COMP6248 Lab 4 Exercise

Adrian Azzareli, aab1g18@soton.ac.uk

1 Investigating Wide MLPs o MNIST data

To accomplish this lab a simple MLP was derived (using the *torchbearer* library); parameters are provided in Table 1.

Parameter	Value
Input Dimensions, N	784 (28x28 Image)
Output Dimensions, M	10
Loss, L	Cross Entropy
Optimiser	Adam
Learning rate, lr	0.01
Momentum, p	0.9

Table 1. Parameter Settings for simple MLP

The lab asks us to evaluate the performance of the MLP while varying the number of nodes in a network to determine 'when' a simple model will *overfit*. A large factor to overfitting is the time (number of epochs) we assign to an MLP to train. Considering this, we aim to observe the points at which varying sizes of MLPs begin to overfit and we will accomplish this by evaluating the results the training accuracy vs testing accuracy for each model size across a range of epochs.
The dimensions of all tested networks is provided in Table 2.

Network #	1	2	3	4	5	6	7
Size, n	100	500	1000	2000	5000	10000	20000

Table 2. Dimensions of the tested network: 784-n-10

We initially monitored the testing and training behaviours of all network variants over 50 epochs by measuring both loss and accuracy, as illustrated in Figure 1. These results are interesting as the loss shows a convergence to around -1.58 for all network sizes, whereas the results for accuracy show a divergence from around 0.0198 after epoch 10 for most network sizes.

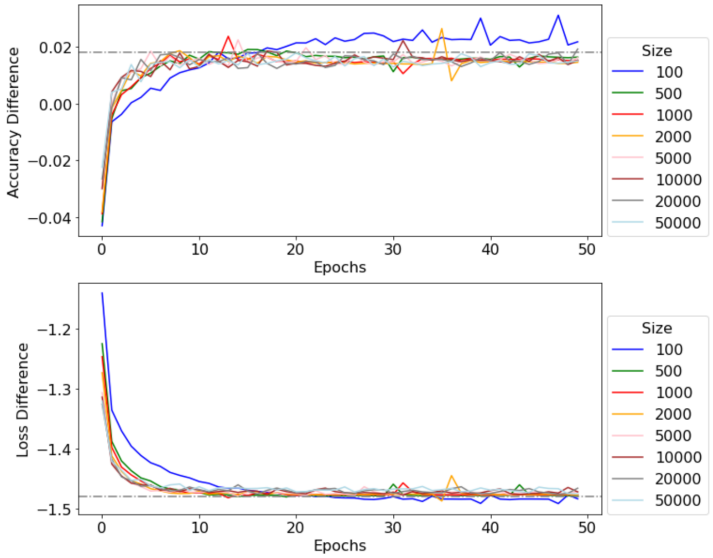

Fig. 1. Difference between training and testing accuracy/loss over 50 epochs

As expected the larger networks converge to incrementally smaller points (shy of -1.58 and 0.0198). Considering that overfitting can be though of as a deviation from some optimum number training-testing difference, we can say, using Figure 1, that overfitting does occur. Though, evidently it does not *devastate* the network[1]. The caveat is that with larger networks, the point of divergence (w.r.t to the accuracy plot) occurs later on. Comparing $n = 500$ and $n = 50000$, Figure 2, we see both the difference in peaks ($n = 500$ has a larger max-difference in training-testing accuracies) and rate of divergence of these peaks ($n = 50000$ has a slow divergence while $n = 500$ diverges slighly quicker after epoch 12. This behaviour is rationalised by larger networks needing more epochs to train, yet being more robust (considering the increase in features the network may be able to differentiate).

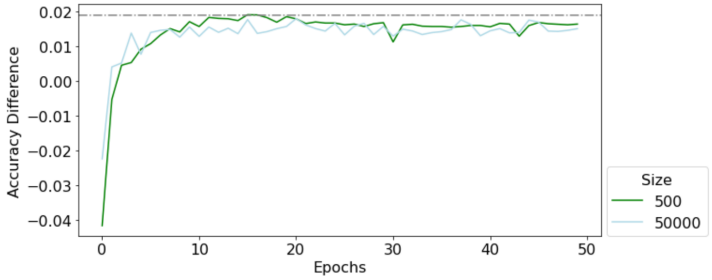

Fig. 2. Difference between training and testing accuracy over 50 epochs

[1]This issue w.r.t MNIST is discussed here: https://datascience.stackexchange.com/questions/19874/why-doesnt-overfitting-devastate-neural-networks-for-mnist-cla fbclid=IwAR1ecHw0cc1C5O9GjvJFICyiPiord0Kn50aXnvxXaGN84xTM5kDvnF