Assignment 4- Text Data

IMDB dataset of movie reviews, labeled as either positive or negative is used. We have implemented the sentiment analysis model using a LSTM neural network. The dataset is preprocessed by vectorizing the text data, and the following preprocessing steps are taken:

- Cutoff reviews after 150 words
- Restrict training samples to 100.
- Validate 10,000 samples.
- Consider only the top 10,000 words.

Q.1.a

First basic sequence model:

The model architecture consists of a one-hot encoding layer, followed by a LSTM layer with 32 hidden units. The output of the LSTM layer is passed through a dropout layer with a rate of 0.5, and then through a dense layer with a single unit and a sigmoid activation function.

The model is trained using the RMSprop optimizer and binary cross-entropy loss function, with accuracy as the evaluation metric. The test accuracy is reported as 0.780. The validation accuracy achieved during training is reported as 0.7612.

Embedding layer:

The model is trained using the same hyperparameters as before with an additional embedding layer, except for the number of epochs which is set to 50. The callbacks are also the same as before. After training, the model is evaluated on the test dataset, resulting in a test accuracy of 0.799.

This improvement in test accuracy over the previous model is likely due to the addition of the embedding layer, which allows the model to learn more from the input words. The embedding layer provides a way to encode into the dense vector representation, which can improve the ability of the model to generalize to new, unseen data.

pretrained Embedding layer:

The model is then trained Using a pretrained Embedding layer on the training dataset, validated on the validation dataset, and evaluated on the test dataset. The results show that the model achieves a test accuracy of 0.786.

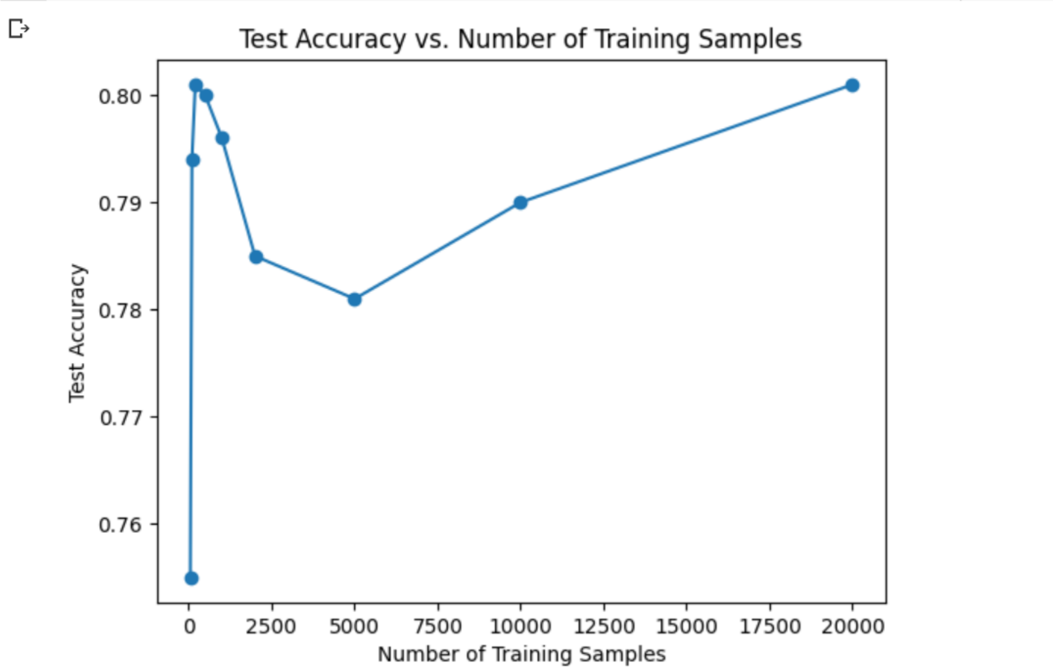
Embedding layer is the best method in terms of test accuracy because it achieved the highest test accuracy of 0.799. This improvement in test accuracy over the first method or base method is likely due to the addition of the embedding layer, which allows the model to learn more from the input words. Pretrained embedding layer achieved a slightly lower test accuracy of 0.786 compared to the second method.

Q.2.a

Effect of Training Sample Size on Embedding Layer Performance:

```
num_train_samples = [50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000]

test_accs = [0.755, 0.794, 0.801, 0.800, 0.796, 0.785, 0.781, 0.790, 0.801
]
```



Based on the above graph, we can determine at what point the embedding layer gives better performance. we can see that the accuracy score starts improving and reaches its peak at 200 training samples. After 200 training samples, the accuracy score starts decreasing slightly as the number of training samples increases. This suggests that 200 training samples are sufficient to achieve good performance, and increasing the number of training samples may not necessarily improve the model's performance.

Conclusion

The use of an embedding layer significantly improves the performance as shown by the increase in test accuracy from 0.780 for the base model to 0.799 for the model with an embedding layer. Furthermore, using a pretrained embedding layer yields a slightly lower test accuracy of 0.786. The effect of training sample size on the performance of the model with an embedding layer suggests that 200 training samples are sufficient to achieve good performance, and increasing the number of training samples may not necessarily improve the model's performance.