1. $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$ ...... (1)

$$\underset{w}{\operatorname{argmin}}\ f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2}\|w - w^{(t)}\|^2 \quad (4)$$

Take derivative of (4) w.r.t. w, & set to 0.

$$0 = \cancel{\frac{\partial f(\vec{w}^{(t)})}{\partial \vec{w}}}^{\,0} + \frac{\partial \langle \vec{w} - \vec{w}^{(t)}, \nabla f(\vec{w}^{(t)}) \rangle}{\partial \vec{w}} + \frac{\partial}{\partial \vec{w}} \frac{\lambda}{2} \|\vec{w} - \vec{w}^{(t)}\|^2$$

$f(\vec{w}^{(t)})$ is not dependent of $\vec{w}$

$$0 = \nabla f(\vec{w}^{(t)})^T + \frac{\lambda}{\cancel{2}} \cdot \cancel{2} \cdot (\vec{w} - \vec{w}^{(t)})^T$$

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

$$w = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)}) \quad (A)$$

comparing (1) and (A),

$$\eta = \frac{1}{\lambda}. \quad (*)$$

The update rule moves toward the opposite direction of the gradient (page 100 of textbook). And when the gradient is zero, w is optimized.

η and λ have an inverse relationship. (*)

2| (eq5) $\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \le \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$

$\rightarrow \left[\sum \left(\langle w^{(t)}, v_t \rangle - \langle w^*, v_t \rangle\right)\right] 2\eta \le \|w^*\|^2 + \eta^2 \sum_{t=1}^{T} \|v_t\|^2$

$= w^{(t+1)} + \eta v_t$ (update rule)

$LHS = \left[\sum_{t=1}^{T} \langle w^{(t+1)}, v_t \rangle + \eta \langle v_t, v_t \rangle - \langle w^*, v_t \rangle\right] 2\eta$

$= 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^*, v_t \rangle + 2\eta^2 \sum_{t=1}^{T} \|v_t\|^2$

$\le RHS = \|w^*\|^2 + \eta^2 \sum_{t=1}^{T} \|v_t\|^2$

move to LHS

$RHS = \|w^*\|^2 \ge 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^*, v_t \rangle + \eta^2 \sum_{t=1}^{T} \|v_t\|^2 = LHS$

Add $\sum_{t=1}^{T} \|w^{(t+1)} - w^*\|^2$ to both sides

$\|w^*\|^2 + \sum_{t=1}^{T} \|w^{(t+1)} - w^*\|^2 \ge \sum_{t=1}^{T} \|w^{(t+1)} - w^*\|^2 + 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^*, v_t \rangle$

$+ \eta^2 \sum_{t=1}^{T} \|v_t\|^2$

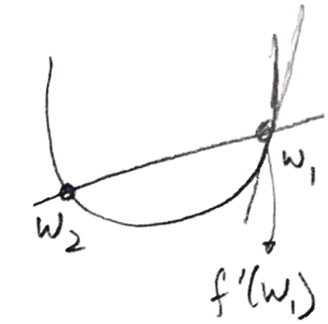
$\|w^*\|^2 + \sum_{t=1}^{T} \|w^{(t+1)} - w^*\|^2 \ge \left(\sum \|w^{(t+1)} - w^*\|^2 + \eta \sum_{t=1}^{T} \|v_t\|^2\right)^2$

3] $f(w^{(1)}) - f(w^*) \leq \langle w^{(1)} - w^*, f'(w^{(1)}) \rangle$

$f(w^{(2)}) - f(w^*) \leq \langle w^{(2)} - w^*, f'(w^{(2)}) \rangle$

$\vdots$

$f(w^{(T)}) - f(w^*) \leq \langle w^{(T)} - w^*, f'(w^{(T)}) \rangle$



for convex function instantaneous slope is greater than average slope

$$\frac{f(w_1) - f(w_2)}{w_1 - w_2} < f'(w_1)$$

Add all LHS's and RHS's together.

$$\left(\sum_{t=1}^{T} f(w^{(t)})\right) - T \cdot f(w^*) \leq \sum \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$$

divide by T,

$$\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$$

by Jensen's Inequality, $f\left(\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})\right) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})$

So,

$$\underbrace{f\left(\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})\right)}_{f(\bar{w})} - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^*)$$

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$$

take result from #2, and divide by T.

$$\frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{T}\left(\frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|v_t\|^2\right) \leq \cdots \rightarrow$$

$$\dashrightarrow \quad \leq \frac{1}{T}\left(\frac{B^2}{\eta} + \frac{\eta}{2} \cdot T\rho^2\right)$$

Now, rewrite (eq 6) using this inequality

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

if $\eta = \sqrt{\frac{B^2}{\rho^2 T}} = \frac{B}{\rho} \frac{1}{\sqrt{T}}$ then

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

$$\leq \frac{BP}{2} \cdot \left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{T}}\right)$$

$$\leq \frac{BP}{2} \cdot \frac{2}{\sqrt{T}}$$

$$f(\bar{w}) - f(w^*) \leq \frac{BP}{\sqrt{T}}$$

Note, $\frac{BP}{\sqrt{T}} \propto \frac{1}{\sqrt{T}}$

4. $f_1(w) = -\ln\left(1 - \frac{1}{1+e^{-w}}\right) = -\ln\left(\frac{e^{-w}}{1+e^{-w}}\right)$

$$\frac{d f_1(w)}{dw} = (-1)\cdot\left(\frac{1+e^{-w}}{e^{-w}}\right)\cdot \frac{e^{-w}(-1)\cdot(1+e^{-w}) - e^{-w}\cdot e^{-w}(-1)}{(1+e^{-w})^2}$$

$$= \frac{-\cancel{(1+e^{-w})}}{\cancel{e^{-w}}}\cdot\frac{-\cancel{e^{-w}} - \cancel{e^{-2w}} + \cancel{e^{-2w}}}{\cancel{(1+e^{-w})^2}_1}$$

$$= \frac{1}{1+e^{-w}} > 0$$

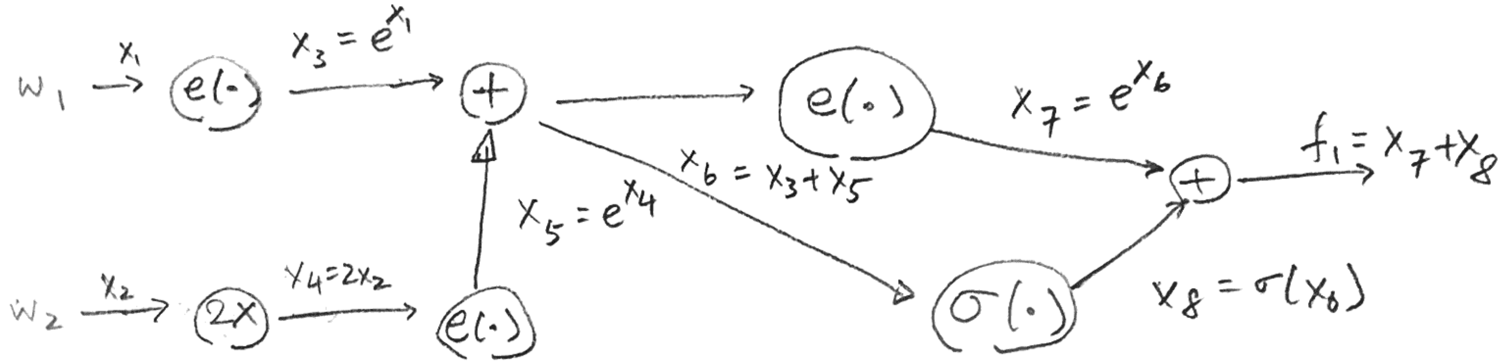
$$f_2(w) = -\ln\left(\frac{1}{1+e^{-w}}\right)$$

$$\frac{d f_2(w)}{dw} = (-1)(1+e^{-w})\cdot\frac{(-1)(e^{-w})(-1)}{(1+e^{-w})^2}$$

$$= \frac{-e^{-w}}{1+e^{-w}} < 0$$

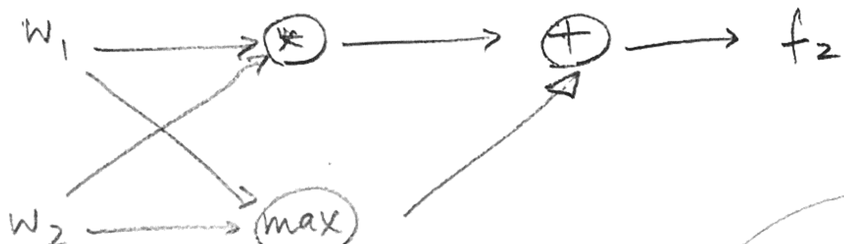
There is <u>no</u> guarantee, since the gradients, depending on which term is picked, is in opposite directions.

5) a) $f_1(w_1, w_2) = e^{e^{w_1} + e^{2w_2}} + \sigma(e^{w_1} + e^{2w_2})$



$f_1(1,-1) = e^{(e+e^{-2})} + \sigma(e+e^{-2}) = 18.30$

$f_2(w_1, w_2) = w_1 w_2 + \max(w_1, w_2)$



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

$\therefore f = \begin{bmatrix} 18.30 \\ 0 \end{bmatrix}$

b) Jacobian $\Rightarrow J_{i,j} = \frac{\partial}{\partial x_j} f_i(x)$

$$\begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix} = \begin{pmatrix} 48.2 & 4.76 \\ 0 & 1 \end{pmatrix}$$

$*$ these values are computed from the following --->

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Set $h = \Delta w = 0.01$, $\vec{w} = (1, -1)$

$$\frac{\partial f_1}{\partial w_1} = \frac{f_1(1.01, -1) - f_1(1, -1)}{0.01} = 48.192$$

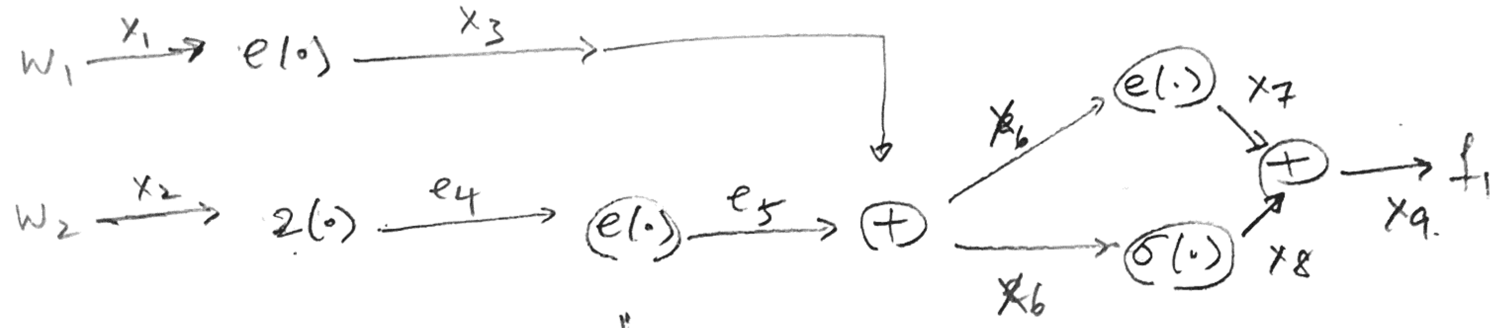
$$\frac{\partial f_1}{\partial w_2} = \frac{f_1(1, -0.99) - f_1(1, -1)}{0.01} = 4.764$$

$$\frac{\partial f_2}{\partial w_1} = \frac{f_2(1.01, -1) - f_2(1, -1)}{0.01} = 0$$

$$\frac{\partial f_2}{\partial w_2} = \frac{f_2(1, -0.99) - f_2(1, -1)}{0.01} = 1$$

---

5) c) F_1, forward, so start from x_1, x_2.



equations $\boxed{\text{Table 1}}$	$\frac{\partial f_1}{\partial w_1}$	$\frac{\partial f_1}{\partial f_2}$
$x_1 = w_1$ $x_2 = w_2$	$\dot{x}_1 = 1$ $\dot{x}_2 = 0$	$\dot{x}_1 = 0$ $\dot{x}_2 = 1$
$x_3 = e^{x_1}$	$\dot{x}_3 = e^{x_1} \cdot \dot{x}_1$	
$x_4 = 2x_2$	$\dot{x}_4 = 2\dot{x}_2$	
$x_5 = e^{x_4}$	$\dot{x}_5 = e^{x_4} \cdot \dot{x}_4$	
$x_6 = x_3 + x_5$	$\dot{x}_6 = \dot{x}_3 + \dot{x}_5$	
$x_7 = e^{x_6}$	$\dot{x}_7 = e^{x_6} \cdot \dot{x}_6$	
$x_8 = \sigma(x_6)$	$\dot{x}_8 = \sigma(x_6)(1-\sigma(x_6)) \cdot \dot{x}_6$	
$x_9 = x_7 + x_8$	$\dot{x}_9 = \dot{x}_7 + \dot{x}_8$	

Consider $\frac{\partial f_1}{\partial w_1}$, $\dot{x}_1 = 1$, $\dot{x}_2 = 0$ (from chart above ↑)

$x_1 = w_1 = 1$, $x_2 = w_2 = -1$

$$\frac{\partial f_1}{\partial w_1} = \dot{x}_9 = \dot{x}_7 + \dot{x}_8 = \left(e^{x_6} \cdot \dot{x}_6\right) + \left[\sigma(x_6)\left(1 - \sigma(x_6)\right)\dot{x}_6\right]$$

$$\dot{x}_6 = \dot{x}_3 + \dot{x}_5 = \left(e^{x_1} \cdot \dot{x}_1\right) + \left(e^{x_4} \cdot \dot{x}_4\right)$$

$$= \left(e^{x_1} \cdot \dot{x}_1\right) + \left(e^{x_4} \cdot 2\dot{x}_2\right)$$

$$= \left(e^{x_1} \cdot 1\right) + \left(e^{x_4} \cdot 2(0)\right) = e^{x_1}$$

So, $\frac{\partial f_1}{\partial w_1} = e^{x_6}\left(e^{x_1}\right) + \left[\sigma(x_6)\left(1 - \sigma(x_6)\right)\right] e^{x_1}$

$x_1 = w_1$

(☆☆) $x_6 = x_3 + x_5 = e^{x_1} + e^{x_4} = e^{x_1} + e^{2x_2} = e^{w_1} + e^{2w_2}$

(A) So, $\frac{\partial f_1}{\partial w_1} = e^{\left(e^{w_1} + e^{2w_2}\right)} \cdot e^{w_1} + \sigma\left(e^{w_1} + e^{2w_2}\right)\left(1 - \sigma\left(e^{w_1} + e^{2w_2}\right)\right) e^{w_1}$

Consider $\frac{\partial f_1}{\partial w_2}$, $\dot{x}_1 = 0$, $\dot{x}_2 = 1$

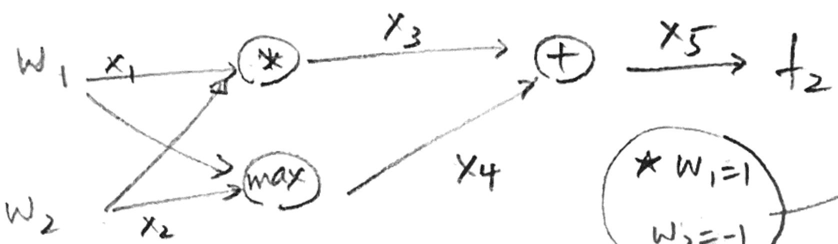
$\dot{x}_3 = 0$, $\dot{x}_4 = 2$, $\dot{x}_5 = 2e^{x_4}$, $\dot{x}_6 = \dot{x}_5$, $\dot{x}_7 = e^{x_6} \cdot 2e^{x_4}$

$\dot{x}_8 = \sigma(x_6)\left(1 - \sigma(x_6)\right) \cdot 2e^{x_4}$

$\dot{x}_9 = e^{x_6} \cdot 2e^{x_4} + \sigma(x_6)\left(1 - \sigma(x_6)\right) \cdot 2e^{x_4}$

(B) ∴ $\frac{\partial f_1}{\partial w_2} = e^{\left(e^{w_1} + e^{2w_2}\right)} \cdot 2e^{2w_2} + \sigma\left(e^{w_1} + e^{2w_2}\right)\left(1 - \sigma\left(e^{w_1} + e^{2w_2}\right)\right) 2e^{2w_2}$

5] c] continued. F2, forward



$\star\ w_1 = 1$
$w_2 = -1$

Table 2

equation	$\frac{\partial f_{w_2}}{\partial w_1}$	$\frac{\partial f_2}{\partial w_2}$
$x_1 = w_1$ $x_2 = w_2$	$\dot{x}_1 = 1$ $\dot{x}_2 = 0$	$\dot{x}_1 = 0$ $x_2 = 1$
$x_3 = x_1 \cdot x_2$	$\dot{x}_3 = \dot{x}_1 x_2 + x_1 \dot{x}_2$	
$x_4 = \max(x_1, x_2)$	$\dot{x}_4 = \begin{cases} \dot{x}_1 & \text{if } x_1 \geq x_2 \\ \dot{x}_2 & \text{o.w.} \end{cases}$	
$x_5 = x_3 + x_4$	$\dot{x}_5 = \dot{x}_3 + \dot{x}_4$	

Consider $\frac{\partial f_2}{\partial w_1}$, $\dot{x}_1 = 1$, $\dot{x}_2 = 0$

$\dot{x}_3 = x_2$, $\dot{x}_4 = \dot{x}_1 = 1$

$\frac{\partial f_2}{\partial w_1} = \dot{x}_5 = x_2 + 1 = w_2 + 1$ (C)

Consider $\frac{\partial f_2}{\partial w_2}$, $\dot{x}_1 = 0$, $\dot{x}_2 = 1$

$\dot{x}_3 = x_1 = w_1$

$\dot{x}_4 = \dot{x}_1 = 0$

$\frac{\partial f_2}{\partial w_2} = \dot{x}_5 = w_1 + 0 = w_1$ (D)

from (A), $\frac{\partial f_1}{\partial w_1} = 47.3$

from (B), $\frac{\partial f_1}{\partial w_2} = 4.71$

from (C), $\frac{\partial f_2}{\partial w_1} = 0$

from (D), $\frac{\partial f_2}{\partial w_2} = 1$

So, Jacobian is:

$$\begin{pmatrix} 47.3 & 4.71 \\ 0 & 1 \end{pmatrix}$$

5] d] F1, backward (so start at $\dot{x}_9$).

$\dot{x}_9 = 1$, $\quad x_9 = x_7 \oplus x_8$, so pass gradient to both.

$\therefore \dot{x}_8 = \dot{x}_9 = 1$ and $\dot{x}_7 = \dot{x}_9 = 1$

Since x_6 goes onto 2 nodes, $\dot{x}_6$ must sum both gradients.

$$\therefore \dot{x}_6 = e^{x_6} \cdot \dot{x}_7 + \sigma(x_6)(1-\sigma(x_6))\dot{x}_8$$

$$= e^{x_6} + \sigma(x_6)(1-\sigma(x_6))$$

$\dot{x}_5 = \dot{x}_6$, and $\dot{x}_3 = \dot{x}_6$, since $x_6 = x_3 \oplus x_5$

$$\dot{x}_4 = e^{x_4} \cdot \dot{x}_5 = e^{x_4}\dot{x}_6 = e^{x_4}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

$$\dot{x}_2 = 2\dot{x}_4 = 2e^{x_4}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

$$\dot{x}_1 = e^{x_1} \cdot \dot{x}_3 = e^{x_1} \cdot \dot{x}_6 = e^{x_1}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

(a) $$\frac{\partial f_1}{\partial w_1} = \dot{x}_1 = e^{w_1}\left(e^{(e^{w_1}+e^{2w_2})} + \sigma(e^{w_1}+e^{2w_2})\left(1-\sigma(e^{w_1}+e^{2w_2})\right)\right)$$

$\boxed{\text{from } (\star\star) \text{ of part c, we know } x_6.}$

(b) $$\frac{\partial f_1}{\partial w_2} = \dot{x}_2 = 2e^{2w_2}\left(e^{(e^{w_1}+e^{2w_2})} + \sigma(e^{w_1}+e^{2w_2})\left(1-\sigma(e^{w_1}+e^{2w_2})\right)\right)$$

5| d| continued. Backward, F_2.

$\dot{x}_5 = 1$, $\dot{x}_4 = \dot{x}_5 = 1$, $\dot{x}_3 = \dot{x}_5 = 1$

$\dot{x}_1 = x_2 \dot{x}_3 + \dot{x}_4 = x_2 = w_2 + 1$

$\dot{x}_2 = x_1 \dot{x}_3 + 0 = w_1$

(c) $\dfrac{\partial f_2}{\partial w_1} = w_2 + 1 = \dot{x}_1$, (d) $\dfrac{\partial f_2}{\partial w_2} = w_1 = \dot{x}_2$.

Note a, b, c, d of part d = A, B, C, D of part c.

i.e. Forward & Backward produce same formula for each entry in ~~Gradient~~. Jacobian

So,

$$J = \begin{pmatrix} 47.3 & 4.71 \\ 0 & 1 \end{pmatrix}$$ "same as part c"

5| e| Yes! I do!

Q6/7: Paper Review

0-Main Contribution

The main contribution of the paper was to challenge the notion that learning weights for neural networks is the only effective way of producing high performing networks. If this notion is removed, then the process for solving neural networks would be reduced to only designing the structure of the network; the weights can be randomly initialized and the network is ready for use. The authors discuss their method for learning the network structure automatically.

1-Strengths

Weight Agnostic Neural Nets (WANN) are immediately useful, and do not require tuning of weights, whereas traditional parameterized models require extensive training to tune good/useful weights.

2-Drawbacks

Looking at results table, we see that there is more variance in performance metrics for WANN than fixed topology models. This might be just a natural by-product of random/shared weights. This variance goes away, however, with weight tuning. Even with tuning, WANN lose to traditional models on benchmark problems like Biped, CarRacing, MNIST(2 layer CNN). However, the authors do note that this is somewhat of an unfair comparison. Many fixed topology networks(they focus on CNNs) are products of years of research and experimentation.

Another important weakness of the paper is that it fails to discuss the time tradeoff between searching for the WANN and tuning optimal parameters. The authors did not show that using WANN produced significant speedup compared to training traditional fixed topologies to find optimal weights. This is not exactly a red flag, but it would have provided a stronger argument for researching WANN.

3-Personal Takeaways

This was very interesting to me, as I had read about NEAT as part of my VIP work. The VIP team works on a genetic programming autoML framework, and some members were interested in incorporating NEAT into the framework. The evolutionary style of searching for WANN makes me think that this is also a good feature to implement.

Conceptually, I find that this provides a nice balance in thinking about neural network design and optimization. On one end of the spectrum is the traditional optimization of parameters and the other end is WANN.

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

In [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
```

In [2]:
```
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`

In [3]:
```
# 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: 2.344749e+00 computed in 0.257661s
loss: 2.344749
sanity check: 2.302585
numerical: 1.965265 analytic: 1.965265, relative error: 4.050525e-08
numerical: 4.230704 analytic: 4.230704, relative error: 1.309142e-08
numerical: -0.892266 analytic: -0.892266, relative error: 5.395355e-08
numerical: 2.159850 analytic: 2.159850, relative error: 6.890465e-09
numerical: 0.289739 analytic: 0.289739, relative error: 1.286889e-07
numerical: -0.882801 analytic: -0.882801, relative error: 5.561869e-08
numerical: 2.027293 analytic: 2.027293, relative error: 1.478793e-08
numerical: -0.146122 analytic: -0.146122, relative error: 3.984646e-07
numerical: -0.920213 analytic: -0.920213, relative error: 4.301624e-08
numerical: -2.725859 analytic: -2.725859, relative error: 3.388619e-08
```

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

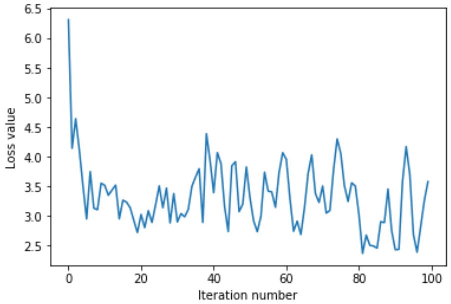
In [4]:
```python
# Now that efficient implementations to calculate loss function and gradient of the softmax are ready,
# use it to train the classifier on the cifar-10 data
# Complete the `train` function in cs231n/classifiers/linear_classifier.py

from cs231n.classifiers.linear_classifier import Softmax

classifier = Softmax()
loss_hist = classifier.train(
    X_train,
    y_train,
    learning_rate=1e-5,
    reg=1e-5,
    num_iters=100,
    batch_size=200,
    verbose=True,
)
# Plot loss vs. iterations
plt.plot(loss_hist)
plt.xlabel("Iteration number")
plt.ylabel("Loss value")
```

```
found batch.
loss + grad computed.
weights updated.
iteration 0 / 100: loss 6.316712
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 10 / 100: loss 3.519489
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 20 / 100: loss 3.028383
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 30 / 100: loss 2.901936
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
```

```
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 40 / 100: loss 3.395893
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 50 / 100: loss 3.307890
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 60 / 100: loss 3.959330
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
```

```
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 70 / 100: loss 3.508856
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 80 / 100: loss 3.023361
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
iteration 90 / 100: loss 2.433835
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
found batch.
loss + grad computed.
weights updated.
```

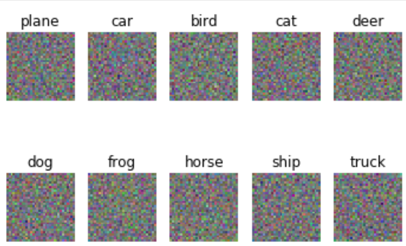
Out[4]: Text(0,0.5,'Loss value')



In [5]:
```
# 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: 0.273000
```

In [6]:
```
# 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])
```



In [ ]:

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.

In [83]:
```
# 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 autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

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.

In [84]:
```
# 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()
```

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.

In [85]:
```
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
```

Forward pass: compute loss

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

In [86]:
```
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:
4.6769255135359344e-12
```

Backward pass

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

In [87]:
```
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)
print(X.shape)
# 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)
    temp = grads[param_name]
    print(f"param: {param_name}")
    print(f"my grad: {temp.shape}")
    print(f"answer: {param_grad_num.shape}")
    print('%s max relative error: %e' % (param_name, rel_error(param_grad_num, grads[param_name])))
    print()
```

```
(5, 4)
param: W2
my grad: (10, 3)
answer: (10, 3)
W2 max relative error: 9.913918e-10

param: b2
my grad: (3,)
answer: (3,)
b2 max relative error: 8.190173e-11

param: W1
my grad: (4, 10)
answer: (4, 10)
W1 max relative error: 4.426512e-09

param: b1
my grad: (10,)
answer: (10,)
b1 max relative error: 5.435431e-08
```

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:

In [88]:
```
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 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=0.0, 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: 0.940686
```

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:

In [89]:
```python
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='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:

In [90]:
```python
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
```

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.

In [91]:
```python
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,)
```

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.

In [73]:
```
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, 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: 0.105000, val 0.092000, 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 2.279695, train: 0.162000, val 0.178000, 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 2.079532, train: 0.237000, val 0.246000, 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 1.920950, train: 0.277000, val 0.288000, 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 1.914854, train: 0.334000, val 0.333000, 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 1.749788, train: 0.363000, val 0.365000, lr 7.737809e-06
finished optimization. best validation accuracy: 0.365000
```

Debug the training

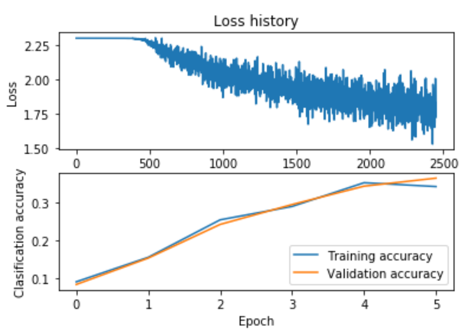
With the default parameters we provided above, you should get a validation accuracy of about 0.37 on the validation set. This isn't very good.

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.

In [56]:
```
# 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('Clasification accuracy')
```

Out[56]:
```
Text(0,0.5,'Clasification accuracy')
```



In [57]:
```
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)
```

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: You 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.).

In [122]:

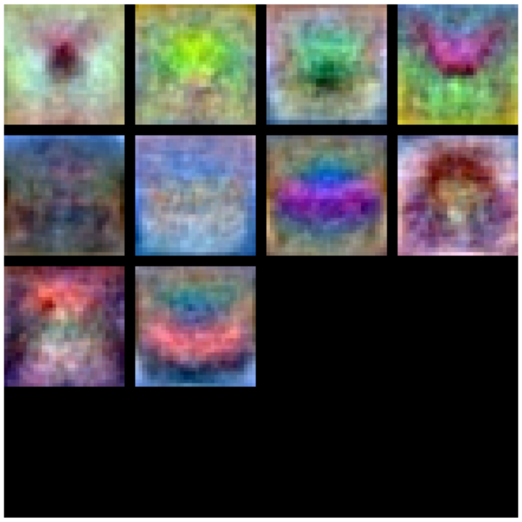
```python
best_model = None # store the best model into this
best_val_acc = 0.0
################################################################################
# TODO: Tune hyperparameters using the validation set. Store your best trained  #
# model in best_model.                                                         #
#                                                                              #
# 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
hidden_size = [10]#[10, 30, 50]
num_epochs = [25, 50]
reg = [2, 3]#[0.95, 0.99] #[0, 0.5, 0.85, 0.9, 0.95, 0.99]
momentum = [.99,.999]#[0.9, 0.99]
learning_rate_decay = [0.85, 0.9]#, 0.95]#[0.9, 0.99, 0.999]
learning_rate = [1e-5]
update = ["rmsprop", "momentum"]#, "momentum"]#["sgd", "momentum"]#["rmsprop", "sgd", "momentum"]

trainer = ClassifierTrainer()
for hidden_layer_size in hidden_size:
    model = init_two_layer_model(32*32*3, hidden_layer_size, 10)
    for epochs in num_epochs:
        for u in update:
            for r in reg:
                for mom in momentum:
                    for decay in learning_rate_decay:
                        for alpha in learning_rate:
#                               print(f"hidden_size: {hidden_layer_size}")
                            print(f"update: {u}")
                            print(f"epochs: {epochs}")
                            print(f"regularize: {r}")
                            print(f"momentum: {mom}")
                            print(f"learning_rate: {alpha}")
                            print(f"learning_rate_decay: {decay}")
                            good_model, loss_history, train_acc, val_acc = trainer.train(X_train, y_train,
                                                 X_val, y_val,
                                                 model, two_layer_net,
                                                 num_epochs=epochs, reg=r,
                                                 momentum=mom,
                                                 learning_rate_decay=decay,
                                                 update=u,
                                                 learning_rate=alpha, verbose=False)
                            print(val_acc)
                            print()
                            good_val_acc = np.max(val_acc)
                            if good_val_acc > best_val_acc:
                                best_val_acc = good_val_acc
                                best_model = good_model


# hidden_size = [10]#[10, 30, 50]
# num_epochs = [50]
# reg = [3.0]#[0.95, 0.99] #[0, 0.5, 0.85, 0.9, 0.95, 0.99]
# momentum = [.99]#[0.9, 0.99]
# learning_rate_decay = [0.85]#[0.85, 0.9, 0.95]#[0.9, 0.99, 0.999]
# learning_rate = [1e-10]#[1e-5, 1e-10, 1e-20] #[1e-1, 1e-5, 1e-10]
# # learning_rate = [1e-6, 1e-7]
# update = [ "momentum" ]

# trainer = ClassifierTrainer()
# for hidden_layer_size in hidden_size:
#     model = init_two_layer_model(32*32*3, hidden_layer_size, 10)
#     for epochs in num_epochs:
#         for u in update:
#             for r in reg:
#                 for mom in momentum:
#                     for decay in learning_rate_decay:
#                         for alpha in learning_rate:
#                             print(f"update: {u}")
#                             print(f"epochs: {epochs}")
#                             print(f"regularize: {r}")
#                             print(f"momentum: {mom}")
#                             print(f"learning_rate: {alpha}")
#                             print(f"learning_rate_decay: {decay}")
#                             good_model, loss_history, train_acc, val_acc = trainer.train(X_train, y_train,
#                                                  X_val, y_val,
#                                                  model, two_layer_net,
#                                                  num_epochs=epochs, reg=r,
#                                                  momentum=mom,
#                                                  learning_rate_decay=decay,
#                                                  update=u,
#                                                  learning_rate=alpha, verbose=False)
#                             print(val_acc)
#                             print()
#                             good_val_acc = np.max(val_acc)
#                             if good_val_acc > best_val_acc:
#                                 best_val_acc = good_val_acc
#                                 best_model = good_model
################################################################################
#                                 END OF YOUR CODE                              #
################################################################################
```

```
update: rmsprop
epochs: 25
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.067, 0.255, 0.281, 0.305, 0.304, 0.322, 0.324, 0.328, 0.333, 0.336, 0.343, 0.34, 0.337, 0.343, 0.341, 0.343, 0.346, 0.347, 0.349, 0.351, 0.351,
0.349, 0.353, 0.351, 0.35, 0.351]

update: rmsprop
epochs: 25
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.352, 0.356, 0.366, 0.38, 0.375, 0.384, 0.377, 0.387, 0.391, 0.388, 0.384, 0.386, 0.388, 0.391, 0.393, 0.39, 0.393, 0.395, 0.396, 0.393, 0.386,
0.396, 0.394, 0.392, 0.395, 0.392]

update: rmsprop
epochs: 25
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.391, 0.391, 0.4, 0.396, 0.397, 0.396, 0.398, 0.395, 0.398, 0.399, 0.398, 0.396, 0.398, 0.398, 0.397, 0.402, 0.401, 0.399, 0.4, 0.399, 0.399, 0.3
99, 0.401, 0.4, 0.402, 0.4]

update: rmsprop
epochs: 25
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.4, 0.397, 0.401, 0.396, 0.404, 0.391, 0.407, 0.401, 0.408, 0.405, 0.405, 0.405, 0.404, 0.405, 0.406, 0.403, 0.403, 0.405, 0.408, 0.408, 0.407,
0.401, 0.407, 0.409, 0.402, 0.408]

update: rmsprop
epochs: 25
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.409, 0.407, 0.405, 0.406, 0.404, 0.409, 0.404, 0.409, 0.404, 0.402, 0.405, 0.404, 0.406, 0.398, 0.402, 0.399, 0.399, 0.397, 0.4, 0.403, 0.402,
0.401, 0.404, 0.402, 0.402, 0.402]

update: rmsprop
epochs: 25
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.405, 0.398, 0.399, 0.399, 0.41, 0.399, 0.405, 0.406, 0.403, 0.407, 0.408, 0.406, 0.404, 0.405, 0.41, 0.404, 0.403, 0.396, 0.407, 0.402, 0.408,
0.404, 0.403, 0.403, 0.405, 0.405]

update: rmsprop
epochs: 25
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.402, 0.406, 0.409, 0.404, 0.404, 0.398, 0.403, 0.406, 0.411, 0.403, 0.407, 0.408, 0.402, 0.404, 0.404, 0.403, 0.402, 0.405, 0.403, 0.404, 0.404,
0.4, 0.402, 0.404, 0.406, 0.403]

update: rmsprop
epochs: 25
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.403, 0.406, 0.403, 0.408, 0.409, 0.402, 0.402, 0.412, 0.405, 0.407, 0.405, 0.407, 0.406, 0.408, 0.408, 0.4, 0.405, 0.407, 0.406, 0.405, 0.402,
0.404, 0.406, 0.407, 0.411, 0.404]

update: momentum
epochs: 25
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.15, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 2
momentum: 0.999
learning_rate: 1e-05
```

```
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 25
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 3
momentum: 0.999
learning_rate: 1e-05
```

```
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: rmsprop
epochs: 50
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 2
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 2
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 3
momentum: 0.99
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.85
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]

update: momentum
epochs: 50
regularize: 3
momentum: 0.999
learning_rate: 1e-05
learning_rate_decay: 0.9
[0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087,
0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087, 0.087]
```

In [123]:
```
# visualize the weights
show_net_weights(best_model)
```



Run on the test set

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

In [124]:
```
scores_test = two_layer_net(X_test, best_model)
print('Test accuracy: ', np.mean(np.argmax(scores_test, axis=1) == y_test))
```

```
Test accuracy:  0.391
```

In [ ]:

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 recieve 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:

```python
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
```

In [1]:
```python
%pip install scipy==1.1.0
```

```
Requirement already satisfied: scipy==1.1.0 in /miniconda3/envs/cs4803/lib/python3.6/site-packages (1.1.0)
Requirement already satisfied: numpy>=1.8.2 in /miniconda3/envs/cs4803/lib/python3.6/site-packages (from scipy==1.1.0) (1.18.1)
WARNING: You are using pip version 19.3.1; however, version 20.0.2 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```python
# As usual, a bit of setup

import numpy as np
import matplotlib.pyplot as plt
from cs231n.gradient_check import eval_numerical_gradient_array, 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))))
```

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:

In [3]:
```python
# 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: ', rel_error(out, correct_out))
```

```
Testing affine_forward function:
difference:  9.769848888397517e-10
```

Affine layer: backward

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

In [4]:
```
# Test the affine_backward function

x = np.random.randn(10, 2, 3)
w = np.random.randn(6, 5)
b = np.random.randn(5)
dout = np.random.randn(10, 5)

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

_, cache = affine_forward(x, w, b)
dx, dw, db = affine_backward(dout, cache)

# The error should be 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))
```

```
dout (10, 5)
x (10, 2, 3)
w (6, 5)
Testing affine_backward function:
dx error:  2.9002938143112204e-10
dw error:  2.9279415839353796e-11
db error:  1.6606257785141173e-11
```

ReLU layer: forward

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

In [5]:
```
# Test the relu_forward function

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

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

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

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

ReLU layer: backward

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

In [6]:
```
x = np.random.randn(10, 10)
dout = np.random.randn(*x.shape)

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

_, cache = relu_forward(x)
dx = relu_backward(dout, cache)

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

```
Testing relu_backward function:
dx error:  3.275590221433177e-12
```

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.

In [7]:
```python
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)

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

# Test svm_loss function. Loss should be around 9 and dx error should be 1e-9
print('Testing svm_loss:')
print('loss: ', 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:  8.997770679279384
dx error:  8.182894472887002e-10

Testing softmax_loss:
loss:  2.3023626383992823
dx error:  8.380339587422198e-09
```

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:

In [8]:
```python
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))
```

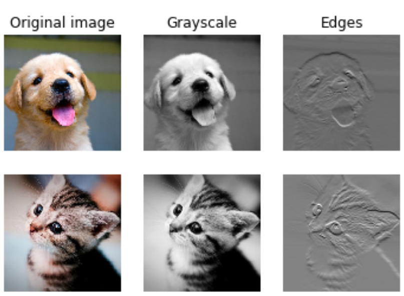
```
Testing conv_forward_naive
difference:  2.212147649671884e-08
```

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.

In [9]:
```python
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()
```

```
/miniconda3/envs/cs4803/lib/python3.6/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
/miniconda3/envs/cs4803/lib/python3.6/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.
/miniconda3/envs/cs4803/lib/python3.6/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()
```



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.

In [10]:
```python
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:  3.5132240751539185e-09
dw error:  5.588793674786507e-10
db error:  8.541292449471228e-11
```

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` .

In [11]:
```python
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(out)
print('Testing max_pool_forward_naive function:')
print('difference: ', rel_error(out, correct_out))
```

```
[[[[-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       ]]]]
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.

In [12]:
```python
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:  3.2756338833305833e-12
```

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 recieves 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:

In [13]:

```
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: 8.384693s
Fast: 0.018816s
Speedup: 445.615193x
Difference:  5.057211705011567e-09

Testing conv_backward_fast:
Naive: 13.820877s
Fast: 0.017050s
Speedup: 810.607302x
dx difference:  9.6336246640444e-10
dw difference:  7.832922414345895e-13
db difference:  1.8295185088741384e-14
```

```python
In [14]: 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('speedup: %fx' % ((t1 - t0) / (t2 - t1)))
print('dx difference: ', rel_error(dx_naive, dx_fast))
```

```
Testing pool_forward_fast:
Naive: 0.851607s
fast: 0.003324s
speedup: 256.197031x
difference:  0.0

Testing pool_backward_fast:
Naive: 0.708919s
speedup: 53.302416x
dx difference:  0.0
```

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:

```python
In [15]: 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:  1.0219630800884883e-08
dw error:  9.972214355734462e-08
db error:  9.43250581562441e-11
```

```python
In [16]: 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:  9.395438932876684e-10
dw error:  2.1027908347897907e-09
db error:  2.580480876527029e-11
```

In [17]:
```
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))
```

```
dout (2, 10)
x (2, 3, 4)
w (12, 10)
Testing affine_relu_forward:
dx error:  2.730720584267022e-10
dw error:  3.9296785544252803e-10
db error:  1.4574102144557293e-11
```

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` .

In [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))))
```

In [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,)
```

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.

In [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)
```

```
dout (100, 10)
x (100, 32, 16, 16)
w (8192, 10)
Sanity check loss (no regularization):  2.302680787589569
dout (100, 10)
x (100, 32, 16, 16)
w (8192, 10)
Sanity check loss (with regularization):  2.344563384742511
```

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 artifical data and a small number of neurons at each layer.

In [5]:
```
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: 1.462154e-06
W2 max relative error: 1.090814e-05
b1 max relative error: 1.783711e-08
b2 max relative error: 2.063144e-09
```

Overfit small data

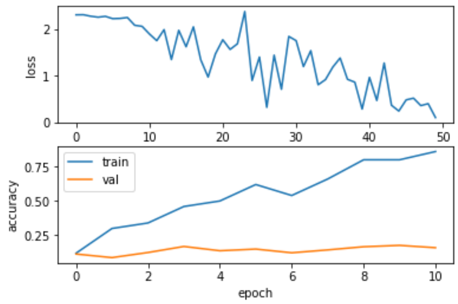
A nice trick is to train your model with just a few training samples. You should be able to overfit small datasets, which will result in very high training accuracy and comparatively low validation accuracy.

In [6]:
```
# 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 2.304942, train: 0.120000, val 0.114000, lr 1.000000e-04
Finished epoch 1 / 10: cost 2.272359, train: 0.300000, val 0.088000, lr 9.500000e-05
Finished epoch 2 / 10: cost 2.058033, train: 0.340000, val 0.125000, lr 9.025000e-05
starting iteration  10
Finished epoch 3 / 10: cost 1.972939, train: 0.460000, val 0.169000, lr 8.573750e-05
Finished epoch 4 / 10: cost 1.468262, train: 0.500000, val 0.138000, lr 8.145062e-05
starting iteration  20
Finished epoch 5 / 10: cost 0.899135, train: 0.620000, val 0.150000, lr 7.737809e-05
Finished epoch 6 / 10: cost 1.842718, train: 0.540000, val 0.123000, lr 7.350919e-05
starting iteration  30
Finished epoch 7 / 10: cost 0.917425, train: 0.660000, val 0.144000, lr 6.983373e-05
Finished epoch 8 / 10: cost 0.286941, train: 0.800000, val 0.167000, lr 6.634204e-05
starting iteration  40
Finished epoch 9 / 10: cost 0.243375, train: 0.800000, val 0.177000, lr 6.302494e-05
Finished epoch 10 / 10: cost 0.103600, train: 0.860000, val 0.160000, lr 5.987369e-05
finished optimization. best validation accuracy: 0.177000
```

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

In [7]:
```
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()
```



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.

In [8]:

```
model = init_two_layer_convnet(filter_size=7)
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=50, num_epochs=1,
          acc_frequency=50, verbose=True)
```

```
starting iteration  0
Finished epoch 0 / 1: cost 2.304729, train: 0.157000, val 0.141000, lr 1.000000e-04
starting iteration  10
starting iteration  20
starting iteration  30
starting iteration  40
starting iteration  50
Finished epoch 0 / 1: cost 2.054563, train: 0.324000, val 0.317000, lr 1.000000e-04
starting iteration  60
starting iteration  70
starting iteration  80
starting iteration  90
starting iteration  100
Finished epoch 0 / 1: cost 1.659939, train: 0.398000, val 0.396000, lr 1.000000e-04
starting iteration  110
starting iteration  120
starting iteration  130
starting iteration  140
starting iteration  150
Finished epoch 0 / 1: cost 1.865885, train: 0.399000, val 0.418000, lr 1.000000e-04
starting iteration  160
starting iteration  170
starting iteration  180
starting iteration  190
starting iteration  200
Finished epoch 0 / 1: cost 1.927011, train: 0.408000, val 0.380000, lr 1.000000e-04
starting iteration  210
starting iteration  220
starting iteration  230
starting iteration  240
starting iteration  250
Finished epoch 0 / 1: cost 1.949246, train: 0.420000, val 0.408000, lr 1.000000e-04
starting iteration  260
starting iteration  270
starting iteration  280
starting iteration  290
starting iteration  300
Finished epoch 0 / 1: cost 1.785747, train: 0.446000, val 0.456000, lr 1.000000e-04
starting iteration  310
starting iteration  320
starting iteration  330
starting iteration  340
starting iteration  350
Finished epoch 0 / 1: cost 1.454454, train: 0.471000, val 0.437000, lr 1.000000e-04
starting iteration  360
starting iteration  370
starting iteration  380
starting iteration  390
starting iteration  400
Finished epoch 0 / 1: cost 1.909681, train: 0.465000, val 0.468000, lr 1.000000e-04
starting iteration  410
starting iteration  420
starting iteration  430
starting iteration  440
starting iteration  450
Finished epoch 0 / 1: cost 1.554669, train: 0.476000, val 0.456000, lr 1.000000e-04
starting iteration  460
starting iteration  470
starting iteration  480
starting iteration  490
starting iteration  500
Finished epoch 0 / 1: cost 1.535584, train: 0.452000, val 0.429000, lr 1.000000e-04
starting iteration  510
starting iteration  520
starting iteration  530
starting iteration  540
starting iteration  550
Finished epoch 0 / 1: cost 1.373795, train: 0.494000, val 0.455000, lr 1.000000e-04
starting iteration  560
starting iteration  570
starting iteration  580
starting iteration  590
starting iteration  600
Finished epoch 0 / 1: cost 1.291351, train: 0.406000, val 0.433000, lr 1.000000e-04
starting iteration  610
starting iteration  620
starting iteration  630
starting iteration  640
starting iteration  650
Finished epoch 0 / 1: cost 1.638062, train: 0.432000, val 0.439000, lr 1.000000e-04
starting iteration  660
starting iteration  670
starting iteration  680
starting iteration  690
starting iteration  700
Finished epoch 0 / 1: cost 1.379444, train: 0.455000, val 0.432000, lr 1.000000e-04
starting iteration  710
starting iteration  720
starting iteration  730
starting iteration  740
starting iteration  750
Finished epoch 0 / 1: cost 1.421884, train: 0.524000, val 0.508000, lr 1.000000e-04
starting iteration  760
starting iteration  770
starting iteration  780
starting iteration  790
starting iteration  800
Finished epoch 0 / 1: cost 1.918442, train: 0.449000, val 0.432000, lr 1.000000e-04
starting iteration  810
starting iteration  820
starting iteration  830
starting iteration  840
starting iteration  850
Finished epoch 0 / 1: cost 1.604052, train: 0.453000, val 0.449000, lr 1.000000e-04
```

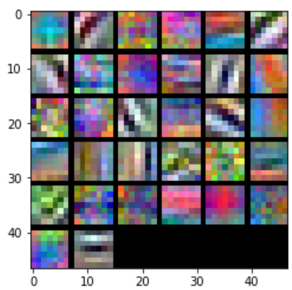
```
starting iteration  860
starting iteration  870
starting iteration  880
starting iteration  890
starting iteration  900
Finished epoch 0 / 1: cost 1.116873, train: 0.519000, val 0.503000, lr 1.000000e-04
starting iteration  910
starting iteration  920
starting iteration  930
starting iteration  940
starting iteration  950
Finished epoch 0 / 1: cost 1.589487, train: 0.435000, val 0.431000, lr 1.000000e-04
starting iteration  960
starting iteration  970
Finished epoch 1 / 1: cost 1.792701, train: 0.473000, val 0.452000, lr 9.500000e-05
finished optimization. best validation accuracy: 0.508000
```

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.

In [9]:
```python
from cs231n.vis_utils import visualize_grid

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

Out[9]: <matplotlib.image.AxesImage at 0x1041d8080>



In [ ]:

Pytorch Q8.5 - Q8.8

Train/Loss Plots, Validation Accuracy Plots, and Weight Visuals for the following:

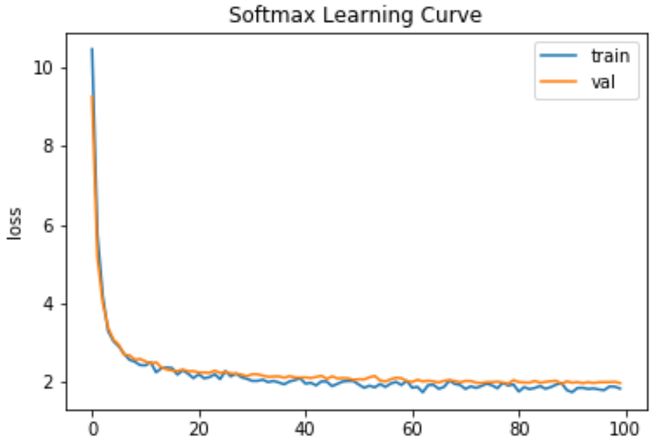
Q8.5: Softmax

Q8.6: Two layer NN

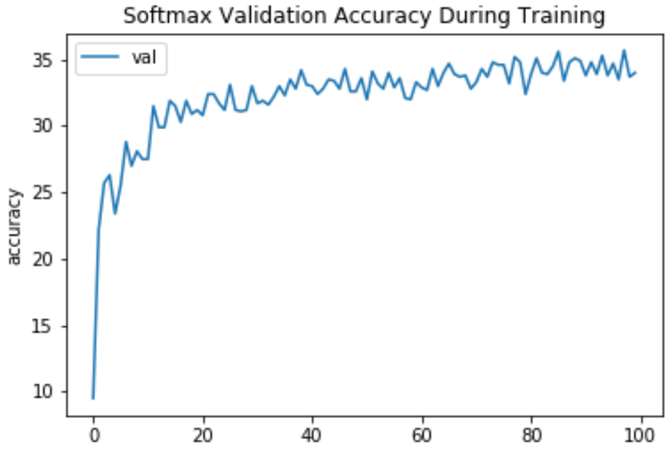
Q8.7: ConvNet

Q8.8: Experiment (MyModel)

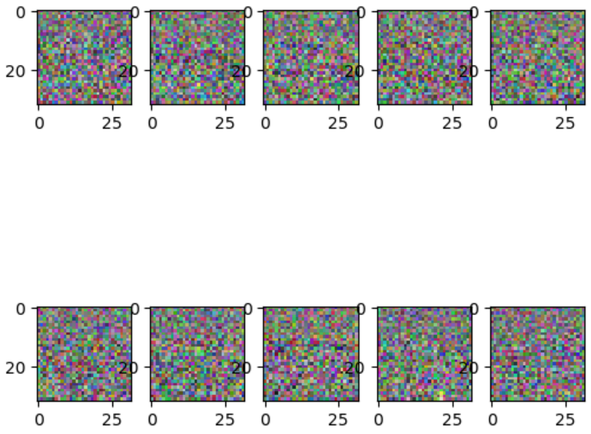
Q8.5: Softmax



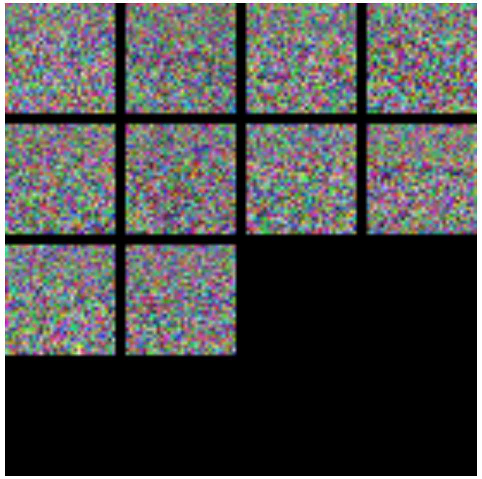
softmax_lossvstrain.png



softmax_valaccuracy.png

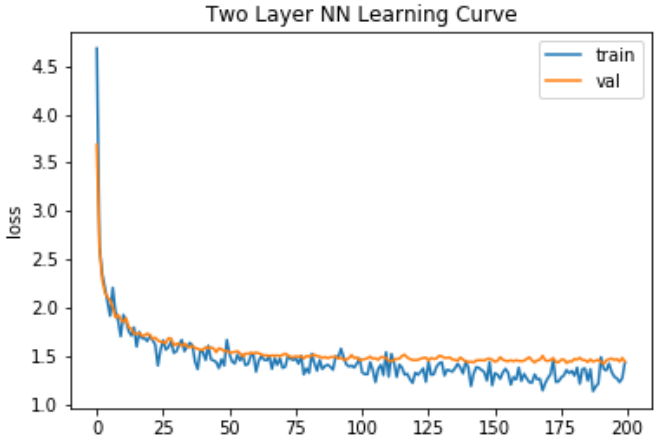


softmax_filt.png

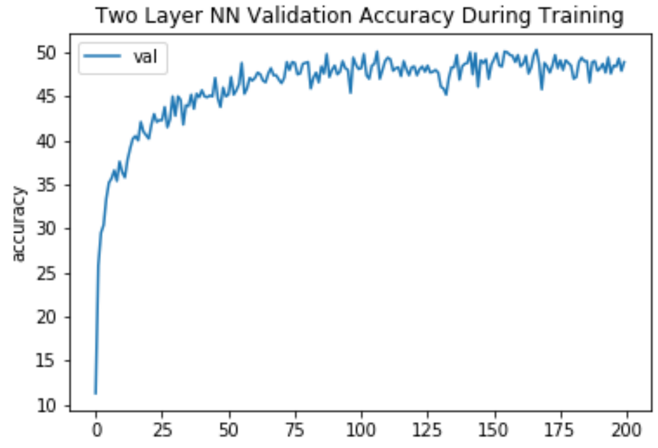


softmax_gridfilt.png

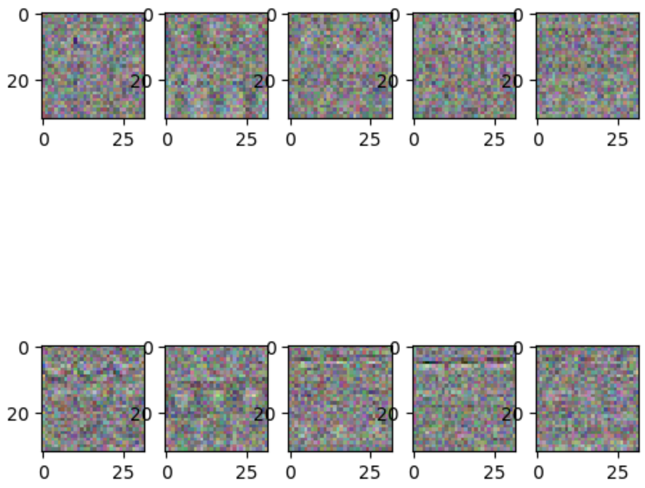
Q8.6: TwoLayerNN



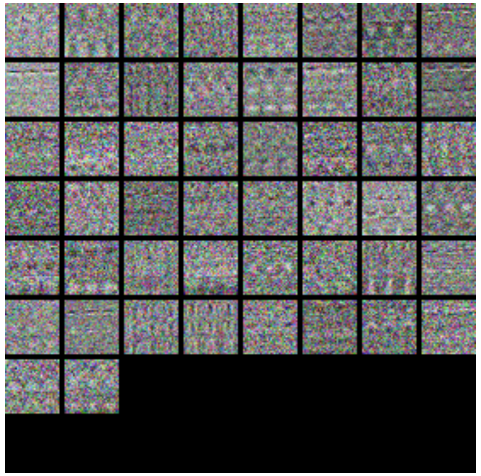
twolayernn_lossvstrain.png



twolayernn _valaccuracy.png

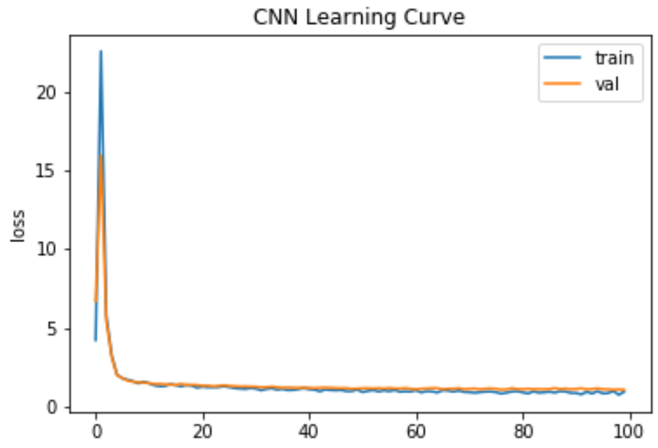


twolayernn _filt.png

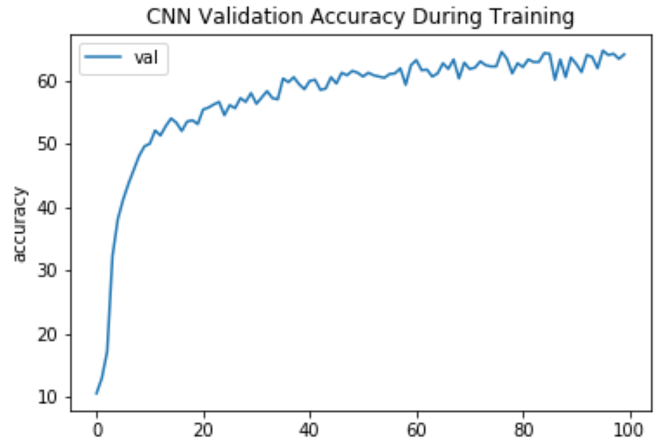


twolayernn _gridfilt.png

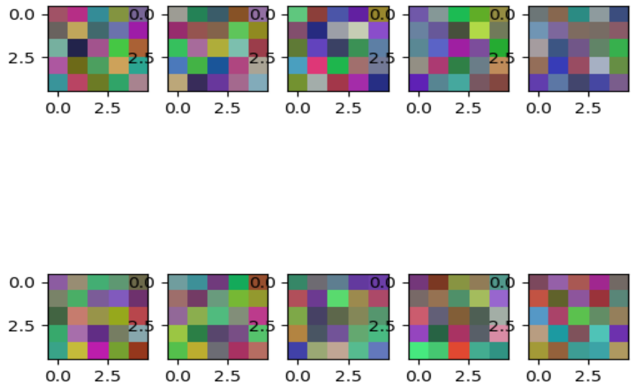
Q8.7: ConvNet



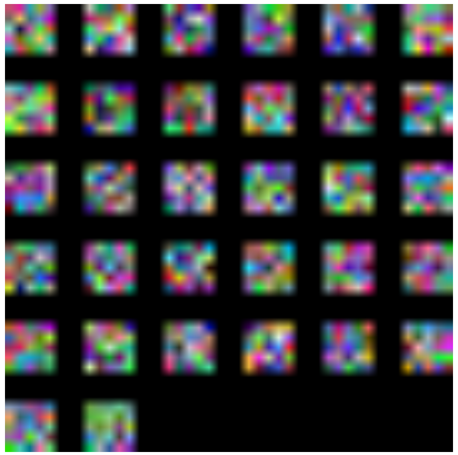
convnet_lossvstrain.png



convnet _valaccuracy.png

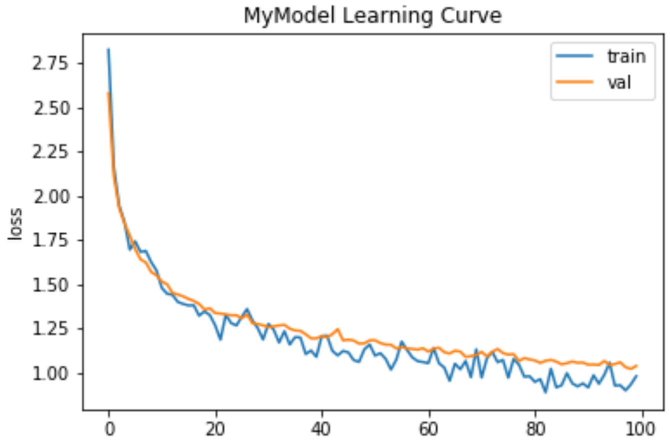


convnet _filt.png

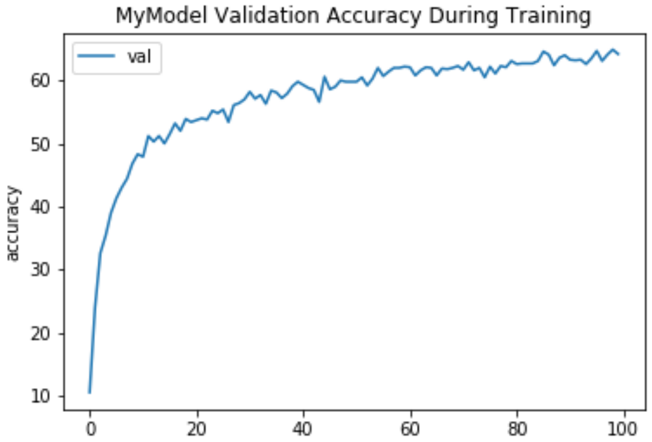


convnet _gridfilt.png

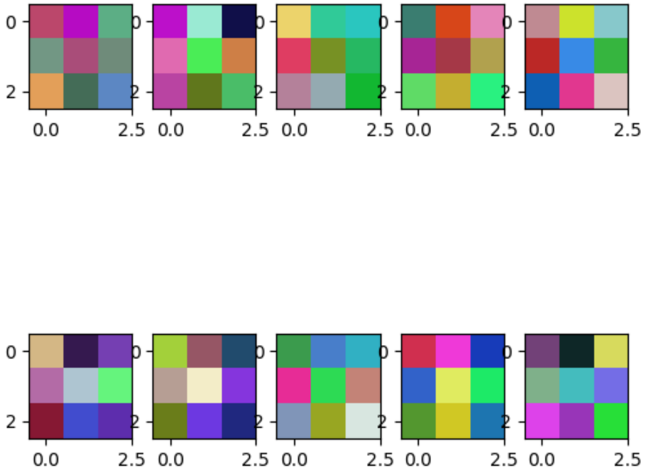
Q8.8: Experiment (MyModel)



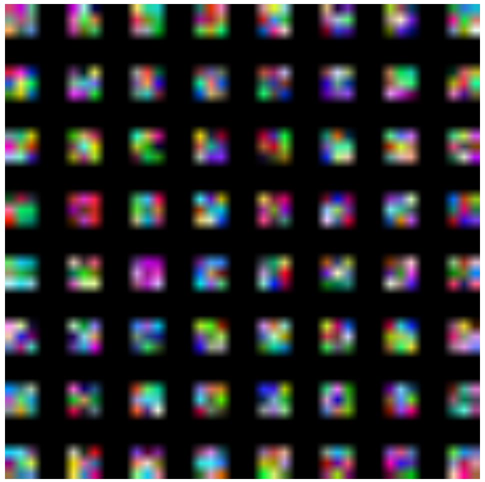
mymodel_lossvstrain.png



mymodel_valaccuracy.png



mymodel _filt.png



mymodel _gridfilt.png