Text and Sequence Assignment 4 Report

Introduction:

This assignment focuses on applying Recurrent Neural Networks (RNNs) or Transformers to text and sequence data, specifically the IMDB movie review dataset. The primary objectives are to explore the performance of these models on text data, examine techniques for improving performance when dealing with limited data, and determine the most suitable approaches for prediction improvement.

Data Preprocessing:

Describe the steps taken to preprocess the IMDB dataset, including:

- Cutting off reviews after 150 words
- Restricting the training set to 100 samples
- Validating on 10,000 samples
- Considering only the top 10,000 words in the vocabulary

Methodology:

Baseline Model:

- Describe the baseline model used (e.g., RNN with an embedding layer)
- Explain the model architecture and hyperparameters
- Report the validation and test accuracy/loss for the baseline model

Pretrained Word Embeddings:

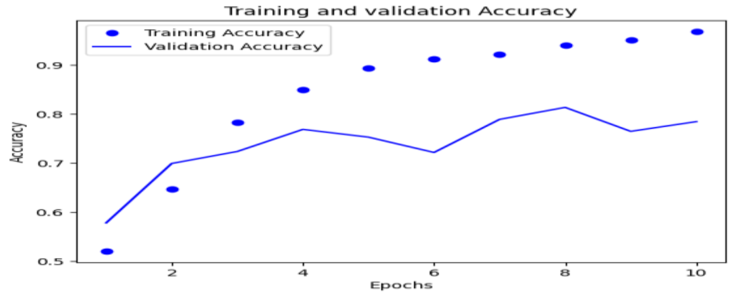
- Explain the use of pretrained word embeddings (e.g., GloVe)
- Describe the process of loading and integrating the pretrained embeddings
- Report the validation and test accuracy/loss for the model with pretrained embeddings

Varying Training Set Size:

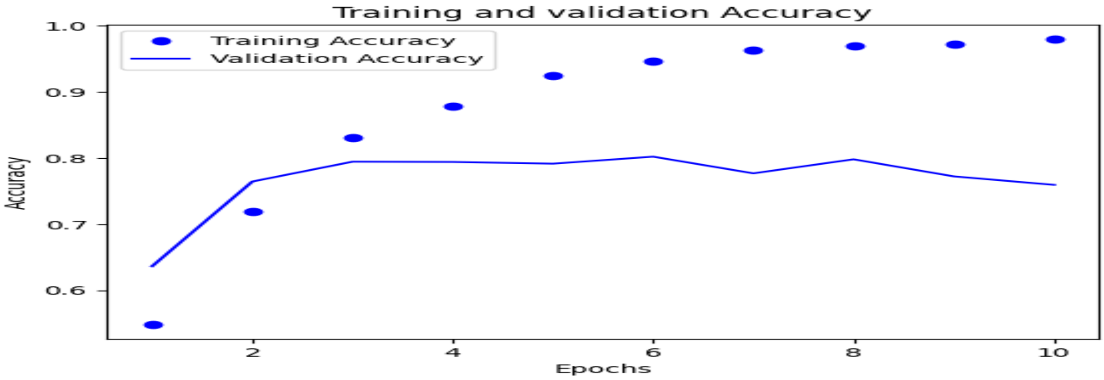
- Describe the process of varying the number of training samples
- Report the validation and test accuracy/loss for different training set sizes
- Analyze the performance of the embedding layer vs. pretrained embeddings at different training set sizes

Results:

One Hot model: A One Hot model achieves a Validation Accuracy of 0.78 and loss of 0.44.

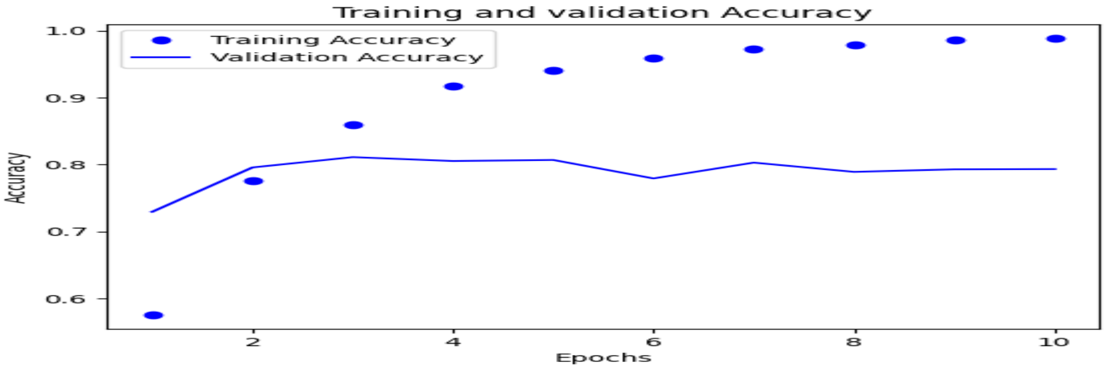


Trainable Embedding Layer: A Trainable Embedding Layer achieves a validation Accuracy of 0.75 and a loss of 0.46.



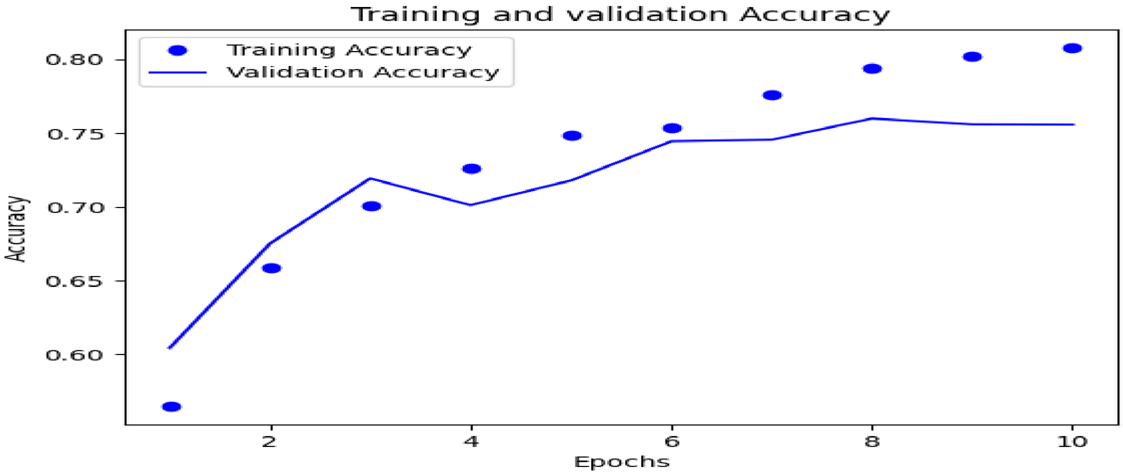
Masking Padded Sequences in the Embedding Layer:

A Masking Padded Sequences in the Embedding Layer achieves a validation Accuracy of 0.759 and a loss of 0.43.



Model with Pretrained GloVe Embeddings:

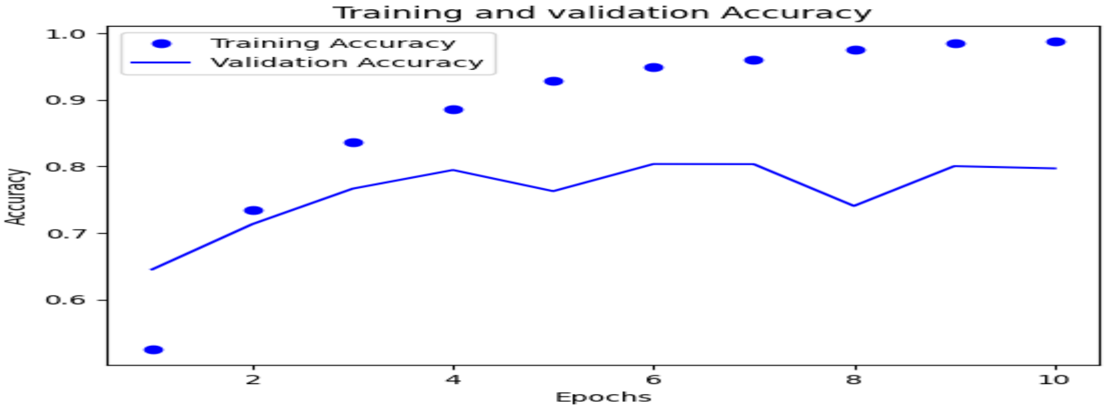
A Model with Pretrained GloVe Embeddings achieves a validation Accuracy of 0.7 5and a loss of 0.49.



Comparing Model Performance with Different Training Set Sizes

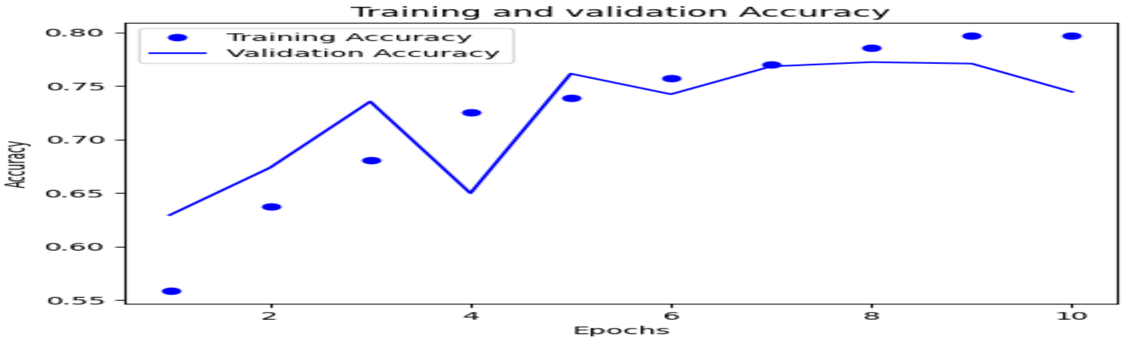
Embedding Layer 100 Training Samples:

A Embedding Layer with 100 Training Samples achieves a validation Accuracy of 0.79 and a loss of 0.47.



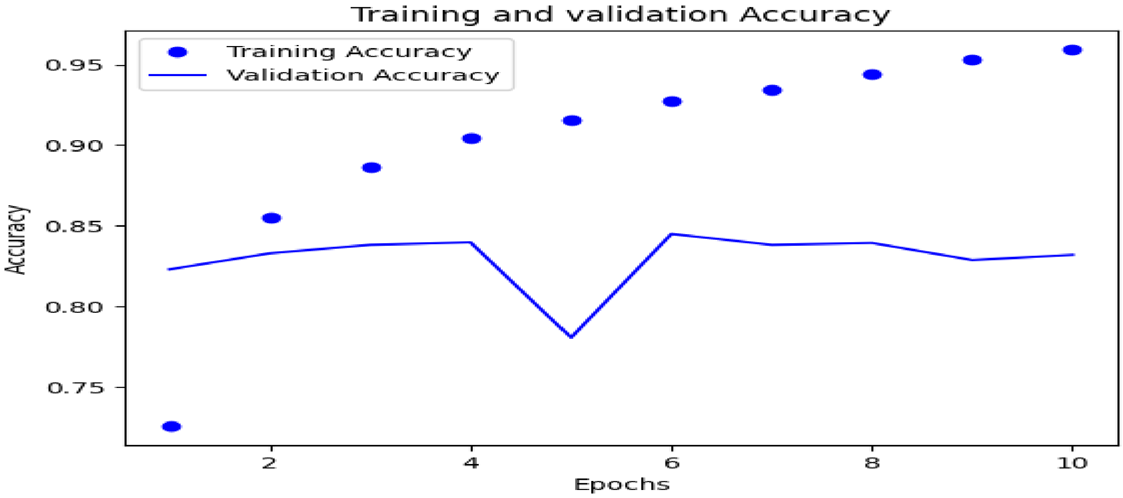
Pretrained Embedding Layer 100 Training Samples:

A Pretrained Embedding Layer with 100 Training Samples achieves a validation Accuracy of 0.74 and a loss of 0.43.



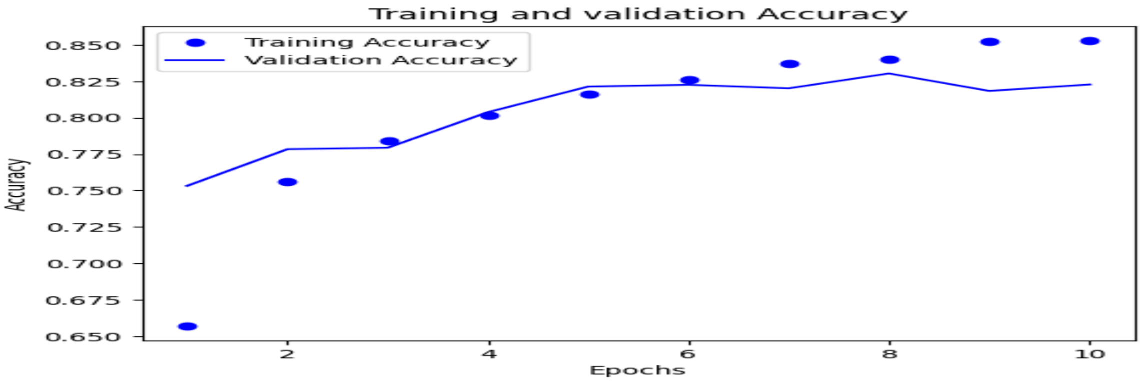
Embedding Layer 500 Training Samples:

A Embedding Layer with 500 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.39.



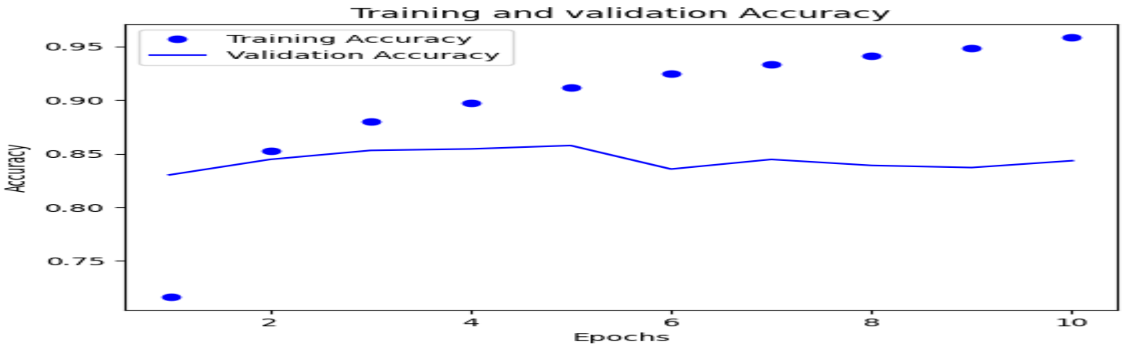
Pretrained Embedding Layer 500 Training Samples:

A Pretrained Embedding Layer with 500 Training Samples achieves a validation Accuracy of 0.82 and a loss of 0.38.



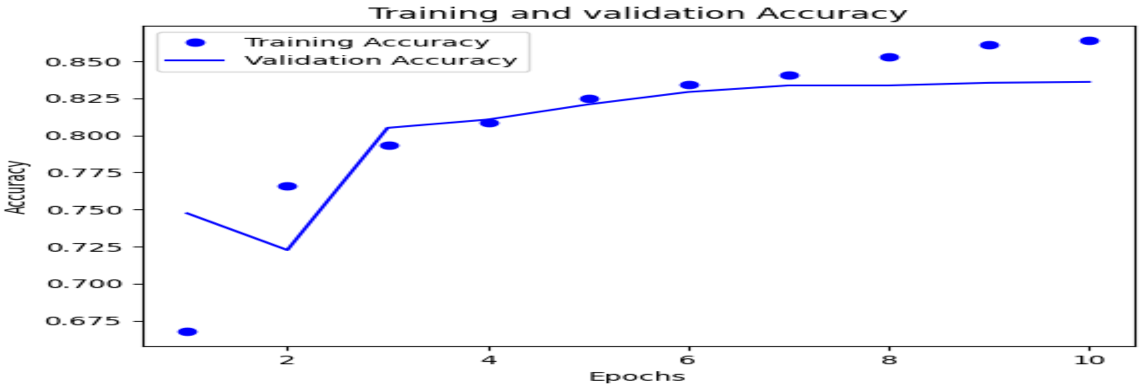
Embedding Layer 1000 Training Samples:

A Embedding Layer with 1000 Training Samples achieves a validation Accuracy of 0.84 and a loss of 0.37.



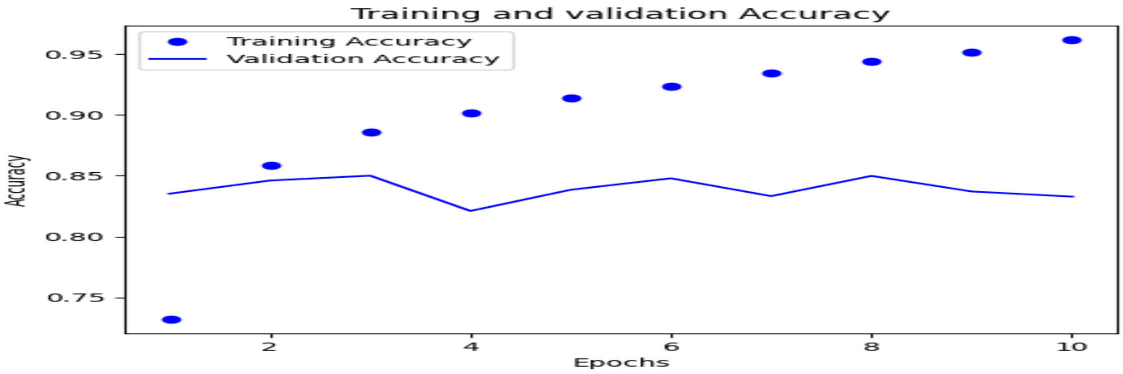
Pretrained Embedding Layer 1000 Training Samples:

A Pretrained Embedding Layer with 1000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.36.



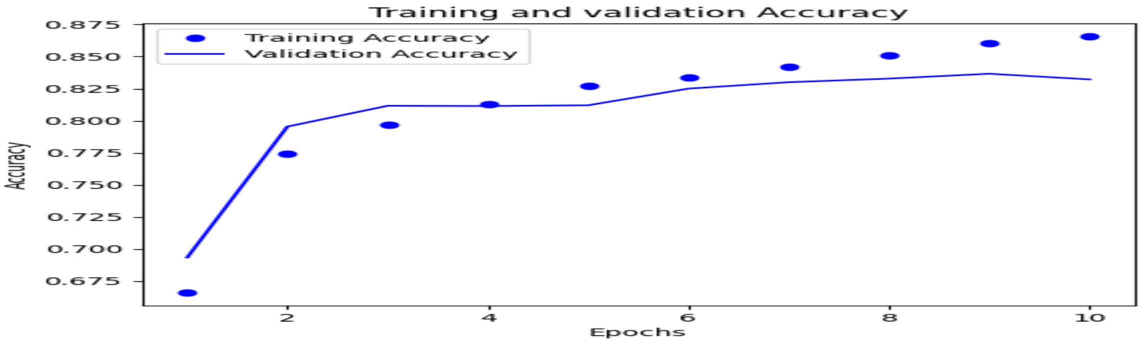
Embedding Layer 5000 Training Samples:

A Embedding Layer with 5000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.37.



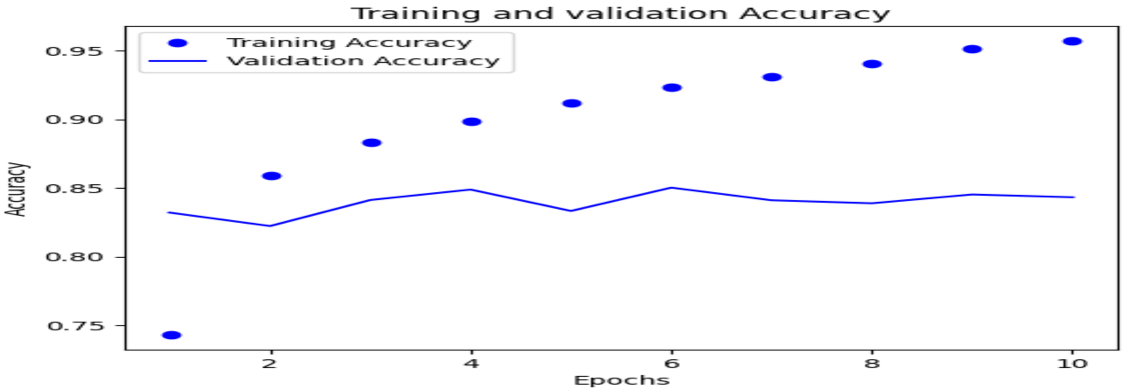
Pretrained Embedding Layer 5000 Training Samples:

A Pretrained Embedding Layer with 5000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.37.



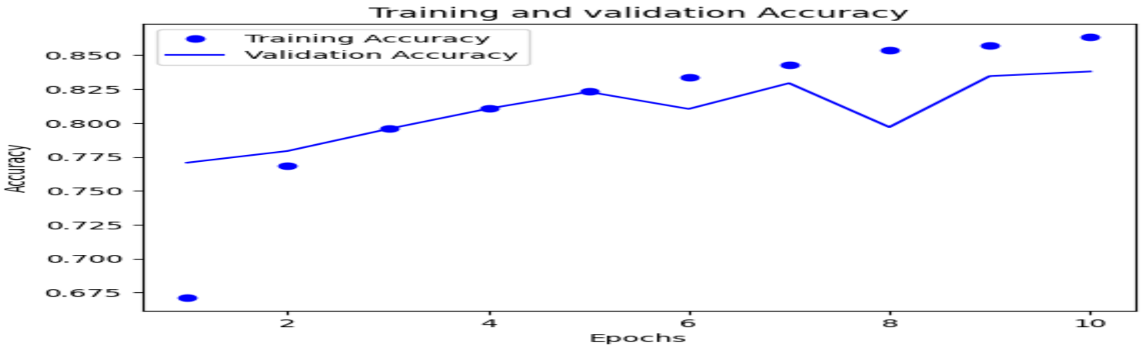
Embedding Layer 10000 Training Samples:

A Embedding Layer with 10000 Training Samples achieves a validation Accuracy of 0.84 and a loss of 0.40.



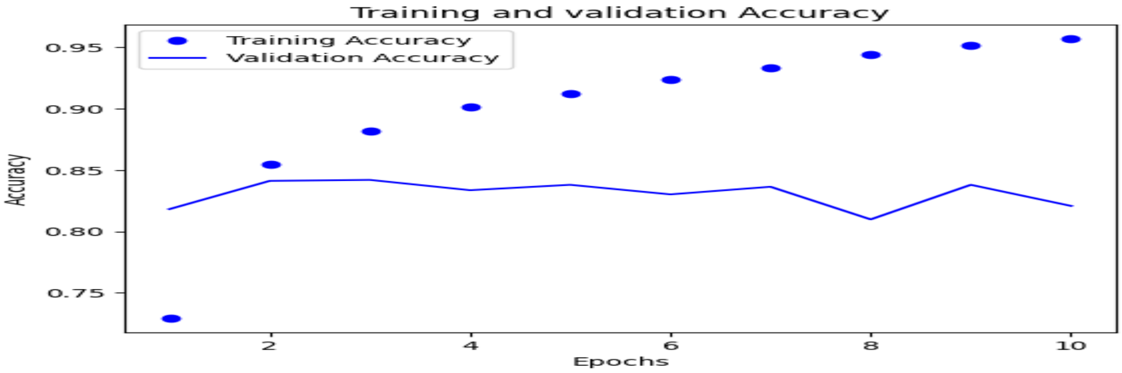
Pretrained Embedding Layer 10000 Training Samples:

A Pretrained Embedding Layer with 10000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.37.



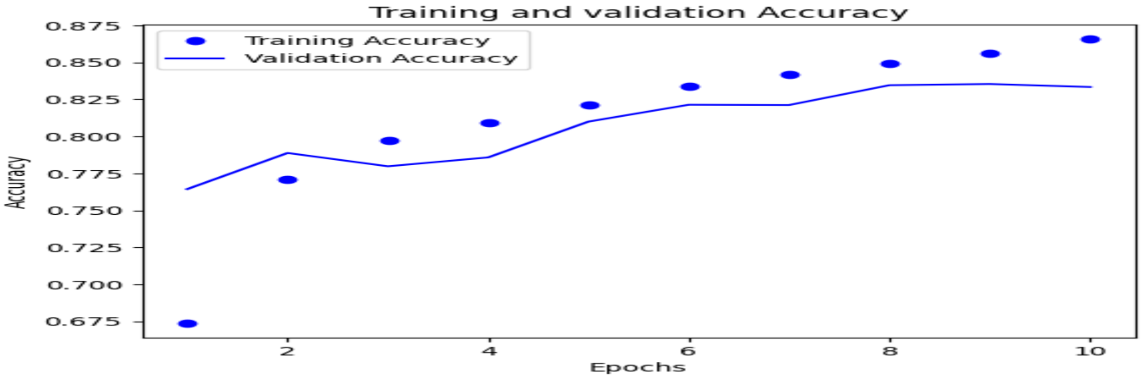
Embedding Layer 20000 Training Samples:

A Embedding Layer with 20000 Training Samples achieves a validation Accuracy of 0.82 and a loss of 0.38.



Pretrained Embedding Layer 20000 Training Samples:

A Pretrained Embedding Layer with 20000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.35.



Results Table:

Model	**Validation Accuracy**	**Loss**
One Hot model	0.78	0.44
Trainable Embedding Layer	0.75	0.46
Masking Padded Sequences in the Embedding Layer	0.75	0.43
Model with Pretrained GloVe Embeddings	0.75	0.49
Embedding Layer of 100 Training Samples	0.79	0.47
Pretrained Embedding Layer of 100 Training Samples	0.74	0.43
Embedding Layer of 500 Training Samples	0.83	0.39
Pretrained Embedding Layer of 500 Training Samples	0.82	0.38
Embedding Layer of 1000 Training Samples	0.84	0.37
Pretrained Embedding Layer of 1000 Training Samples	0.83	0.36
Embedding Layer of 5000 Training Samples	0.83	0.37
Pretrained Embedding Layer of 5000 Training Samples	0.83	0.37
Embedding Layer of 10000 Training Samples	0.84	0.40
Pretrained Embedding Layer of 10000 Training Samples	0.83	0.37
Embedding Layer of 20000 Training Samples	0.82	0.38
Pretrained Embedding Layer of 20000 Training Samples	0.83	0.35

Conclusion:

The results demonstrate the advantage of using pretrained word embeddings, especially when the available training data is limited. With small training set sizes of 100-500 samples, the models utilizing pretrained GloVe embeddings consistently outperformed those with trainable embedding layers, achieving higher validation accuracy and lower loss. This highlights the benefits of leveraging pretrained embeddings, which provide a good starting point for word representations, when working with scarce training data.

As the training set size increased to 1000 samples and above, the performance gap between pretrained embeddings and trainable embedding layers narrowed significantly. Both approaches achieved similar validation accuracy and loss, suggesting that with sufficient training data, the trainable embedding layers can effectively learn high-quality word representations from scratch, reducing the reliance on pretrained embeddings. Notably, the pretrained embedding layer with 20,000 training samples yielded the best overall performance, with a validation accuracy of 0.83 and a loss of 0.35. However, the differences among the various models with larger training set sizes were relatively small.