DD2424 Deep Learning in Data Science
Assignment 2

Gianluigi Silvestri
giasil@kth.se 940106-4614

April 18, 2018

1 Check the gradients

To test the correctness of the analytical derivatives I have used the centered difference gradient for W_1 ,b_1,W_2 and b_2, and then compared the results as suggested in the assignment with:

$$\frac{| g_a - g_n |}{\max(eps, | g_a | + | g_n |)} \tag{1}$$

where eps is $1\exp(-7)$ and the result has to be smaller than $1\exp(-3)$. Another way I used to test the gradient was, as suggested in the assignment, to train my network on a 100 training data without regularization and learning rate $\eta = 0.1$. After training for 200 epochs, the obtained loss was 0, as shown in figure 1.

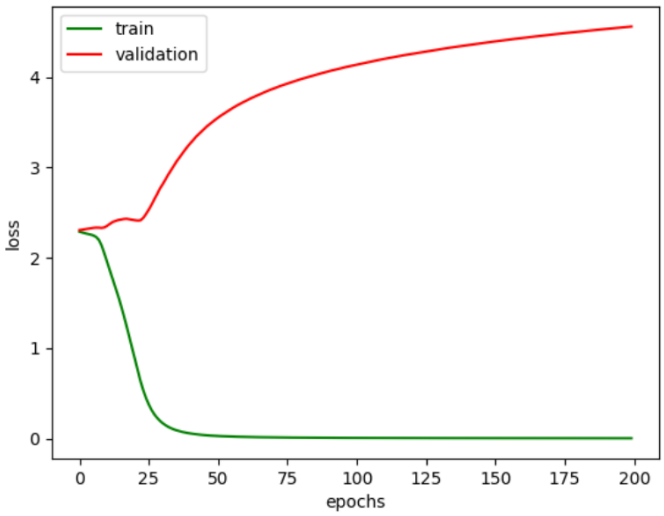

Figure 1:

2 Momentum

I have tested the momentum using one CFAIR batch, with 10 epochs, a batch size of 100, $\eta = 0.01$ and no regularization.The application of a momentum=0.9 makes the training faster, as shown in figure 2.

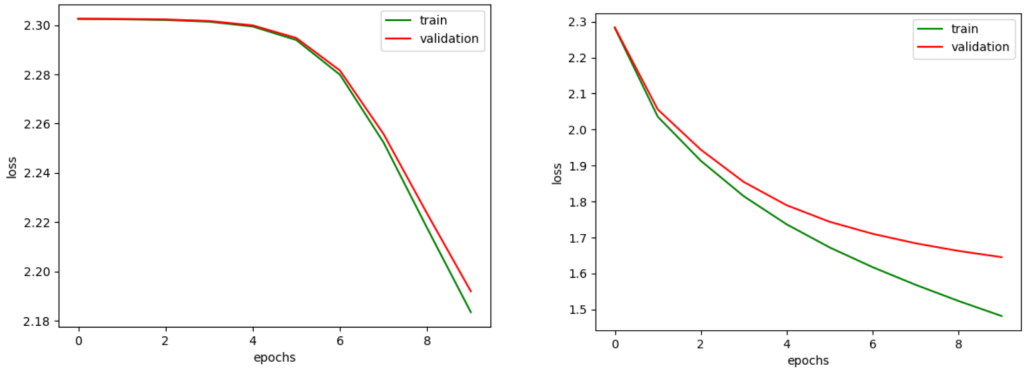

Figure 2: Left: no momentum. Right: momentum = 0.9

3 Coarse search

In all the searches I have used the first CFAIR batch, a batch size in the training of 100, weight decay of 0.92 for each epoch, and a momentum of 0.9. In the first experiment with 10 epochs, I didn't use regularization, and the learning rate values were sampled between 0.1 and $10 \exp -5$. The 3 best results are shown in figure 3.

0.0243997705263	43.83
0.0208232434544	43.84
0.0335802100624	43.85

Figure 3: The left column contains the learning rate, while the right columns contains the accuracy.

In the second experiment, the search for learning rate was limited between 0.01 and 0.001, and the search for regularization term λ was introduced with values between 0.1 and $10 \exp -6$. The 3 best results on 10 epochs are shown in figure 4.

0.0365531209119	0.00324155520282	44.0
0.0421860816558	0.00223324103752	44.08
0.0315283145411	0.00138810207887	44.13

Figure 4: The left column contains the learning rate, the central one contains the regularization term, and the right columns contains the accuracy.

Since many good results were obtained also with a very small λ, another search has been done with the same values but increasing the number of epochs to 20. In this way, the results were more dependent on the on the regularization term. The results are shown in figure 5.

0.0221705781239	7.23888094158e-06	44.93
0.0736139020259	0.00331719907997	44.99
0.0278576043574	0.00272339262514	45.06

Figure 5: The left column contains the learning rate, the central one contains the regularization term, and the right columns contains the accuracy.

In the new results, there is still some good combination obtained with a small λ. For the fine search, however, I have decided to exclude those values, since they were associated with small learning rates and with more epochs may lead to overfitting.

4 Fine Search

The last search has been done with 200 samples between 0.01 and 0.001 for η and between 0.001 and 0.0001 for λ. The training was executed for 30 epochs. The results are shown in figure 6.

0.0554883268965	0.00498454585669	46.35
0.0730628223498	0.00461635610078	46.43
0.084124905016	0.00262003632199	46.61

Figure 6: The left column contains the learning rate, the central one contains the regularization term, and the right columns contains the accuracy.

Using the parameters that gave the best results in the last search, I couldn't obtain an accuracy grater than 44% with 10 epochs as required in the assignment. However, due to the weight decay, I prefer to use a bigger initial learning rate that might be beneficial if the training is done for a lot of epochs.

5 Training with all the data

The training as been done for 30 epochs. The parameters were: $\lambda \approx 0.00262$, $\eta \approx 0.084$, $\rho = 0.9$, weight decay=0.92 per epoch. The accuracy was 55.38% for the training set, 50.8% for the validation and 51.97% for the test set. The loss results are shown in figure 7, and it is possible to see how probably training for more epochs would lead to a better result.

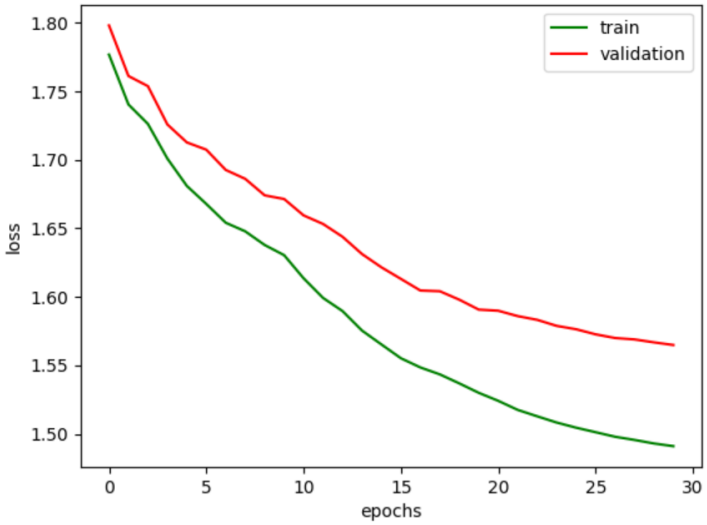
train
validation
loss
epochs
1.80
1.75
1.70
1.65
1.60
1.55
1.50
0
5
10
15
20
25
30

Figure 7: