Deep Learning in Data Science - Assignment 2 - Bonus Tasks

Fernando García Sanz

April 6, 2020

Abstract

The scope of this assignment consists in building and training a two layer network which has multiple outputs. This network is used to classify the images contained in the CIFAR-10 [1] dataset. The network is trained using mini-batch gradient descent over a cost function which computes the cross-entropy loss of the classifier applied to the labelled training data and an L_2 regularization term over the weight matrix. Bonus tasks can be found in this document.

1 Optimize the performance of the network

Three different methods have been employed to improve the performance of the network.

1.1 Exhaustive random search

Do a more exhaustive random search to find good values for the amount of regularization, the length of the cycles, number of cycles, etc.

To accomplish this task, the number of lambda values which are considered has been doubled, going from eight to sixteen. Moreover, the number of η update cycles employed have been increased too, using four cycles in both the coarse search and the fine search.

As it can be seen, the results obtained by means of this procedure slightly improve those obtained employing the configuration without the optimization:

- Best accuracies in validation: [0.526, 0.5246, 0.5238]
- Best lambdas in validation ($\log_{10}$ scale): [-2.88736766, -2.82894768, -2.7488288]
- Improved accuracies in validation: [0.528, 0.526, 0.526]
- Improved lambdas in validation ($\log_{10}$ scale): [-2.8472087, -2.88736766, -2.83107602]

Using the best lambda obtained during the search process, and performing a training process in which 49000 samples are used for training and just 1000 for validation, the results have been the following:

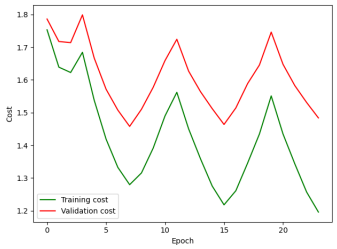

Figure 1: Cost over training epochs.

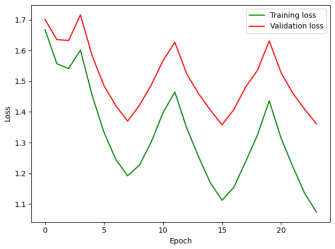

Figure 2: Loss over training epochs.

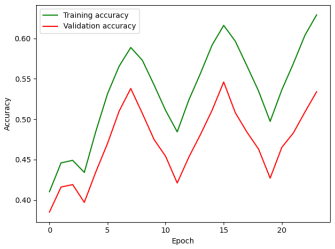

Figure 3: Accuracy over training epochs.

The accuracy obtained over the test set has been equal to Accuracy on test data: 52.5%, being slightly better than the one obtained without this more exhaustive search. Even though, the value obtained without the improvement was equal to 52.21%, suggesting that performing a more exhaustive search will not significantly suppose a great enhancement of the network performance.

1.2 Increment of the hidden layer number of nodes

You could also explore whether having more hidden nodes improves the final classification rate. One would expect that with more hidden nodes then the amount of regularization would have to increase.

Increasing the number of nodes that compose the hidden layer supposes a more proficient behaviour of the network. As more nodes exist, more features of the dataset can be described, although it is necessary to be careful, as the overfitting risk also grows.

Using 100 nodes returns better results when training the network. Using the same procedure as before, implementing the cyclical η and the lambda search, the outcomes have been the following:

- Best accuracies in validation: [0.5314, 0.5298, 0.5286]
- Best lambdas in validation ($\log_{10}$ scale): [-4.7154769, -3.94603659, -2.37877684]
- Improved accuracies in validation: [0.5406, 0.5392, 0.5368]
- Improved lambdas in validation ($\log_{10}$ scale): [-4.53748595, -4.31180968, -4.55533657]

As it can be seen, the results obtained in the validation set are better than before. Nevertheless, the best values of lambda turn out to be lower than before. Therefore, the features represented by the different nodes are highly considered, and not soften by means of regularization.

When finally training with the best value of lambda, the results have been the following:

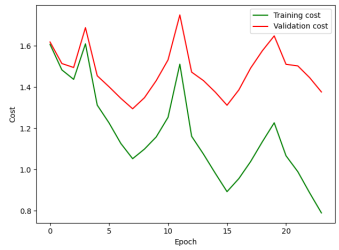

Figure 4: Cost over training epochs.

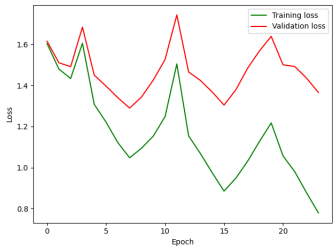

Figure 5: Loss over training epochs.

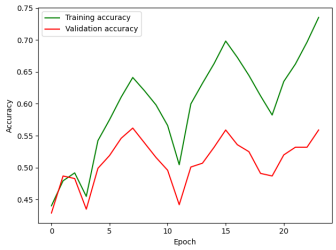

Figure 6: Accuracy over training epochs.

The accuracy obtained over the test set has been equal to 52.61%, slightly better than the one obtained without this optimization.

1.3 Dropout

Apply dropout to your training if you have a high number of hidden nodes and you feel you need more regularization.

Dropout mechanism only makes sense when the number of neurons is big enough, since otherwise it might worsen the network performance. Therefore, a total number of 100 nodes has been employed in the hidden layer.

Performing a training process in which cyclical η and lambda search are employed, without performing dropout, provides the following results:

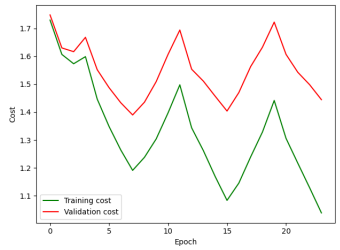

Figure 7: Cost over training epochs.

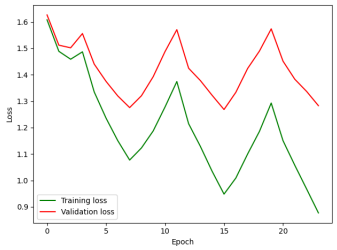

Figure 8: Loss over training epochs.

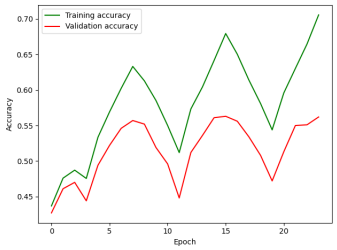

Figure 9: Accuracy over training epochs.

When dropout is used, applying a probability of a 20%, the following results are retrieved:

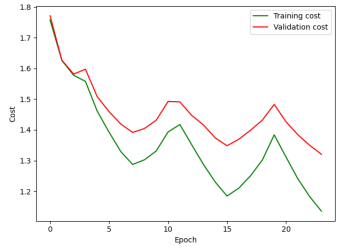

Figure 10: Cost over training epochs.

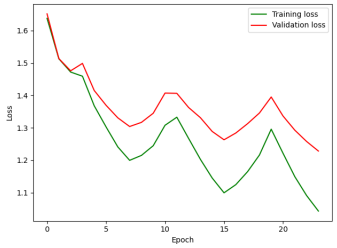

Figure 11: Loss over training epochs.

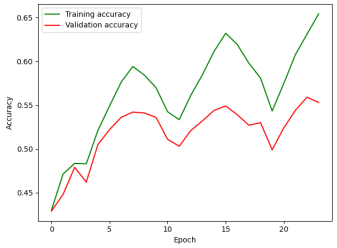

Figure 12: Accuracy over training epochs.

The accuracy obtained without employing dropout has been equal to 53.54%. When using dropout, accuracy has been equal to 54.22%. As it can be seen, employing dropout has resulted in an improvement of almost a 1% in the accuracy obtained over the test set, keeping the same configuration settings.

2 Look for η_{min} and η_{max} according to *Smith, 2015*

In order to compute the best values of η as stated in [2], we define a sufficiently large interval of values η can take, and then compute the accuracy obtained over the test set of each network trained with each respective value of η.

To perform this computation, the minimum value, η_{min}, has been set to ≈ 0, and the maximum value, η_{max}, to 0.18. A hundred values from this interval separated by a step of $(\eta_{max} - \eta_{min})/100$ has been employed. The accuracies obtained are represented in the graph on the right.

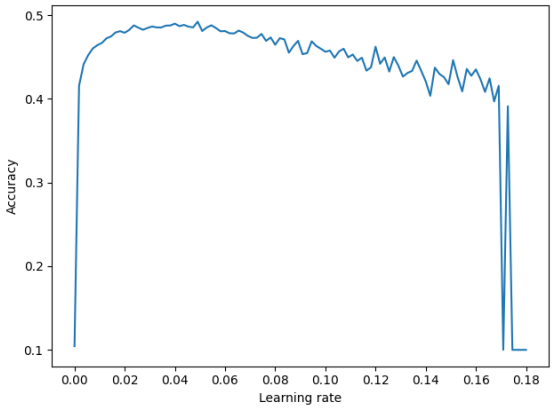

Figure 13: Accuracy over test set given η.

To properly choose the values for η_{min} and η_{max}, the first one should be set when the accuracy starts to increase, and the second one, once the accuracy starts to stabilize. Therefore, a value close to zero, as the suggested in the assignment, 1e-5, should suffice. The accuracy seems to be stabilized around 0.04 - 0.05, so setting a value which belongs to this interval, being in this case 0.05 the chosen one, should be acceptable.

Using these original boundaries of the interval as η_{min} and η_{max} returns the following:

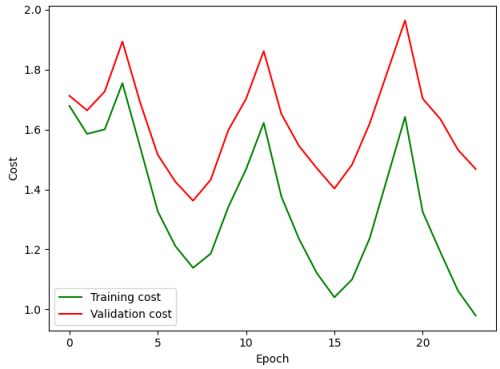

Figure 14: Cost over training epochs.

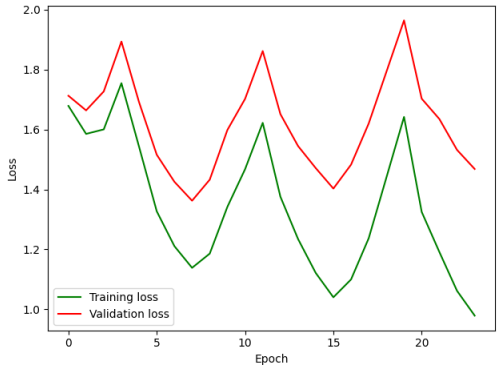

Figure 15: Loss over training epochs.

When using the found values as η_{min} and η_{max}, the results are as follows:

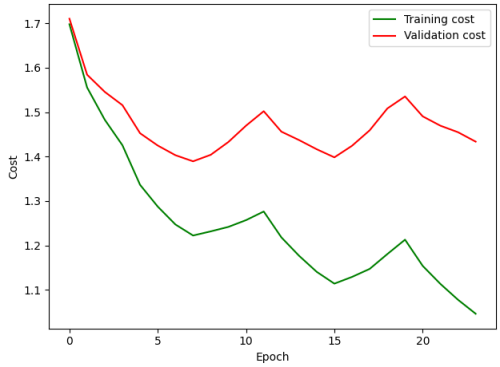

Figure 16: Cost over training epochs.

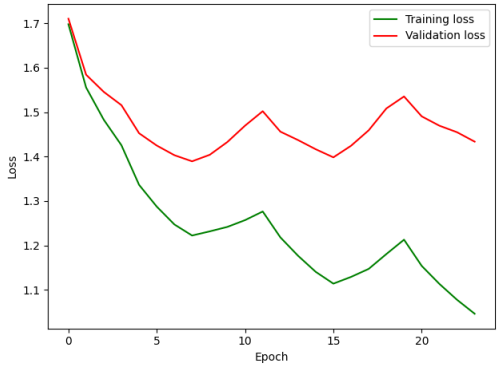

Figure 17: Loss over training epochs.

As it can be seen, the shapes of the plots are not that steeped in the second ones, as a consequence of a reduced interval of values defined by the new η_{min} and η_{max}.

References

[1] A. Krizhevsky. The cifar-10 dataset. `https://www.cs.toronto.edu/~kriz/cifar.html`. Accessed: 21-03-2020.

[2] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.