DD2424 Deep Learning

Assignment 1

Yunpeng Ma

1. Introduction

The purpose is to train a neural network using mini-batch gradient descent which is applied to a cross-entropy loss function classifier with regulation.

2. Methodology

In order to train a multi-linear classifier, we have dataset of CIFAR-10, the output will be vector of probabilities. The following equations will be used for this training proce:

$$s = Wx + b$$

$$p = \frac{\exp(s)}{1^T \exp(s)}$$

During the training process, we will learn the parameters of W and b of the cost function:

$$W^*, b^* = \arg\min_{W,b} J(D, \lambda, W, b)$$

J is a cost function of cross-entropy with a regularization term:

$$J(D, \lambda, W, b) = \frac{1}{|D|} \sum_{x,y \in D} \log(y^T p) + \lambda \sum_{i,j} W_{ij}^2$$

We use mini-batch gradient descent method to learn W and b.

3. Exercise 1

3.1 Check gradients

By computing the relative error between a numerically computed gradient value g_n and an analytically computed gradient value g_a

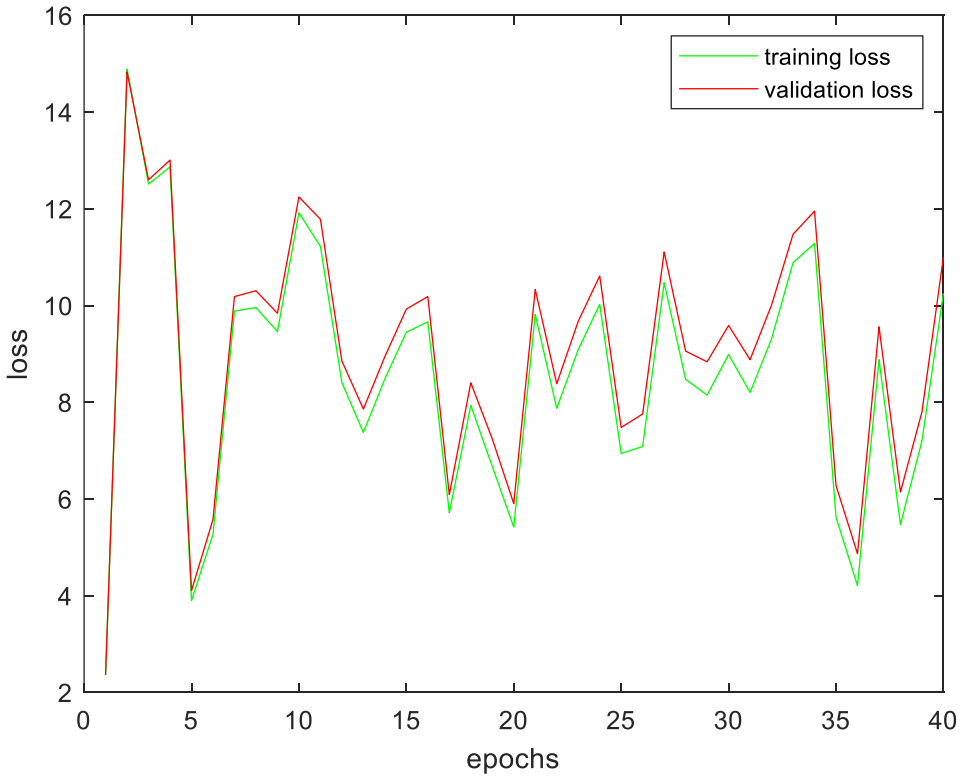
$$\frac{|g_a - g_n|}{\max(eps, |g_a| + |g_n|)}$$

If this value is small, we can believe the calculation is correct.
During the training process, I get relative error of W is less than 10^{-3}, relative error of b is less than 10^{-6}.

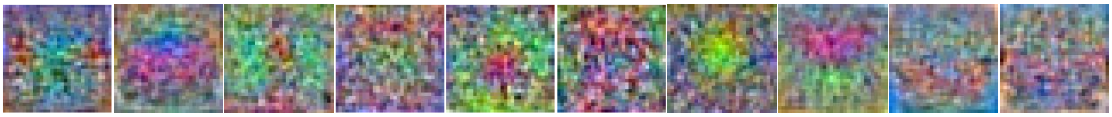
3.2 Check loss, accuracy and weights visualization

No regulation:
lambda=0, n epochs=40, n batch=100, eta=.1

- Loss



- Weights:

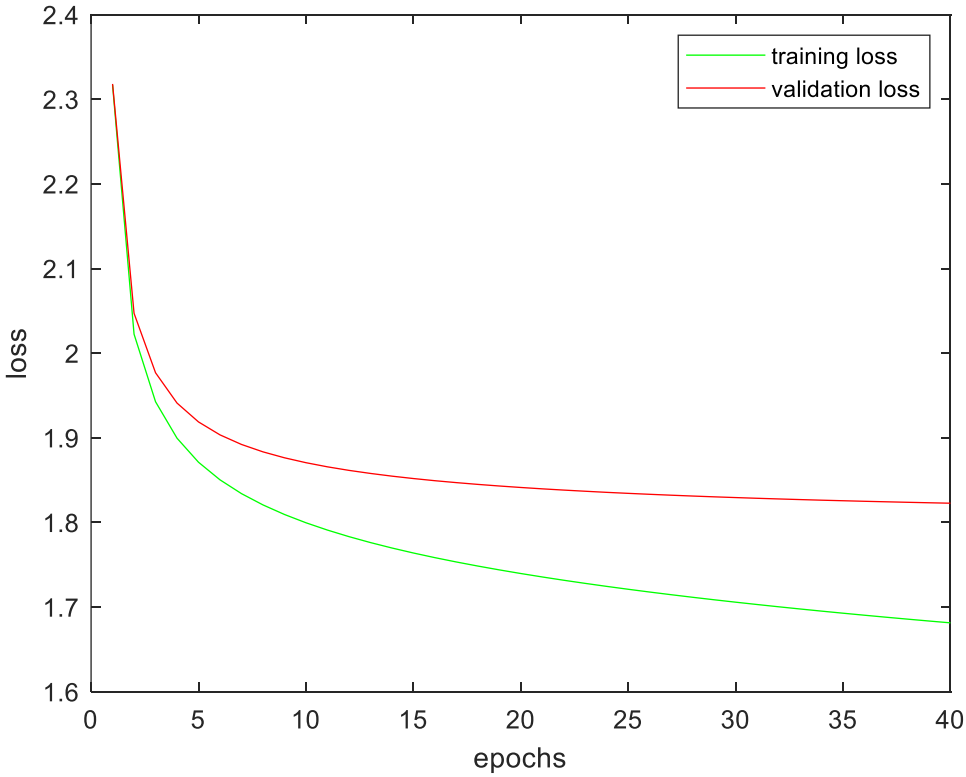


- Accuracy:

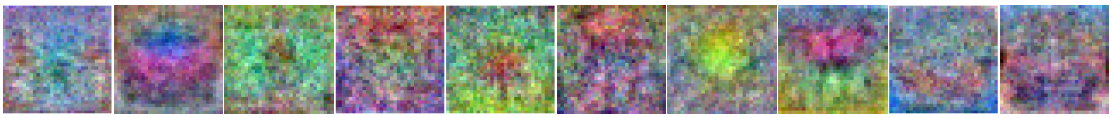
Training Accuracy:28.81%
Test Accuracy:24.78%

lambda=0, n epochs=40, n batch=100, eta=.01

- Loss



- Weights:

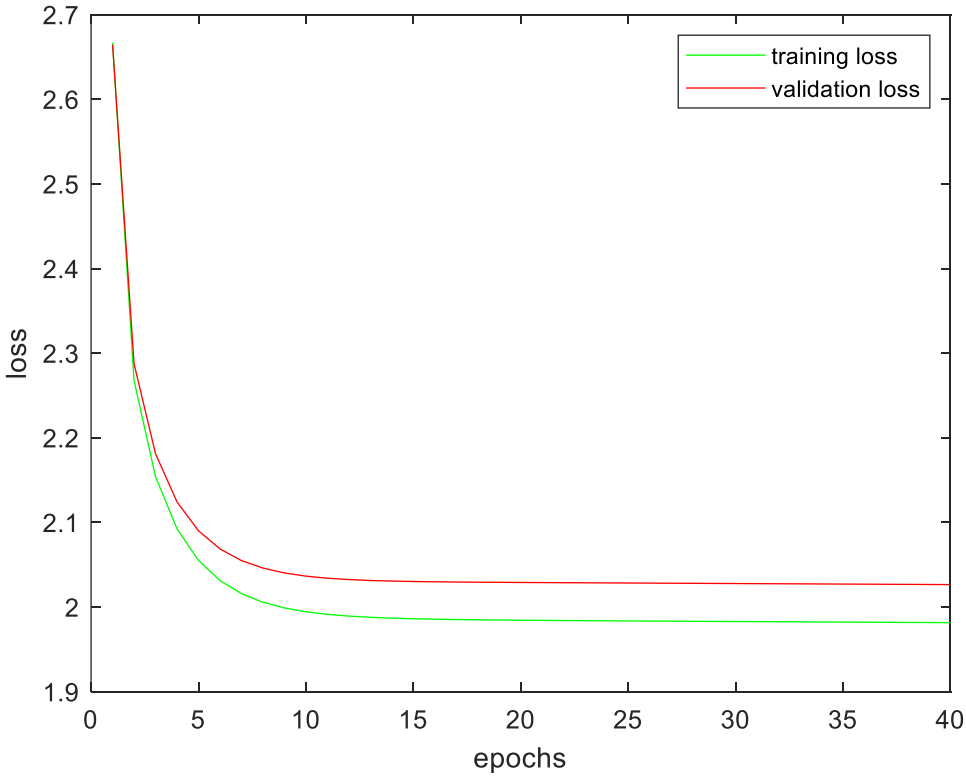


- Accuracy:

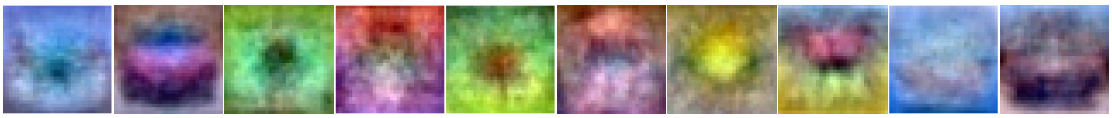
Training Accuracy: 41.6%
Test Accuracy: 36.87%

With Regulation:
lambda=0.1, n epochs=40, n batch=100, eta=.01

- Loss



- Weights:

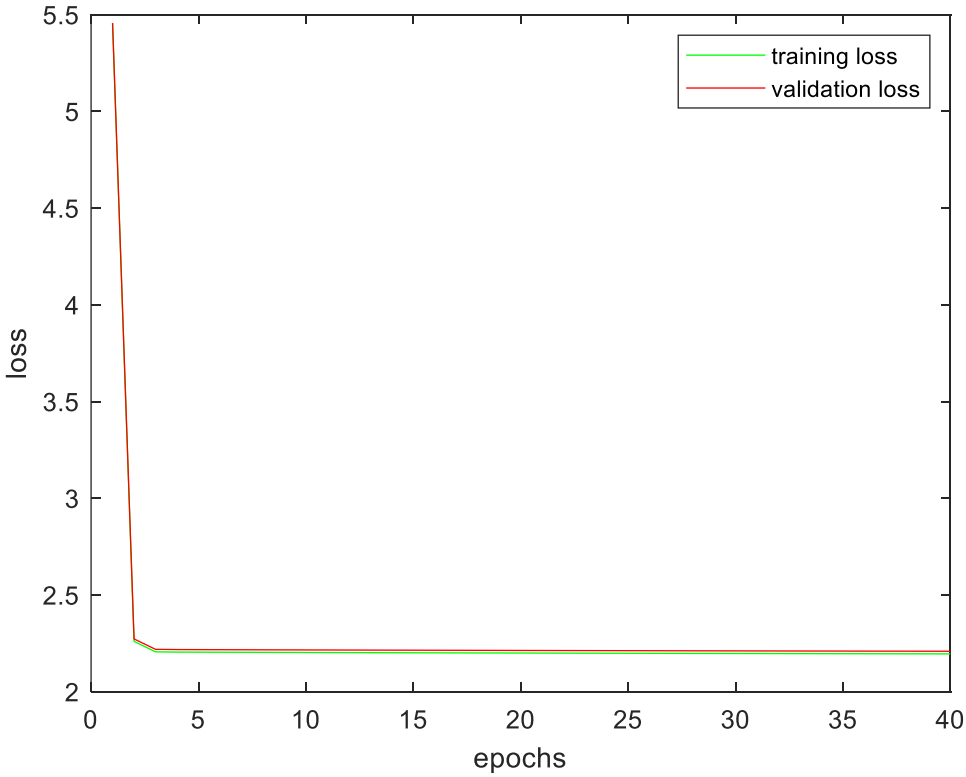


- Accuracy:

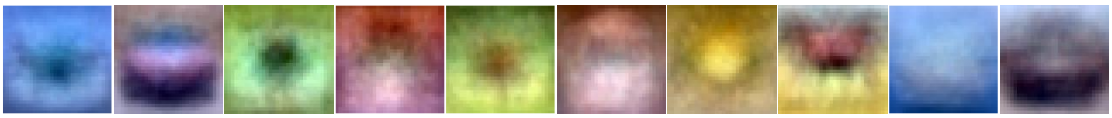
Training Accuracy: 34.19%
Test Accuracy: 33.38%

lambda=1, n epochs=40, n batch=100, eta=.01

- Loss



- Weights:



- Accuracy:

Training Accuracy: 22.27%
Test Accuracy: 21.92%

4. Conclusions

1. The learning rate will affect the learning process. High learning rate will lead to a fast convergence, but if it is set too high, it might jump over the local minima and make the learning process unstable.
2. The regularization term can effectively reduce overfitting. Nevertheless, too high λ may decrease W, that may increase the bias.