Assignment 4

Advanced Machine Learning - Text Data

Purpose:

The purpose of this assignment is to apply RNNs to text and sequence data. It is Demonstrate how to improve performance of the network, especially when dealing with limited data and to check which approaches are more suitable for prediction improvement. Specifically, answer the questions provided in the assignment.

Dataset:

Here we are using the data set which is mentioned in the assignment which is the data set of IMDB.

Firstly, the test accuracy of predictions of the whole data set is observed followed by having various observations by using different samples.

We first note the value of the initially given code and various modifications will be made to this base code to find different results by using techniques pf embedding layer, padding, and masking and pretrained embedding layer: Glove to observe the approach that works best.

Analysis:
For this assignment, first the code given in the module is executed.
Once the initial code results are observed, the code is updated with the changes asked in question.

The following changes have been made to the original file:
1)Cutoff reviews after 150 words
2) Restrict training samples to 100
3) Validate on 10,000 samples
4) Consider only the top 10,000 words
5) Before the layers.Bidirectional layer, consider a) an embedding layer, and
b) a pretrained word embedding
The below questions have been answered.
Which approach works better?
2. Now try changing the number of training samples to determine at what point the embedding layer gives better performance.
After making the necessary changes the code is observed with different samples ranging from 250,650,1050,5050,9050.

Observations:

First the test accuracy of the initial file is observed as below.

	Test Accuracy			
Program Model	Basic Sequence	Embedding layer	Padding and Masking	Pre-trained Embedding layer
Initial	0.87	0.86	0.86	0.87

This gives the accuracy of prediction of 86% for the embedding layer and pretrained embedding layer gives 87% which is higher than basic embedding layer when taking all the samples present in the dataset.

Case I:

The code is updated with the following changes.
Cut off: 150.
Samples: 100
Words: 10000
Validation samples:10000
Pre-trained: GloVe

Updated Cut off: 150. Samples: 100 Words: 10000 Validation samples:10000 Pre-trained: GloVe	0.81	0.79	0.78	0.77

This gives the accuracy of prediction of 78% for the embedding layer and pretrained embedding layer gives 77% which is lesser than basic embedding layer when taking 100 samples present in the dataset.

Case 2:

The code is updated with the following changes.
Cut off: 150.
Samples: 250,650,1050,5050,9050
Words: 10000
Validation samples:10000
Pre-trained: GloVe
Here the model is trained with different samples and test accuracy is observed.

Updated Cut off: 150. Samples: 250 Words: 10000 Validation samples:10000 Pre-trained: GloVe	0.81	0.78	0.79	0.77

Updated Cut off: 150. Samples: 650 Words: 10000 Validation samples:10000 Pre-trained: GloVe	0.80	0.78	0.79	0.76
Updated Cut off: 150. Samples: 1050 Words: 10000 Validation samples:10000 Pre-trained: GloVe	0.77	0.78	0.80	0.77
Updated Cut off: 150. Samples: 5050 Words: 10000 Validation samples:10000 Pre-trained: GloVe	0.80	0.78	0.79	0.77

From the results we can see that accuracy of prediction of 78% for the embedding layer and pretrained embedding layer gives 77% which is lesser than basic embedding layer when taking different samples present in the dataset.

It can be observed that while testing on 650 samples there is a slight decrease in pretrained embedding layer which gives 76% accuracy. However, there is no change in embedding layer which gave 78% accuracy. Therefore, at this point embedding layer is giving better performance than the pre-trained embedding layer.

Conclusion:

From the observations, it is evident that among the original embedding layer an Pre-trained embedding layer, the original embedding layer approach is better as it give 78% of accuracy than the latter. Its performance increases with the rise of samples and works efficiently from 650 samples.