr
7.737809e-06
finished optimization. best validation accuracy: 0.360000

```

8 Debug the training

With the default parameters we provided above, you should get a validation accuracy of about 0.37 on the validation set. This isn't very good.

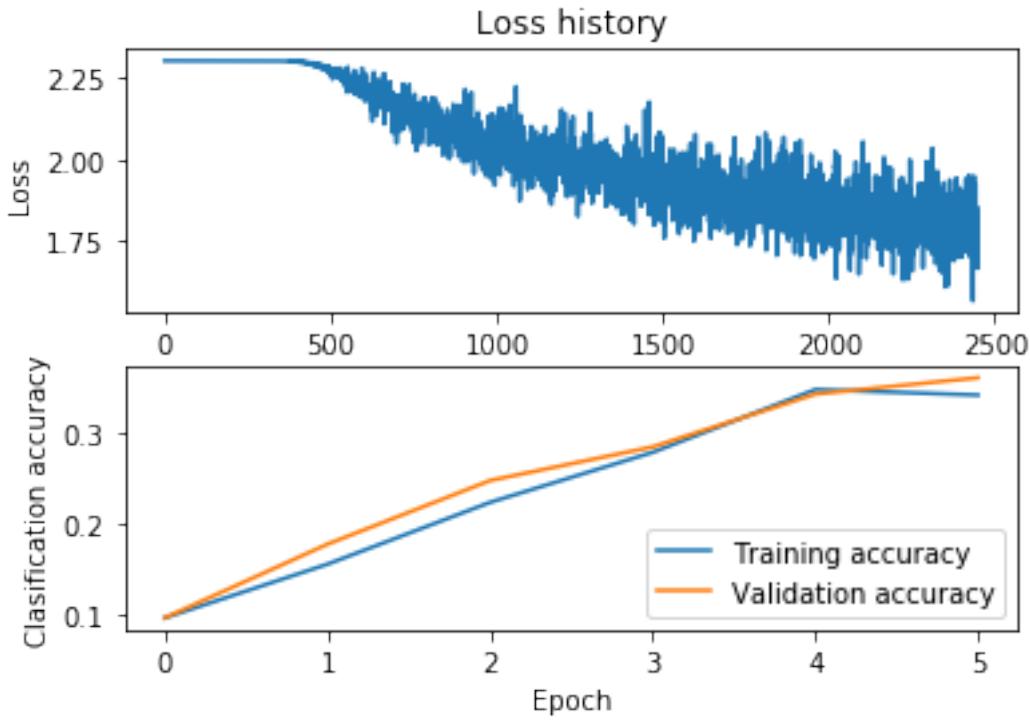
One strategy for getting insight into what's wrong is to plot the loss function and the accuracies on the training and validation sets during optimization.

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

```
[12]: # Plot the loss function and train / validation accuracies
plt.subplot(2, 1, 1)
plt.plot(loss_history)
plt.title('Loss history')
plt.xlabel('Iteration')
plt.ylabel('Loss')

plt.subplot(2, 1, 2)
plt.plot(train_acc)
plt.plot(val_acc)
plt.legend(['Training accuracy', 'Validation accuracy'], loc='lower right')
plt.xlabel('Epoch')
plt.ylabel('Classification accuracy')

[12]: Text(0, 0.5, 'Classification accuracy')
```

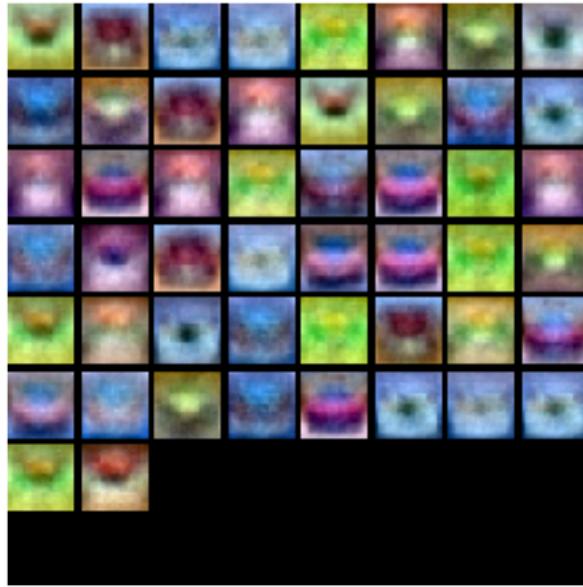


```
[13]: from cs231n.vis_utils import visualize_grid

# Visualize the weights of the network

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

show_net_weights(model)
```



9 Tune your hyperparameters

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

Tuning. Tuning the hyperparameters and developing intuition for how they affect the final performance is a large part of using Neural Networks, so we want you to get a lot of practice. Below, you should experiment with different values of the various hyperparameters, including hidden layer size, learning rate, numer of training epochs, and regularization strength. You might also consider tuning the momentum and learning rate decay parameters, but you should be able to get good performance using the default values.

Approximate results. You should be aim to achieve a classification accuracy of greater than 50% on the validation set. Our best network gets over 56% on the validation set.

Experiment: Your goal in this exercise is to get as good of a result on CIFAR-10 as you can, with a fully-connected Neural Network. For every 1% above 56% on the Test set we will award you with one extra bonus point. Feel free implement your own techniques (e.g. PCA to reduce dimensionality, or adding dropout, or adding features to the solver, etc.).

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

```
#####
```

```

# TODO: Tune hyperparameters using the validation set. Store your best trained model in best_model.
→#
# To help debug your network, it may help to use visualizations similar to the ones we used above; these visualizations will have significant qualitative differences from the ones we saw above for the poorly tuned network.
→#
# Tweaking hyperparameters by hand can be fun, but you might find it useful to write code to sweep through possible combinations of hyperparameters automatically like we did on the previous assignment.
→#
#####
# input size, hidden size, number of classes
model = init_two_layer_model(32*32*3, 1250, 10)
trainer = ClassifierTrainer()
best_model, loss_history, train_acc, val_acc = trainer.train(X_train, y_train,
                                                               X_val, y_val,
                                                               model, two_layer_net,
                                                               num_epochs=15, reg=1e-3,
                                                               momentum=0.9, batch_size=250,
                                                               learning_rate_decay=0.95,
                                                               learning_rate=1e-4, verbose=True)

# Plot the loss function and train / validation accuracies
plt.subplot(2, 1, 1)
plt.plot(loss_history)
plt.title('Loss history')
plt.xlabel('Iteration')
plt.ylabel('Loss')

plt.subplot(2, 1, 2)
plt.plot(train_acc)
plt.plot(val_acc)
plt.legend(['Training accuracy', 'Validation accuracy'], loc='lower right')
plt.xlabel('Epoch')
plt.ylabel('Classification accuracy')
#####

```

```
#  
→#
```

END OF YOUR CODE



```
#####
starting iteration 0
Finished epoch 0 / 15: cost 2.302586, train: 0.107000, val 0.077000, lr
1.000000e-04
starting iteration 10
starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
starting iteration 100
starting iteration 110
starting iteration 120
starting iteration 130
starting iteration 140
starting iteration 150
starting iteration 160
starting iteration 170
starting iteration 180
starting iteration 190
Finished epoch 1 / 15: cost 1.811845, train: 0.307000, val 0.356000, lr
9.500000e-05
starting iteration 200
starting iteration 210
starting iteration 220
starting iteration 230
starting iteration 240
starting iteration 250
starting iteration 260
starting iteration 270
starting iteration 280
starting iteration 290
starting iteration 300
starting iteration 310
starting iteration 320
starting iteration 330
starting iteration 340
starting iteration 350
starting iteration 360
starting iteration 370
starting iteration 380
```

```
starting iteration 390
Finished epoch 2 / 15: cost 1.695008, train: 0.421000, val 0.421000, lr
9.025000e-05
starting iteration 400
starting iteration 410
starting iteration 420
starting iteration 430
starting iteration 440
starting iteration 450
starting iteration 460
starting iteration 470
starting iteration 480
starting iteration 490
starting iteration 500
starting iteration 510
starting iteration 520
starting iteration 530
starting iteration 540
starting iteration 550
starting iteration 560
starting iteration 570
starting iteration 580
Finished epoch 3 / 15: cost 1.540828, train: 0.424000, val 0.455000, lr
8.573750e-05
starting iteration 590
starting iteration 600
starting iteration 610
starting iteration 620
starting iteration 630
starting iteration 640
starting iteration 650
starting iteration 660
starting iteration 670
starting iteration 680
starting iteration 690
starting iteration 700
starting iteration 710
starting iteration 720
starting iteration 730
starting iteration 740
starting iteration 750
starting iteration 760
starting iteration 770
starting iteration 780
Finished epoch 4 / 15: cost 1.319719, train: 0.510000, val 0.474000, lr
8.145062e-05
starting iteration 790
starting iteration 800
```

```
starting iteration 810
starting iteration 820
starting iteration 830
starting iteration 840
starting iteration 850
starting iteration 860
starting iteration 870
starting iteration 880
starting iteration 890
starting iteration 900
starting iteration 910
starting iteration 920
starting iteration 930
starting iteration 940
starting iteration 950
starting iteration 960
starting iteration 970
Finished epoch 5 / 15: cost 1.330010, train: 0.518000, val 0.487000, lr
7.737809e-05
starting iteration 980
starting iteration 990
starting iteration 1000
starting iteration 1010
starting iteration 1020
starting iteration 1030
starting iteration 1040
starting iteration 1050
starting iteration 1060
starting iteration 1070
starting iteration 1080
starting iteration 1090
starting iteration 1100
starting iteration 1110
starting iteration 1120
starting iteration 1130
starting iteration 1140
starting iteration 1150
starting iteration 1160
starting iteration 1170
Finished epoch 6 / 15: cost 1.505263, train: 0.534000, val 0.499000, lr
7.350919e-05
starting iteration 1180
starting iteration 1190
starting iteration 1200
starting iteration 1210
starting iteration 1220
starting iteration 1230
starting iteration 1240
```

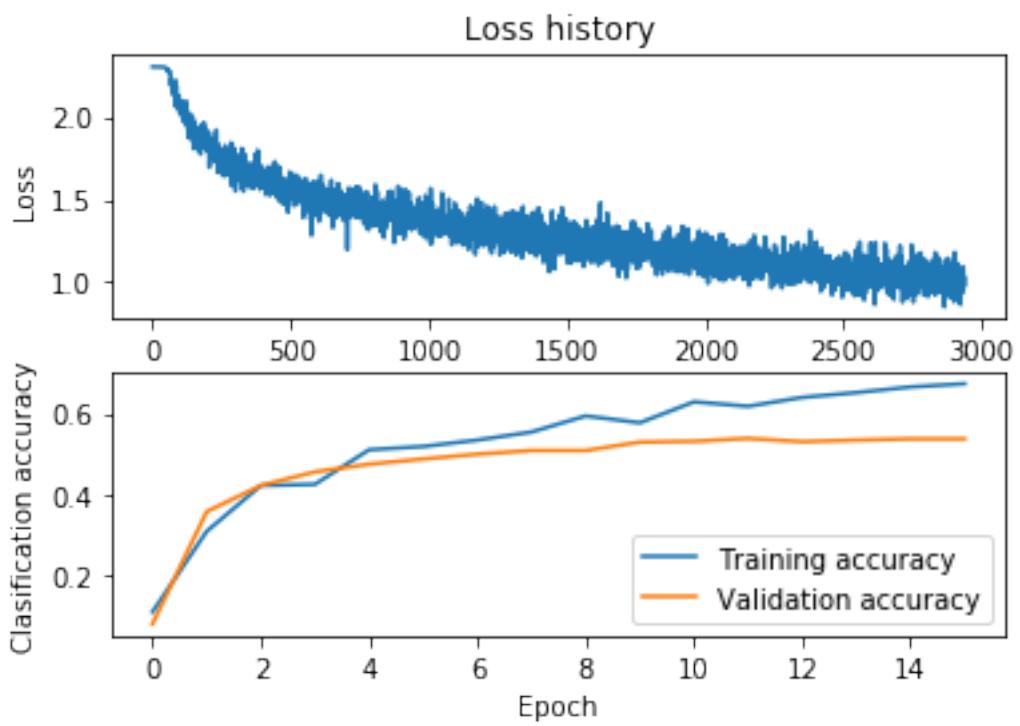
```
starting iteration 1250
starting iteration 1260
starting iteration 1270
starting iteration 1280
starting iteration 1290
starting iteration 1300
starting iteration 1310
starting iteration 1320
starting iteration 1330
starting iteration 1340
starting iteration 1350
starting iteration 1360
starting iteration 1370
Finished epoch 7 / 15: cost 1.412038, train: 0.554000, lr
6.983373e-05
starting iteration 1380
starting iteration 1390
starting iteration 1400
starting iteration 1410
starting iteration 1420
starting iteration 1430
starting iteration 1440
starting iteration 1450
starting iteration 1460
starting iteration 1470
starting iteration 1480
starting iteration 1490
starting iteration 1500
starting iteration 1510
starting iteration 1520
starting iteration 1530
starting iteration 1540
starting iteration 1550
starting iteration 1560
Finished epoch 8 / 15: cost 1.203628, train: 0.594000, lr
6.634204e-05
starting iteration 1570
starting iteration 1580
starting iteration 1590
starting iteration 1600
starting iteration 1610
starting iteration 1620
starting iteration 1630
starting iteration 1640
starting iteration 1650
starting iteration 1660
starting iteration 1670
starting iteration 1680
```

```
starting iteration 1690
starting iteration 1700
starting iteration 1710
starting iteration 1720
starting iteration 1730
starting iteration 1740
starting iteration 1750
starting iteration 1760
Finished epoch 9 / 15: cost 1.156691, train: 0.577000, val 0.529000, lr
6.302494e-05
starting iteration 1770
starting iteration 1780
starting iteration 1790
starting iteration 1800
starting iteration 1810
starting iteration 1820
starting iteration 1830
starting iteration 1840
starting iteration 1850
starting iteration 1860
starting iteration 1870
starting iteration 1880
starting iteration 1890
starting iteration 1900
starting iteration 1910
starting iteration 1920
starting iteration 1930
starting iteration 1940
starting iteration 1950
Finished epoch 10 / 15: cost 1.245950, train: 0.629000, val 0.531000, lr
5.987369e-05
starting iteration 1960
starting iteration 1970
starting iteration 1980
starting iteration 1990
starting iteration 2000
starting iteration 2010
starting iteration 2020
starting iteration 2030
starting iteration 2040
starting iteration 2050
starting iteration 2060
starting iteration 2070
starting iteration 2080
starting iteration 2090
starting iteration 2100
starting iteration 2110
starting iteration 2120
```

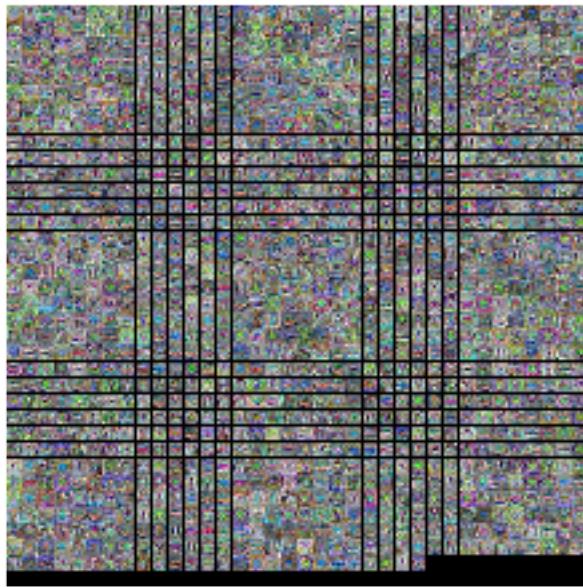
```
starting iteration 2130
starting iteration 2140
starting iteration 2150
Finished epoch 11 / 15: cost 1.067791, train: 0.618000, val 0.538000, lr
5.688001e-05
starting iteration 2160
starting iteration 2170
starting iteration 2180
starting iteration 2190
starting iteration 2200
starting iteration 2210
starting iteration 2220
starting iteration 2230
starting iteration 2240
starting iteration 2250
starting iteration 2260
starting iteration 2270
starting iteration 2280
starting iteration 2290
starting iteration 2300
starting iteration 2310
starting iteration 2320
starting iteration 2330
starting iteration 2340
starting iteration 2350
Finished epoch 12 / 15: cost 1.146589, train: 0.640000, val 0.530000, lr
5.403601e-05
starting iteration 2360
starting iteration 2370
starting iteration 2380
starting iteration 2390
starting iteration 2400
starting iteration 2410
starting iteration 2420
starting iteration 2430
starting iteration 2440
starting iteration 2450
starting iteration 2460
starting iteration 2470
starting iteration 2480
starting iteration 2490
starting iteration 2500
starting iteration 2510
starting iteration 2520
starting iteration 2530
starting iteration 2540
Finished epoch 13 / 15: cost 1.101773, train: 0.652000, val 0.534000, lr
5.133421e-05
```

```
starting iteration 2550
starting iteration 2560
starting iteration 2570
starting iteration 2580
starting iteration 2590
starting iteration 2600
starting iteration 2610
starting iteration 2620
starting iteration 2630
starting iteration 2640
starting iteration 2650
starting iteration 2660
starting iteration 2670
starting iteration 2680
starting iteration 2690
starting iteration 2700
starting iteration 2710
starting iteration 2720
starting iteration 2730
starting iteration 2740
Finished epoch 14 / 15: cost 1.083652, train: 0.666000, val 0.537000, lr
4.876750e-05
starting iteration 2750
starting iteration 2760
starting iteration 2770
starting iteration 2780
starting iteration 2790
starting iteration 2800
starting iteration 2810
starting iteration 2820
starting iteration 2830
starting iteration 2840
starting iteration 2850
starting iteration 2860
starting iteration 2870
starting iteration 2880
starting iteration 2890
starting iteration 2900
starting iteration 2910
starting iteration 2920
starting iteration 2930
Finished epoch 15 / 15: cost 1.023649, train: 0.674000, val 0.537000, lr
4.632912e-05
finished optimization. best validation accuracy: 0.538000
```

[14]: Text(0, 0.5, 'Clasification accuracy')



```
[15]: # visualize the weights
show_net_weights(best_model)
```



10 Run on the test set

When you are done experimenting, you should evaluate your final trained network on the test set.

```
[16]: scores_test = two_layer_net(X_test, best_model)
      print('Test accuracy: ', np.mean(np.argmax(scores_test, axis=1) == y_test))
```

Test accuracy: 0.523

layers

February 12, 2020

1 Modular neural nets

In the previous exercise, we computed the loss and gradient for a two-layer neural network in a single monolithic function. This isn't very difficult for a small two-layer network, but would be tedious and error-prone for larger networks. Ideally we want to build networks using a more modular design so that we can snap together different types of layers and loss functions in order to quickly experiment with different architectures.

In this exercise we will implement this approach, and develop a number of different layer types in isolation that can then be easily plugged together. For each layer we will implement forward and backward functions. The forward function will receive data, weights, and other parameters, and will return both an output and a cache object that stores data needed for the backward pass. The backward function will receive upstream derivatives and the cache object, and will return gradients with respect to the data and all of the weights. This will allow us to write code that looks like this:

```
def two_layer_net(X, W1, b1, W2, b2, reg):
    # Forward pass; compute scores
    s1, fc1_cache = affine_forward(X, W1, b1)
    a1, relu_cache = relu_forward(s1)
    scores, fc2_cache = affine_forward(a1, W2, b2)

    # Loss functions return data loss and gradients on scores
    data_loss, dscores = svm_loss(scores, y)

    # Compute backward pass
    da1, dW2, db2 = affine_backward(dscores, fc2_cache)
    ds1 = relu_backward(da1, relu_cache)
    dX, dW1, db1 = affine_backward(ds1, fc1_cache)

    # A real network would add regularization here

    # Return loss and gradients
    return loss, dW1, db1, dW2, db2
```

[1]: # As usual, a bit of setup

```
import numpy as np
```

```

import matplotlib.pyplot as plt
from cs231n.gradient_check import eval_numerical_gradient, u
→eval_numerical_gradient
from cs231n.layers import *

%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))))

```

2 Affine layer: forward

Open the file `cs231n/layers.py` and implement the `affine_forward` function.

Once you are done we will test your can test your implementation by running the following:

[2]: # 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), u
→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: ', rel_error(out, correct_out))

```

```
Testing affine_forward function:  
difference: 9.769847728806635e-10
```

3 Affine layer: backward

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

```
[3]: # 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, u  
→dout)  
dw_num = eval_numerical_gradient_array(lambda w: affine_forward(x, w, b)[0], w, u  
→dout)  
db_num = eval_numerical_gradient_array(lambda b: affine_forward(x, w, b)[0], b, u  
→dout)  
  
, cache = affine_forward(x, w, b)  
dx, dw, db = affine_backward(dout, cache)  
  
# The error should be less than 1e-10  
print('Testing affine_backward function:')  
print('dx error: ', rel_error(dx_num, dx))  
print('dw error: ', rel_error(dw_num, dw))  
print('db error: ', rel_error(db_num, db))
```

```
Testing affine_backward function:  
dx error: 7.30632070020297e-10  
dw error: 7.34672694737665e-11  
db error: 1.3037618198032586e-11
```

4 ReLU layer: forward

Implement the `relu_forward` function and test your implementation by running the following:

```
[4]: # 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: ', rel_error(out, correct_out))
```

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

5 ReLU layer: backward

Implement the `relu_backward` function and test your implementation using numeric gradient checking:

```
[5]: 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: ', rel_error(dx_num, dx))
```

```
Testing relu_backward function:
dx error: 3.2756390066177367e-12
```

6 Loss layers: Softmax and SVM

You implemented these loss functions in the last assignment, so we'll give them to you for free here. It's still a good idea to test them to make sure they work correctly.

```
[6]: 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 sum_loss function. Loss should be around 9 and dx error should be 1e-9
print('Testing svm_loss:')
print('loss: ', loss)
print('dx error: ', 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: ', loss)
print('dx error: ', rel_error(dx_num, dx))
```

Testing svm_loss:
loss: 8.999118178169411
dx error: 1.4021566006651672e-09

Testing softmax_loss:
loss: 2.302497336692487
dx error: 7.40973039369766e-09

7 Convolution layer: forward naive

We are now ready to implement the forward pass for a convolutional layer. Implement the function `conv_forward_naive` in the file `cs231n/layers.py`.

You don't have to worry too much about efficiency at this point; just write the code in whatever way you find most clear.

You can test your implementation by running the following:

```
[12]: x_shape = (2, 3, 4, 4)
w_shape = (3, 3, 4, 4)
x = np.linspace(-0.1, 0.5, num=np.prod(x_shape)).reshape(x_shape)
w = np.linspace(-0.2, 0.3, num=np.prod(w_shape)).reshape(w_shape)
b = np.linspace(-0.1, 0.2, num=3)

conv_param = {'stride': 2, 'pad': 1}
out, _ = conv_forward_naive(x, w, b, conv_param)
correct_out = np.array([[[[-0.08759809, -0.10987781],
                         [-0.18387192, -0.2109216]],
                        [[ 0.21027089,  0.21661097],
                         [ 0.22847626,  0.23004637]],
                        [[ 0.50813986,  0.54309974],
                         [ 0.64082444,  0.67101435]]],
                       [[[ -0.98053589, -1.03143541],
                         [-1.19128892, -1.24695841]],
                        [[ 0.69108355,  0.66880383],
                         [ 0.59480972,  0.56776003]],
                        [[ 2.36270298,  2.36904306],
                         [ 2.38090835,  2.38247847]]]]])

# Compare your output to ours; difference should be around 1e-8
print('Testing conv_forward_naive')
print('difference: ', rel_error(out, correct_out))
```

```
Testing conv_forward_naive
difference: 2.2121476575931688e-08
```

8 Aside: Image processing via convolutions

As fun way to both check your implementation and gain a better understanding of the type of operation that convolutional layers can perform, we will set up an input containing two images and manually set up filters that perform common image processing operations (grayscale conversion and edge detection). The convolution forward pass will apply these operations to each of the input images. We can then visualize the results as a sanity check.

```
[13]: from scipy.misc import imread, imresize

kitten, puppy = imread('kitten.jpg'), imread('puppy.jpg')
# kitten is wide, and puppy is already square
d = kitten.shape[1] - kitten.shape[0]
kitten_cropped = kitten[:, d//2:-d//2, :]

img_size = 200 # Make this smaller if it runs too slow
x = np.zeros((2, 3, img_size, img_size))
x[0, :, :, :] = imresize(puppy, (img_size, img_size)).transpose((2, 0, 1))
x[1, :, :, :] = imresize(kitten_cropped, (img_size, img_size)).transpose((2, 0, 1))

# Set up a convolutional weights holding 2 filters, each 3x3
w = np.zeros((2, 3, 3, 3))

# The first filter converts the image to grayscale.
# Set up the red, green, and blue channels of the filter.
w[0, 0, :, :] = [[0, 0, 0], [0, 0.3, 0], [0, 0, 0]]
w[0, 1, :, :] = [[0, 0, 0], [0, 0.6, 0], [0, 0, 0]]
w[0, 2, :, :] = [[0, 0, 0], [0, 0.1, 0], [0, 0, 0]]

# Second filter detects horizontal edges in the blue channel.
w[1, 2, :, :] = [[1, 2, 1], [0, 0, 0], [-1, -2, -1]]

# Vector of biases. We don't need any bias for the grayscale
# filter, but for the edge detection filter we want to add 128
# to each output so that nothing is negative.
b = np.array([0, 128])

# Compute the result of convolving each input in x with each filter in w,
# offsetting by b, and storing the results in out.
out, _ = conv_forward_naive(x, w, b, {'stride': 1, 'pad': 1})

def imshow_noax(img, normalize=True):
    """ Tiny helper to show images as uint8 and remove axis labels """

```

```

if normalize:
    img_max, img_min = np.max(img), np.min(img)
    img = 255.0 * (img - img_min) / (img_max - img_min)
plt.imshow(img.astype('uint8'))
plt.gca().axis('off')

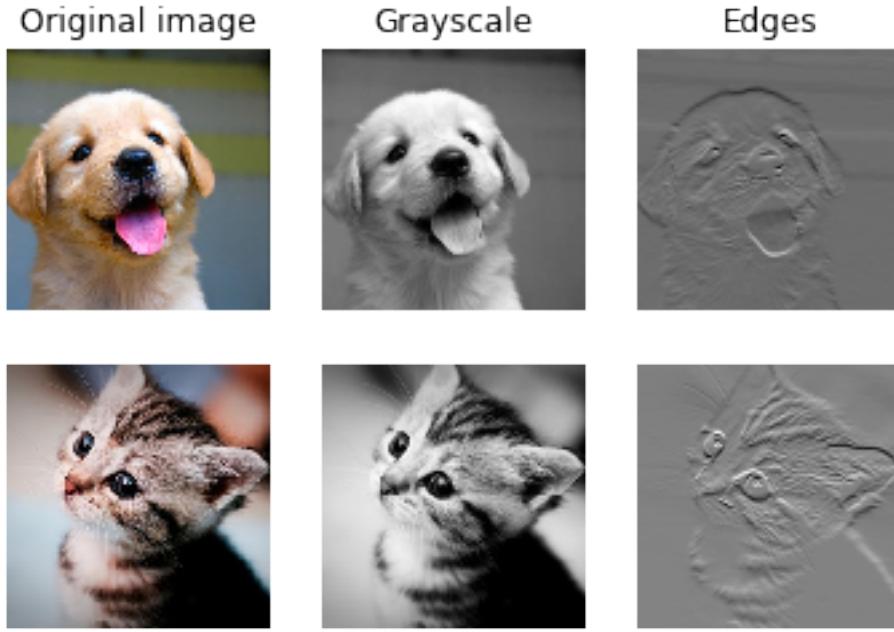
# Show the original images and the results of the conv operation
plt.subplot(2, 3, 1)
imshow_noax(puppy, normalize=False)
plt.title('Original image')
plt.subplot(2, 3, 2)
imshow_noax(out[0, 0])
plt.title('Grayscale')
plt.subplot(2, 3, 3)
imshow_noax(out[0, 1])
plt.title('Edges')
plt.subplot(2, 3, 4)
imshow_noax(kitten_cropped, normalize=False)
plt.subplot(2, 3, 5)
imshow_noax(out[1, 0])
plt.subplot(2, 3, 6)
imshow_noax(out[1, 1])
plt.show()

```

```

C:\Users\david\Miniconda3\envs\cs7643\lib\site-packages\ipykernel_launcher.py:3:
DeprecationWarning: `imread` is deprecated!
`imread` is deprecated in SciPy 1.0.0, and will be removed in 1.2.0.
Use ``imageio.imread`` instead.
    This is separate from the ipykernel package so we can avoid doing imports
until
C:\Users\david\Miniconda3\envs\cs7643\lib\site-
packages\ipykernel_launcher.py:10: DeprecationWarning: `imresize` is deprecated!
`imresize` is deprecated in SciPy 1.0.0, and will be removed in 1.2.0.
Use ``skimage.transform.resize`` instead.
    # Remove the CWD from sys.path while we load stuff.
C:\Users\david\Miniconda3\envs\cs7643\lib\site-
packages\ipykernel_launcher.py:11: DeprecationWarning: `imresize` is deprecated!
`imresize` is deprecated in SciPy 1.0.0, and will be removed in 1.2.0.
Use ``skimage.transform.resize`` instead.
    # This is added back by InteractiveShellApp.init_path()

```



9 Convolution layer: backward naive

Next you need to implement the function `conv_backward_naive` in the file `cs231n/layers.py`. As usual, we will check your implementation with numeric gradient checking.

```
[14]: x = np.random.randn(4, 3, 5, 5)
w = np.random.randn(2, 3, 3, 3)
b = np.random.randn(2,)
dout = np.random.randn(4, 2, 5, 5)
conv_param = {'stride': 1, 'pad': 1}

dx_num = eval_numerical_gradient_array(lambda x: conv_forward_naive(x, w, b, ↴conv_param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_forward_naive(x, w, b, ↴conv_param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_forward_naive(x, w, b, ↴conv_param)[0], b, dout)

out, cache = conv_forward_naive(x, w, b, conv_param)
dx, dw, db = conv_backward_naive(dout, cache)

# Your errors should be around 1e-9'
print('Testing conv_backward_naive function')
print('dx error: ', rel_error(dx, dx_num))
```

```

print('dw error: ', rel_error(dw, dw_num))
print('db error: ', rel_error(db, db_num))

```

Testing conv_backward_naive function
dx error: 1.0
dw error: 1.0
db error: 7.338357481916887e-13

10 Max pooling layer: forward naive

The last layer we need for a basic convolutional neural network is the max pooling layer. First implement the forward pass in the function `max_pool_forward_naive` in the file `cs231n/layers.py`.

```

[15]: x_shape = (2, 3, 4, 4)
x = np.linspace(-0.3, 0.4, num=np.prod(x_shape)).reshape(x_shape)
pool_param = {'pool_width': 2, 'pool_height': 2, 'stride': 2}

out, _ = max_pool_forward_naive(x, pool_param)

correct_out = np.array([[[[-0.26315789, -0.24842105],
                         [-0.20421053, -0.18947368]],
                        [[-0.14526316, -0.13052632],
                         [-0.08631579, -0.07157895]],
                        [[-0.02736842, -0.01263158],
                         [ 0.03157895,  0.04631579]]],
                       [[[ 0.09052632,  0.10526316],
                         [ 0.14947368,  0.16421053]],
                        [[ 0.20842105,  0.22315789],
                         [ 0.26736842,  0.28210526]],
                        [[ 0.32631579,  0.34105263],
                         [ 0.38526316,  0.4        ]]]])

# Compare your output with ours. Difference should be around 1e-8.
print('Testing max_pool_forward_naive function:')
print('difference: ', rel_error(out, correct_out))

```

Testing max_pool_forward_naive function:
difference: 4.1666665157267834e-08

Max pooling layer: backward naive Implement the backward pass for a max pooling layer in the function `max_pool_backward_naive` in the file `cs231n/layers.py`. As always we check the correctness of the backward pass using numerical gradient checking.

```

[16]: x = np.random.randn(3, 2, 8, 8)
dout = np.random.randn(3, 2, 4, 4)
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

```

```

dx_num = eval_numerical_gradient_array(lambda x: max_pool_forward_naive(x,
    pool_param)[0], x, dout)

out, cache = max_pool_forward_naive(x, pool_param)
dx = max_pool_backward_naive(dout, cache)

# Your error should be around 1e-12
print('Testing max_pool_backward_naive function:')
print('dx error: ', rel_error(dx, dx_num))

```

Testing max_pool_backward_naive function:
dx error: 3.2756392551857143e-12

11 Fast layers

Making convolution and pooling layers fast can be challenging. To spare you the pain, we've provided fast implementations of the forward and backward passes for convolution and pooling layers in the file `cs231n/fast_layers.py`.

The fast convolution implementation depends on a Cython extension; to compile it you need to run the following from the `cs231n` directory:

```
python setup.py build_ext --inplace
```

The API for the fast versions of the convolution and pooling layers is exactly the same as the naive versions that you implemented above: the forward pass receives data, weights, and parameters and produces outputs and a cache object; the backward pass receives upstream derivatives and the cache object and produces gradients with respect to the data and weights.

NOTE: The fast implementation for pooling will only perform optimally if the pooling regions are non-overlapping and tile the input. If these conditions are not met then the fast pooling implementation will not be much faster than the naive implementation.

You can compare the performance of the naive and fast versions of these layers by running the following:

```
[18]: from cs231n.fast_layers import conv_forward_fast, conv_backward_fast
from time import time

x = np.random.randn(100, 3, 31, 31)
w = np.random.randn(25, 3, 3, 3)
b = np.random.randn(25,)
dout = np.random.randn(100, 25, 16, 16)
conv_param = {'stride': 2, 'pad': 1}

t0 = time()
out_naive, cache_naive = conv_forward_naive(x, w, b, conv_param)
t1 = time()
out_fast, cache_fast = conv_forward_fast(x, w, b, conv_param)
t2 = time()
```

```

print('Testing conv_forward_fast:')
print('Naive: %fs' % (t1 - t0))
print('Fast: %fs' % (t2 - t1))
print('Speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('Difference: ', rel_error(out_naive, out_fast))

t0 = time()
dx_naive, dw_naive, db_naive = conv_backward_naive(dout, cache_naive)
t1 = time()
dx_fast, dw_fast, db_fast = conv_backward_fast(dout, cache_fast)
t2 = time()

print('\nTesting conv_backward_fast:')
print('Naive: %fs' % (t1 - t0))
print('Fast: %fs' % (t2 - t1))
print('Speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('dx difference: ', rel_error(dx_naive, dx_fast))
print('dw difference: ', rel_error(dw_naive, dw_fast))
print('db difference: ', rel_error(db_naive, db_fast))

```

```

Testing conv_forward_fast:
Naive: 0.055973s
Fast: 0.020001s
Speedup: 2.798489x
Difference:  0.0

```

□
↳-----

```

IndexError                                     Traceback (most recent call □
↳last)

<ipython-input-18-d0e7ac55f1f2> in <module>
    21
    22 t0 = time()
---> 23 dx_naive, dw_naive, db_naive = conv_backward_naive(dout, cache_naive)
    24 t1 = time()
    25 dx_fast, dw_fast, db_fast = conv_backward_fast(dout, cache_fast)
```

□
↳~\Documents\GitHub\CS7643_DeepLearning\ps1\hw1_q8_starter\assignment\1_cs231n\cs231n\layers.py in conv_backward_naive(dout, cache)
 194 for c in range(0, padded_x.shape[3] - filter_width, □
 195 curr_conv = padded_x[img, :, r:r+filter_height, □
 196 c:c+filter_width]

```

--> 196 dw[curr_filter] += curr_conv * dout[img, □
curr_filter, r, c]
197 dx_pad[img, :, r:r+filter_height, c:
curr_filter_width] += w[curr_filter] * dout[img, curr_filter, r, c]
198 # take the values of dx_pad and remove the padding

```

IndexError: index 16 is out of bounds for axis 3 with size 16

[19]: `from cs231n.fast_layers import max_pool_forward_fast, max_pool_backward_fast`

```

x = np.random.randn(100, 3, 32, 32)
dout = np.random.randn(100, 3, 16, 16)
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

t0 = time()
out_naive, cache_naive = max_pool_forward_naive(x, pool_param)
t1 = time()
out_fast, cache_fast = max_pool_forward_fast(x, pool_param)
t2 = time()

print('Testing pool_forward_fast:')
print('Naive: %fs' % (t1 - t0))
print('fast: %fs' % (t2 - t1))
print('speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('difference: ', rel_error(out_naive, out_fast))

t0 = time()
dx_naive = max_pool_backward_naive(dout, cache_naive)
t1 = time()
dx_fast = max_pool_backward_fast(dout, cache_fast)
t2 = time()

print('\nTesting pool_backward_fast:')
print('Naive: %fs' % (t1 - t0))
print('fast: %fs' % ((t1 - t0) / (t2 - t1)))
print('dx difference: ', rel_error(dx_naive, dx_fast))

```

```

Testing pool_forward_fast:
Naive: 0.109343s
fast: 0.004004s
speedup: 27.308324x
difference:  0.0

```

```

Testing pool_backward_fast:
Naive: 0.316000s

```

```
speedup: 39.501744x
dx difference: 0.0
```

12 Sandwich layers

There are a couple common layer "sandwiches" that frequently appear in ConvNets. For example convolutional layers are frequently followed by ReLU and pooling, and affine layers are frequently followed by ReLU. To make it more convenient to use these common patterns, we have defined several convenience layers in the file `cs231n/layer_utils.py`. Lets grad-check them to make sure that they work correctly:

```
[20]: from cs231n.layer_utils import conv_relu_pool_forward, conv_relu_pool_backward

x = np.random.randn(2, 3, 16, 16)
w = np.random.randn(3, 3, 3, 3)
b = np.random.randn(3,)
dout = np.random.randn(2, 3, 8, 8)
conv_param = {'stride': 1, 'pad': 1}
pool_param = {'pool_height': 2, 'pool_width': 2, 'stride': 2}

out, cache = conv_relu_pool_forward(x, w, b, conv_param, pool_param)
dx, dw, db = conv_relu_pool_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: conv_relu_pool_forward(x, w, b, conv_param, pool_param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_relu_pool_forward(x, w, b, conv_param, pool_param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_relu_pool_forward(x, w, b, conv_param, pool_param)[0], b, dout)

print('Testing conv_relu_pool_forward:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
Testing conv_relu_pool_forward:
dx error: 7.62487651942435e-09
dw error: 2.6197688241937565e-09
db error: 2.6690956577210883e-11
```

```
[21]: from cs231n.layer_utils import conv_relu_forward, conv_relu_backward

x = np.random.randn(2, 3, 8, 8)
w = np.random.randn(3, 3, 3, 3)
b = np.random.randn(3,)
dout = np.random.randn(2, 3, 8, 8)
conv_param = {'stride': 1, 'pad': 1}
```

```

out, cache = conv_relu_forward(x, w, b, conv_param)
dx, dw, db = conv_relu_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: conv_relu_forward(x, w, b, conv_param)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: conv_relu_forward(x, w, b, conv_param)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: conv_relu_forward(x, w, b, conv_param)[0], b, dout)

print('Testing conv_relu_forward:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))

```

```

Testing conv_relu_forward:
dx error:  5.854747818684639e-09
dw error:  4.4562615648434886e-10
db error:  2.628715883419431e-11

```

[22]:

```

from cs231n.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:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))

```

```

Testing affine_relu_forward:
dx error:  1.7472216289582935e-10
dw error:  1.052184901093154e-09
db error:  7.82669531646424e-12

```

convnet

February 12, 2020

1 Train a ConvNet!

We now have a generic solver and a bunch of modularized layers. It's time to put it all together, and train a ConvNet to recognize the classes in CIFAR-10. In this notebook we will walk you through training a simple two-layer ConvNet and then set you free to build the best net that you can to perform well on CIFAR-10.

Open up the file `cs231n/classifiers/convnet.py`; you will see that the `two_layer_convnet` function computes the loss and gradients for a two-layer ConvNet. Note that this function uses the "sandwich" layers defined in `cs231n/layer_utils.py`.

[1]: # As usual, a bit of setup

```
import numpy as np
import matplotlib.pyplot as plt
from cs231n.classifier_trainer import ClassifierTrainer
from cs231n.gradient_check import eval_numerical_gradient
from cs231n.classifiers.convnet import *

%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))))
```

[2]: 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 = 'cs231n/datasets/cifar-10-batches-py'
X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

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

# Transpose so that channels come first
X_train = X_train.transpose(0, 3, 1, 2).copy()
X_val = X_val.transpose(0, 3, 1, 2).copy()
X_test = X_test.transpose(0, 3, 1, 2).copy()

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

```

2 Sanity check loss

After you build a new network, one of the first things you should do is sanity check the loss. When we use the softmax loss, we expect the loss for random weights (and no regularization) to be about $\log(C)$ for C classes. When we add regularization this should go up.

```
[3]: model = init_two_layer_convnet()

X = np.random.randn(100, 3, 32, 32)
y = np.random.randint(10, size=100)

loss, _ = two_layer_convnet(X, model, y, reg=0)

# Sanity check: Loss should be about log(10) = 2.3026
print('Sanity check loss (no regularization): ', loss)

# Sanity check: Loss should go up when you add regularization
loss, _ = two_layer_convnet(X, model, y, reg=1)
print('Sanity check loss (with regularization): ', loss)
```

Sanity check loss (no regularization): 2.3025661192636675
Sanity check loss (with regularization): 2.345170379324019

3 Gradient check

After the loss looks reasonable, you should always use numeric gradient checking to make sure that your backward pass is correct. When you use numeric gradient checking you should use a small amount of artificial data and a small number of neurons at each layer.

```
[4]: num_inputs = 2
input_shape = (3, 16, 16)
reg = 0.0
num_classes = 10
X = np.random.randn(num_inputs, *input_shape)
y = np.random.randint(num_classes, size=num_inputs)

model = init_two_layer_convnet(num_filters=3, filter_size=3, input_shape=input_shape)
loss, grads = two_layer_convnet(X, model, y)
for param_name in sorted(grads):
    f = lambda _: two_layer_convnet(X, model, y)[0]
    param_grad_num = eval_numerical_gradient(f, model[param_name], verbose=False, h=1e-6)
    e = rel_error(param_grad_num, grads[param_name])
    print('%s max relative error: %e' % (param_name, rel_error(param_grad_num, grads[param_name])))
```

W1 max relative error: 3.147468e-02
W2 max relative error: 2.327145e-04

```
b1 max relative error: 2.032773e-07
b2 max relative error: 9.228404e-10
```

4 Overfit small data

A nice trick is to train your model with just a few training samples. You should be able to overfit small datasets, which will result in very high training accuracy and comparatively low validation accuracy.

[5]: # Use a two-layer ConvNet to overfit 50 training examples.

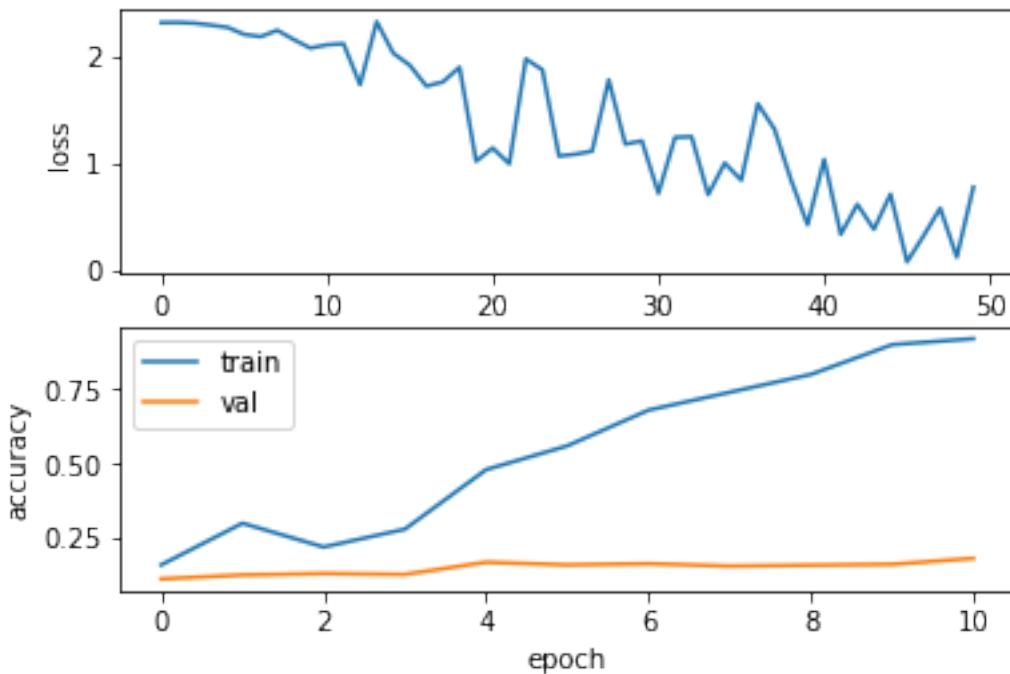
```
model = init_two_layer_convnet()
trainer = ClassifierTrainer()
best_model, loss_history, train_acc_history, val_acc_history = trainer.train(
    X_train[:50], y_train[:50], X_val, y_val, model, two_layer_convnet,
    reg=0.001, momentum=0.9, learning_rate=0.0001, batch_size=10, ↴
    num_epochs=10,
    verbose=True)
```

```
starting iteration 0
Finished epoch 0 / 10: cost 2.323287, train: 0.160000, val 0.113000, lr
1.000000e-04
Finished epoch 1 / 10: cost 2.278602, train: 0.300000, val 0.126000, lr
9.500000e-05
Finished epoch 2 / 10: cost 2.083403, train: 0.220000, val 0.131000, lr
9.025000e-05
starting iteration 10
Finished epoch 3 / 10: cost 2.035731, train: 0.280000, val 0.128000, lr
8.573750e-05
Finished epoch 4 / 10: cost 1.018502, train: 0.480000, val 0.169000, lr
8.145062e-05
starting iteration 20
Finished epoch 5 / 10: cost 1.067086, train: 0.560000, val 0.160000, lr
7.737809e-05
Finished epoch 6 / 10: cost 1.210232, train: 0.680000, val 0.164000, lr
7.350919e-05
starting iteration 30
Finished epoch 7 / 10: cost 1.001124, train: 0.740000, val 0.156000, lr
6.983373e-05
Finished epoch 8 / 10: cost 0.420666, train: 0.800000, val 0.159000, lr
6.634204e-05
starting iteration 40
Finished epoch 9 / 10: cost 0.708382, train: 0.900000, val 0.162000, lr
6.302494e-05
Finished epoch 10 / 10: cost 0.774000, train: 0.920000, val 0.182000, lr
5.987369e-05
finished optimization. best validation accuracy: 0.182000
```

Plotting the loss, training accuracy, and validation accuracy should show clear overfitting:

```
[6]: plt.subplot(2, 1, 1)
plt.plot(loss_history)
plt.xlabel('iteration')
plt.ylabel('loss')

plt.subplot(2, 1, 2)
plt.plot(train_acc_history)
plt.plot(val_acc_history)
plt.legend(['train', 'val'], loc='upper left')
plt.xlabel('epoch')
plt.ylabel('accuracy')
plt.show()
```



5 Train the net

Once the above works, training the net is the next thing to try. You can set the `acc_frequency` parameter to change the frequency at which the training and validation set accuracies are tested. If your parameters are set properly, you should see the training and validation accuracy start to improve within a hundred iterations, and you should be able to train a reasonable model with just one epoch.

Using the parameters below you should be able to get around 50% accuracy on the validation set.

```
[7]: model = init_two_layer_convnet(filter_size=5)
trainer = ClassifierTrainer()
best_model, loss_history, train_acc_history, val_acc_history = trainer.train(
    X_train, y_train, X_val, y_val, model, two_layer_convnet,
    reg=0.001, momentum=0.9, learning_rate=0.0001, batch_size=250, ↴
    ↪num_epochs=5,
    learning_rate_decay=0.95, verbose=True)

# Plot the loss function and train / validation accuracies
plt.subplot(2, 1, 1)
plt.plot(loss_history)
plt.title('Loss history')
plt.xlabel('Iteration')
plt.ylabel('Loss')

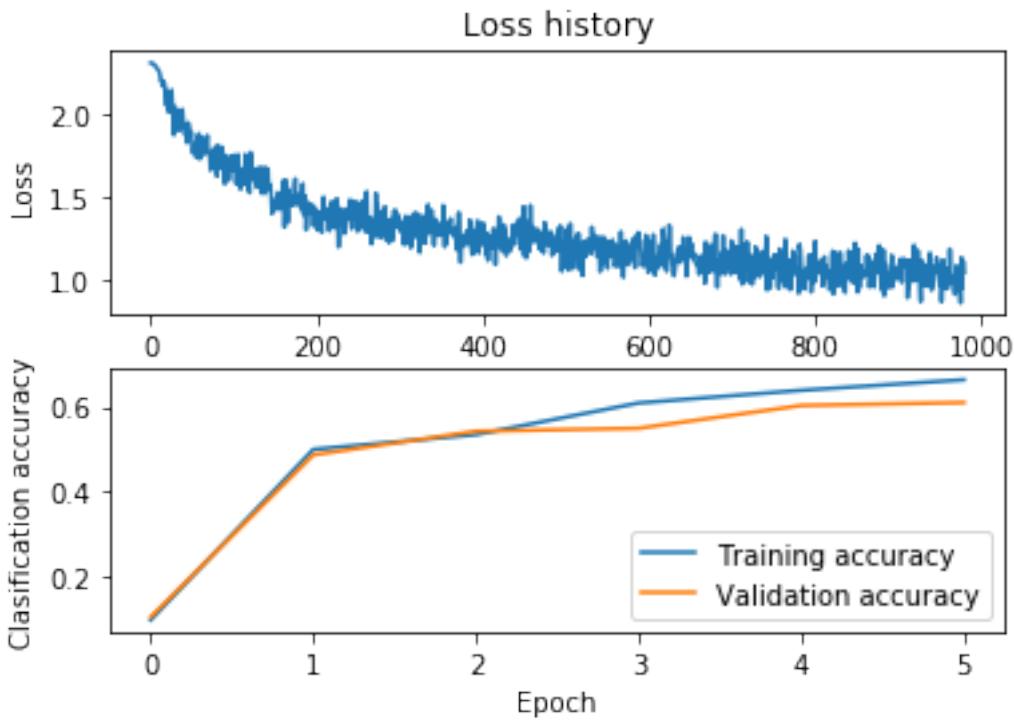
plt.subplot(2, 1, 2)
plt.plot(train_acc_history)
plt.plot(val_acc_history)
plt.legend(['Training accuracy', 'Validation accuracy'], loc='lower right')
plt.xlabel('Epoch')
plt.ylabel('Classification accuracy')
```

```
starting iteration 0
Finished epoch 0 / 5: cost 2.304040, train: 0.095000, val 0.102000, lr
1.000000e-04
starting iteration 10
starting iteration 20
starting iteration 30
starting iteration 40
starting iteration 50
starting iteration 60
starting iteration 70
starting iteration 80
starting iteration 90
starting iteration 100
starting iteration 110
starting iteration 120
starting iteration 130
starting iteration 140
starting iteration 150
starting iteration 160
starting iteration 170
starting iteration 180
starting iteration 190
Finished epoch 1 / 5: cost 1.418221, train: 0.499000, val 0.486000, lr
9.500000e-05
starting iteration 200
```

```
starting iteration 210
starting iteration 220
starting iteration 230
starting iteration 240
starting iteration 250
starting iteration 260
starting iteration 270
starting iteration 280
starting iteration 290
starting iteration 300
starting iteration 310
starting iteration 320
starting iteration 330
starting iteration 340
starting iteration 350
starting iteration 360
starting iteration 370
starting iteration 380
starting iteration 390
Finished epoch 2 / 5: cost 1.248672, train: 0.534000, val 0.542000, lr
9.025000e-05
starting iteration 400
starting iteration 410
starting iteration 420
starting iteration 430
starting iteration 440
starting iteration 450
starting iteration 460
starting iteration 470
starting iteration 480
starting iteration 490
starting iteration 500
starting iteration 510
starting iteration 520
starting iteration 530
starting iteration 540
starting iteration 550
starting iteration 560
starting iteration 570
starting iteration 580
Finished epoch 3 / 5: cost 1.284732, train: 0.609000, val 0.549000, lr
8.573750e-05
starting iteration 590
starting iteration 600
starting iteration 610
starting iteration 620
starting iteration 630
starting iteration 640
```

```
starting iteration 650
starting iteration 660
starting iteration 670
starting iteration 680
starting iteration 690
starting iteration 700
starting iteration 710
starting iteration 720
starting iteration 730
starting iteration 740
starting iteration 750
starting iteration 760
starting iteration 770
starting iteration 780
Finished epoch 4 / 5: cost 1.069853, train: 0.639000, val 0.603000, lr
8.145062e-05
starting iteration 790
starting iteration 800
starting iteration 810
starting iteration 820
starting iteration 830
starting iteration 840
starting iteration 850
starting iteration 860
starting iteration 870
starting iteration 880
starting iteration 890
starting iteration 900
starting iteration 910
starting iteration 920
starting iteration 930
starting iteration 940
starting iteration 950
starting iteration 960
starting iteration 970
Finished epoch 5 / 5: cost 1.045888, train: 0.664000, val 0.610000, lr
7.737809e-05
finished optimization. best validation accuracy: 0.610000
```

[7]: Text(0, 0.5, 'Clasification accuracy')



6 Visualize weights

We can visualize the convolutional weights from the first layer. If everything worked properly, these will usually be edges and blobs of various colors and orientations.

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

grid = visualize_grid(best_model['W1'].transpose(0, 2, 3, 1))
plt.imshow(grid.astype('uint8'))
```

[8]: <matplotlib.image.AxesImage at 0x26033b60208>

