DD2424 Assignment 2

Levent Güner
leventg@kth.se

Introduction

This assignment is about implementing two-layer Neural Network.

Data

CIFAR-10 dataset is used, which has 10 class labels representing different objects and 10,000 different 32x32 images per class.

Methods

There is a two-layer neural network generated and tested with different configurations. The learning rate is using cyclical learning rate. λ is optimized using course and fine search method.

Cyclic learning rate method makes the learning rate changing between η_{min} and η_{max} over determined step number. During the learning process $\eta_{min} = 10^{-5}$ and $\eta_{max} = 10^{-1}$ values are taken.

Results

```
grad(W1)
--------
Analytic grad(W1): 0.815013016313191
Numerical grad(W1): 0.8150130162309921
Sum of absolute diff.: 1.16e-08

grad(W2)
--------
Analytic grad(W2): -0.08988912929373827
Numerical grad(W2): -0.08988912969520868
Sum of absolute diff.: 5.65e-09

grad(b1)
--------
Analytic grad(b1): 0.14604829285871845
Numerical grad(b1): 0.14604829294828636
Sum of absolute diff.: 3.58e-10

grad(b2)
--------
Analytic grad(b2): 2.3592239273284576e-16
Numerical grad(b2): 2.7755575615628914e-17
Sum of absolute diff.: 6.76e-11
```

Table 1: Gradient results

I have checked the gradients with Centered Difference Formula and got pretty accurate results can be seen in Table 1.

Levent Güner
leventg@kth.se

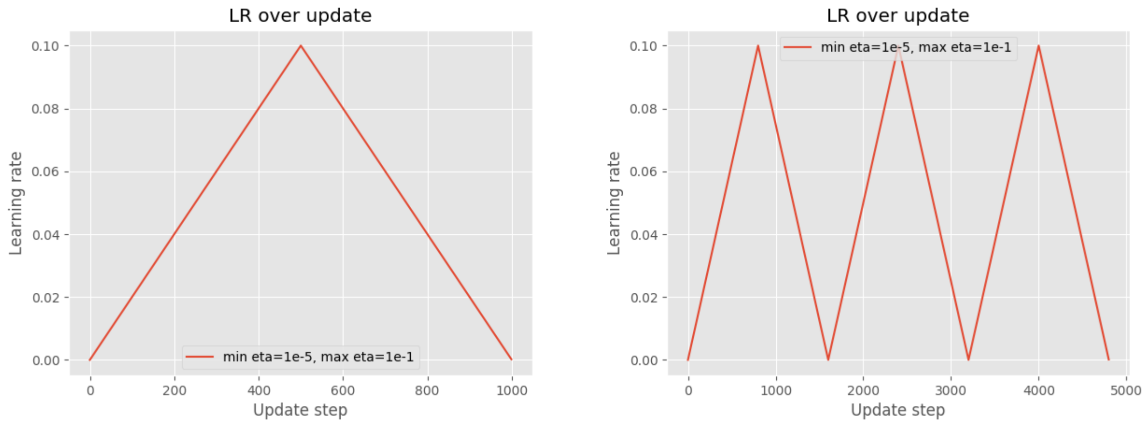
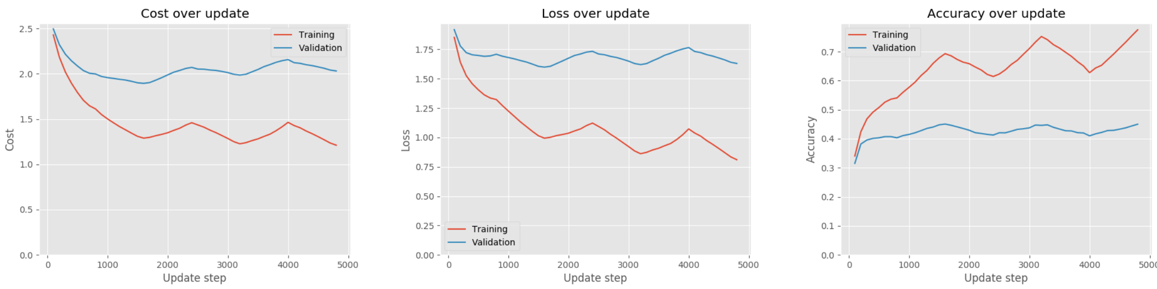


Figure 1: Learning rates over update step for different configurations

Figure 1 shows how cyclical learning rate changes learning rate for each update. The cycle is adjusted to 3 and η is changing between 10^{-5} and 10^{-1}.

1 cycle, 500 steps



3 cycles, 800 steps

Figure 2: Training curves (cost, loss, accuracy) for different configurations

Figure 2 shows the cost, loss and accuracy curves for training processes. The difference of cyclic eta can be seen easily on graphs. The total epoch number is generated by $\frac{2 \times no.cycles \times no.steps}{batch\ size}$. Which makes 10 epochs for the first process and 48 epochs for the second one.

λ	η_{min}	η_{max}	No. cycles	No. of steps	Final testing accuracy
0.01	10^{-5}	10^{-1}	1	500	0.45
0.01	10^{-5}	10^{-1}	3	800	0.46

Table 2: Accuracies for different configurations

Levent Güner
leventg@kth.se

Table 2 shows different λ and learning rate configurations and their results. The accuracy increases when we increase number of steps and η cycles.

Course and Fine Search

λ	Accuracy
2.26e-3	0.4992
1.46e-6	0.5034
1.22e-3	0.5048
2.21e-3	0.507
1.17e-7	**0.5084**
8.19e-4	0.5042
5.76e-7	0.5070
5.05e-6	0.50
1.59e-6	0.5022
2.29e-6	0.5022

Table 3: Course search

λ	Accuracy
0.0	0.5028
1.30e-8	0.5028
2.60e-8	0.5024
3.90e-8	0.5004
5.19e-8	0.4986
6.49e-8	0.4986
7.79e-8	0.5078
9.09e-8	0.5078
1.04e-7	0.5084
1.17e-7	**0.5084**

Table 4: Fine search

Table 3 and shows course and fine search results for 1 cycle. With the best λ result, 3 and 4 cycles were tested and the following results are yield:

2 cycles:
Validation accuracy: 52.8%, Test accuracy: 49.94%

3 cycles:
Validation accuracy: 52.6%, Test accuracy: 49.6%

4 cycles:

Validation accuracy: 52.4%, Test accuracy: 49.18%

Conclusion and Discussions

In this assignment, there is a double-layer neural network model trained with different optimizations. In comparison to single-layer neural network model, the effects of λ and learning rate are seen more.

As mentioned in Assignment 1, L2 regularization factor, which is determined by λ can be seen here:

$$\lambda \sum_{i,j} {W_{ij}}^2$$

During the processes I experienced the importance of λ more, specially data size/λ magnitude effect. As one can see from the results, overfitting starts when we increase the cyclic learning rate cycle. I have found the best results with λ=1.17e-7 and 2 cycles of learning rate.

The model obviously gave better results than a single-layer model. However, when there are more optimizations are made it is more tend to give even better results.