**Assignment 4 - Text and Sequence**

**-Tejaswini Kolli**

Used binary\_crossentropy loss function, accuracy metric and rmsprop optimizer to learn how different training samples affect the performance of the models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.no | maxlen | Training Set | Validation Set | Test Set | Embedding layer | Pretrained Network | Loss and Accuracy on Test |
| 1 | 150 | 10000 | 10000 | 5000 | Yes | Yes | loss: 0.7875 - acc: 0.7972 |
| 2 | 150 | 8000 | 10000 | 5000 | Yes | Yes | loss: 0.9575 - acc: 0.5472 |
| 3 | 150 | 5000 | 10000 | 5000 | Yes | Yes | loss: 1.9113 - acc: 0.3644 |
| 4 | 150 | 500 | 10000 | 5000 | Yes | Yes | loss: 0.5067 - acc: 0.7488 |
| 5 | 150 | 100 | 5000 | 5000 | Yes | Yes | loss: 0.5220 - acc: 0.9102 |
| 6 | 150 | 100 | 10000 | 5000 | Yes | Yes | loss: 0.1585 - acc: 0.9866 |
| 7 | 150 | 100 | 10000 | 5000 | No | Yes | loss: 0.2808 - acc: 0.8902 |

**Summary:**

* The validation set size is kept constant at 10,000 and Test 5,000 across most experiments, allowing for consistent evaluation during training.
* The training set is changed from the range of 100 to 10,000
* Training set 100 has the lowest loss of 0.1585 and the maximum accuracy of 0.9866; as training sets increased, accuracy decreased, and loss increased.

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

* Training set 5000 has the highest loss of 1.9113 and the minimum accuracy of 0.3644

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

* The combination of Training set 100, Validation set 10,000 which does not use an embedding layer, still achieves reasonably good accuracy, indicating that the model can learn effectively even without embeddings.

**Conclusion:**

Neural network models for text classification require careful consideration of pretrained networks, embeddings, and training set size to achieve excellent performance. These factors provide insightful information that may be used to improve and fine-tune models in a variety of scenarios. Pretrained networks, particularly those trained on large datasets, enable the model to capture complex linguistic patterns and semantic representations. These learnt representations, extracted from large amounts of textual input, are useful for tasks like text classification. Embeddings are essential for providing machine learning models with context and word meaning. They allow for a more sophisticated comprehension of the relationships between words, which improves the model's text categorization abilities. A key aspect affecting a model's ability to generalize is the size of the training set.

**Recommendations:**

* Smaller training sets can still yield good results when combined with proper embedding and pretrained networks.
* The use of embeddings and pretrained networks contributes positively to the model's ability to generalize.