COMP 5212 HA5 - BERT

Leung Pak Hei 20690382

April 2023

1 Introduction

Default Baseline performance: The default output with bert(sentid, attention mask=mask, return dict=False), Linear (1286), Relu(), Linear(768,512), Relu(), dropout(0.1), Relu(), Linear(512,2), log softmax(output=2).The maximum length of the sequence will use is 15 The baseline weighted accuracy is 0.82, which will then be used as the benchmark weighted accuracy.
In the following, different similar neural network structures will be compared using different parameters such as max length sequence, learning rate, neural network, and dropout rate.

2 Comparison using different neural network layers:

This section will explore the weighted accuracy of the prediction using different numbers of hidden layers in the same epoch with a learning rate =0.1. The optimizer is SGD. The result is as follows

Comparison using different neural network layers (others hold equal)	
Neural Network Structure	Weighted Accuracy
(Baseline structure)	0.82
remove one relu (726,512)	0.72
Add one more relu layer before softmax	0.86
Add two more relu layer before softmax	0.8
Add three more relu layers before softmax with big size	0.66

From the results, it is clear that increasing more number of layers does not nessarily increase the weighted accuracy of the BERT model. It is suspected that neural network with more hidden layers tend to have higher chance of overfitting so the accuracy for the testing data is lower. Similarly, with too few layers, the model may underperform as it does not capture enough details. However, it may be also affected by the size of the layer and other parameters.

3 Comparison using different dropout rate

This section will explore the weighted accuracy of the prediction using different number dropout rate in the same epoch with a learning rate =0.1. The optimizer is SGD. The result is as follows

Comparison using different dropout rate	
dropout rate	Weighted Accuracy
0.1	0.84
0.2	0.86
0.3	0.82
0.8	0.58
0.9	0.54

From the above experiments, given the default neural network, learning rate, and batch size, etc, the lower dropou rate tends to have better accuracy on the sentence. However, if the dropout rate is too high, the model may suffer from underfitting due to very strong effect of dropout.

4 Comparison using different max length of the text sequence

This section will explore the weighted accuracy of the prediction using different numbers of max length of the text sequence in the same epoch with a learning rate =0.1. The optimizer is SGD. The result is as follows

Comparison using different numbers of maximum length sequence (others hold equal)	
maximum length sequence	Accuracy
5	0.68
10	0.76
15	0.84
20	0.86
25	0.89

From the above experiments, given the default neural network, learning rate, and batch size, etc, the maximum length sequence of the sentence trained tends to have better accuracy on the sentence and smaller optimum sentence length can reduce the training time of the BERT.

5 Comparison using Different numbers of learning rate

This section will explore the weighted accuracy of the prediction using different numbers of hidden layers in the same epoch. The optimizer is SGD. The result is as follows

Comparison using different learning rate (others hold equal)	
Learning rate + same hidden layers	Accuracy
0.05	0.9
0.1	0.88
0.5	0.84
1	0.74
5	0.62

The above results show that the accuracy tends to decrease with a larger learning rate. However, the training time of the BERT is much higher in the low learning rate case.