UE Machine Learning: Supervised Techniques

Exercise 8 Report

Name: Omar Amr

Matrikel-Nr: k11776960

1. Variable Importance and Model Selection Procedure:

A. Main structures and variable importance:

Two main types of neural networks are trained in this assignment. The first abstract architecture of the first neural network is as follows:

1. Input layer with 300 nodes, a node for each character in the encoded sequence.
2. A hidden layer with variable number of nodes. The accuracy of each neural network is computed and compared together in order to come up with the best number of nodes.
3. Output layer with 1 node, which represents the model's classification for the input.

The second abstract architecture consists of:

1. Input layer with 300 nodes, a node for each character in the encoded sequence.
2. A hidden layer with variable number of nodes. The accuracy of each neural network is computed and compared together in order to come up with the best number of nodes.
3. Another hidden layer with also variable number of nodes to find the best model.
4. Output layer with 1 node, which represents the model's classification for the input.

Concerning the parameters of the neural network that affects the performance of the models dramatically, two main parameters are put into consideration:

1. Learning rate: which indicated how quickly the neural network changes its model depending on the values of new inputs. Higher learning rate makes the network change its weights more that lower learning rates.
2. Dropout rate: this variable protects the neural network from overfitting by ignoring some nodes in the process of changing weights. In this way, the network's weights will not over-fit the data because of the weights that was not updated according to the new input.

B. Model Selection Procedure:

The search strategy followed to find the best combination of parameters is grid search where each variable is assigned a set of values. Then All the combinations are tried together, and the accuracy of the neural network is calculated every time and compared to previous results. Eventually the parameters that bring about the best accuracy are used to train the final model.

Concerning neural network with one hidden layer, the following parameters are included in the grid search:

1. Number of nodes per hidden layer (5 values are tested).
2. Learning rate (5 values are tested).
3. Dropout rate (5 values are tested.

Which sums up to 125 tests, the neural network with best accuracy is chosen at the end of the search.

Concerning neural network with Two hidden layers, the exact same parameters are included in the search in addition to another variable that represents the number of nodes in the second hidden layer. This sums up to 625 tests. And as before, the neural network with best accuracy is chosen at the end of the search.

2. Model Validation and Evaluation:

Validation sets are used to evaluate the performance of neural network during the search process. Before descripting the details of the steps of validation, two terms need to be clarified first:

A. Epoch: an epoch is when the data is passed through the network forward and backward once.

B. Batch size: since the data might be very big to be used by the network at once, the data is divided into smaller chunks called batches. Batch size determine the number of input data entries that are used to validate the model once.

Many epochs are used in order to guarantee that the network updates its weight in a sufficient manner as usually one pass is not enough. Using too few epochs

causes the model to under-fit the data and using too many causes the model to over-fit the data.

100 Epochs are used to validate the model with batch size of 40. The accuracy of the last batch during the model training is very close to its previous batches which means overfitting is not considerably noticed using these parameters.

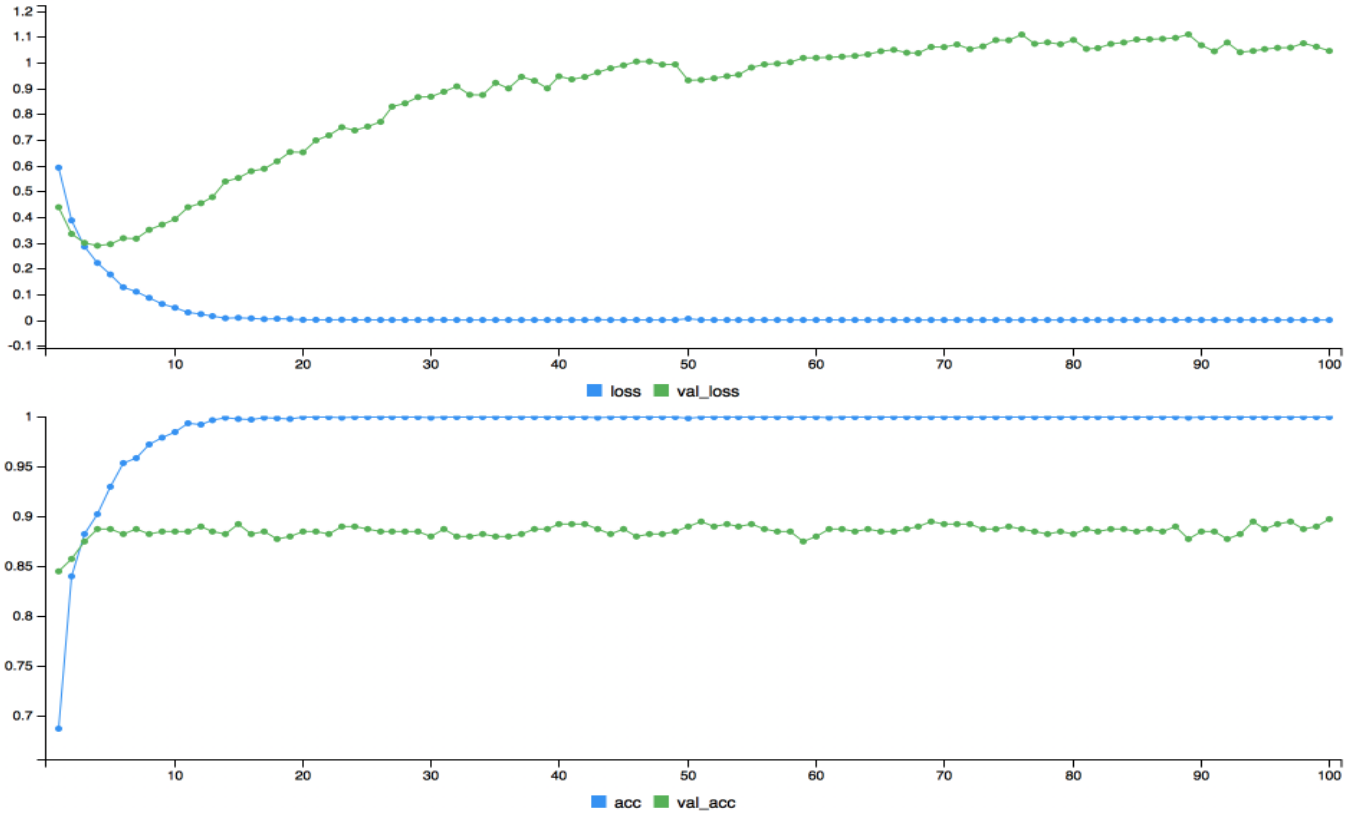
3. Summary of results:

The final models chosen by the grid search, representing the best models, for both the one hidden layer model and the two hidden layers model have very close accuracy values. The following table shows the parameters that produces the best models:

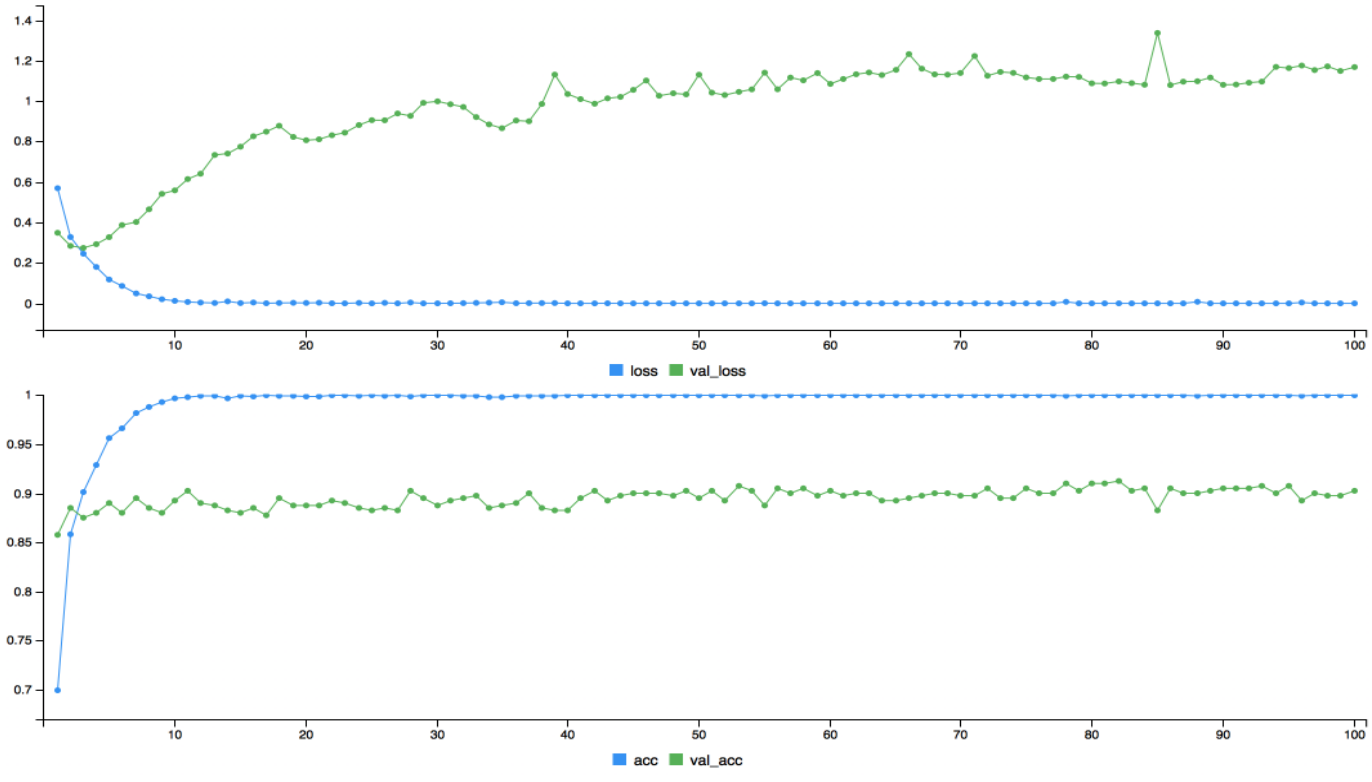
Input Layer	1st Hidden Layer	2nd Hidden Layer	Output Layer	Learning Rate	Dropout Rate	Model Accuracy
300	100	--	1	0.001	0.4	90%
300	200	100	1	0.001	0.3	90.5%
100 epochs with batch size of 40 are used to train both models.						

The graphs in the next page show the accuracy of both models during the training process. The green line in lower half of each figure shows the accuracy of each model during each epoch.

Neural Network with 1 hidden layer Accuracy graph:



Neural Network with 2 hidden layers Accuracy graph:



4. Summary on Preferred Model and Limitations:

The neural network with 2 hidden layers has better accuracy than the one with only 1 hidden layer. However, I personally prefer the model trained with only one hidden layer. The reasons behind that are:

A. By comparing the "val_acc" graph for each model in the figures above, one can see that the graph for the 1-hidden-layer network is more consistent and more settled towards the end of the training.
B. The "val_acc" graph for the 2-hidden-layer network might show a better final accuracy, but considerably viewable outliers can be noticed in the graph. P.S. One might think this happens because of the learning rate for the 2-hidden-layer model is larger than the other model. In fact, the two models that achieved the best results are trained with the same learning rate.
C. The difference between the accuracy of the 2 model is very small. Only 0.5% difference. On the other hand, the 1-hidden-layer model fulfils values of accuracy with low variance every time. The 2-hidden-layer model might get a very high value at one time but the next time the model is retrained, the variance is pretty notable.

More justifications why more hidden layers does not mean better accuracy:

1. Since the 2-hidden-layer neural network has more nodes, this means it has more random weights at the beginning of the training. The dataset is not large to adequate the training of the model.
2. Layers far from the output layer learns slower than the once closer to it. This means that the second layer adjusts its values quicker than the first one. And again, as a result of the dataset not being large enough, the first layer does not get the chance to adjust its parameters to the best possible way.