ASSIGNMENT 3

TEXT AND SEQUENCE

NITHIN VARMA

MOTIVE: To get a better interpretation of different parameters that influence the performance of the model, and to check which adaptations work under various conditions. Especially, the changes include limiting the number of training samples, and taking only the top 10,000 words, and differentiating the performance of a model using the embedding layer to one that uses a pre-trained model.

Training sample sizes and their results

Models	Training Sample Size	Scratch or pre-trained	Test accuracy	Test Loss
1		Scratch	0.842	0.342
2	100	Scratch	0.531	0.703
3	100	Pre-trained	0.526	3.4
4	1000	Embedding layer	0.572	0.694
5	15000	Embedding layer and convID	0.826	0.412
6	30000	Embedding layer and convID	0.813	0.443
7	15000	Pre- trained	0.514	3.9
8	30000	Pre- Trained	0.523	0.93

- Differentiating the model accuracy of 1 and 2, where no enhancements were made, it achieved a test accuracy of 84% , stating that it was precisely classifying reviews as either positive or negative based on text content. And , reviews of the train were = 100 samples , the accuracy of the model decreased crucially , test accuracy of 53%
- The training samples was changed to determine when the embedding layer under-performed the pre-trained word embedding. The embedding layer fared better with 1,000 training samples, obtaining a test accuracy of 54% , contrasted to a test accuracy of 52% with the pre-trained word embedding at training sample 100.
- Using cutoff reviews at 150 words, a training sample of 100, validation on 10,000 samples, and just the top 10,000 words considered, both models with an embedding layer and a pre-trained word embedding achieved nearly the same test accuracy of 52%.
- An increase in the training sample size had a little impact on the accuracy where we have used the convID with the embedding layers also increasing the training sample size 15,000 and 30,000 . Resulted in a better accuracy that is 82% and 81%
- . Using 15000 and 30000 training data, the embedding layer and Conv1d models beat the pre-trained word embedding test accuracies.

Conclusion

Therefore, the accuracies state that the performance of the model is proportional to the parameters used, including the total training samples, the max review length, etc. In whole, when you are using an embedding layer while working on smaller datasets, it may be more effective.