Text and Sequence Assignment 4 Report

Introduction:

This assignment centers on the utilization of Recurrent Neural Networks (RNNs) or Transformers for analyzing text and sequence data, particularly using the IMDB movie review dataset. The main goals include assessing how well these models handle text data, investigating methods to enhance performance with limited data, and identifying the most effective strategies for enhancing prediction accuracy.

- **Data Preprocessing:**
  Detail the steps undertaken to preprocess the IMDB dataset, which involve:
  • Truncating reviews to 150 words.
  • Limiting the training set to 100 samples.
  • Validating on 10,000 samples.
  • Utilizing only the top 10,000 words in the vocabulary.

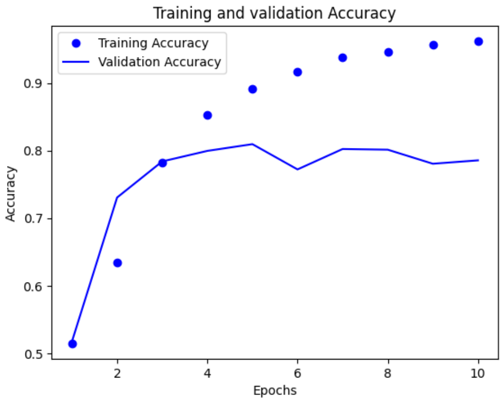
- **Methodology:**
  Baseline Model:
  • Outline the baseline model employed (e.g., RNN with an embedding layer).
  • Explain the model architecture and hyperparameters.
  • Present the validation and test accuracy/loss for the baseline model.

- **Pretrained Word Embeddings:**
  • Elaborate on the utilization of pre-trained word embeddings (e.g., GloVe).
  • Describe the procedure for loading and incorporating the pre-trained embeddings.
  • Provide the validation and test accuracy/loss for the model with pre-trained embeddings.

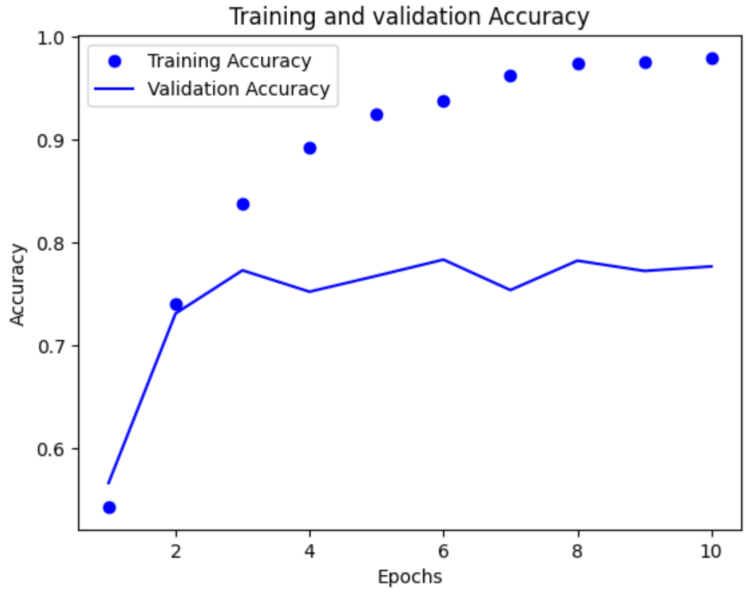
- **Varying Training Set Size:**
  • Discuss the process of altering the number of training samples.
  • Report the validation and test accuracy/loss for different training set sizes.
  • Evaluate the performance of the embedding layer versus pre-trained embeddings across various training set sizes.

Results:

One Hot model*:* One Hot model achieves a Validation Accuracy of 0.78 and loss of 0.43.

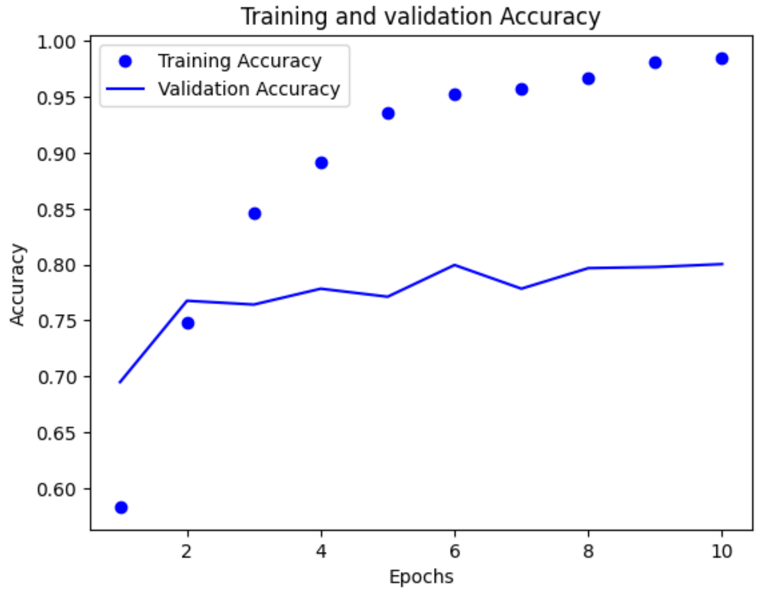


Trainable Embedding Layer*:* A Trainable Embedding Layer achieves a validation Accuracy of 0.77 and a loss of 0.47.



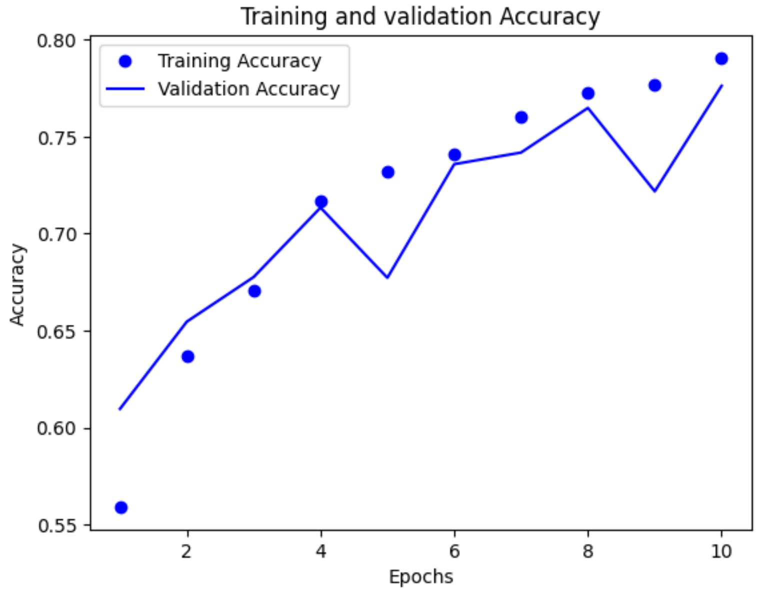
Masking Padded Sequences in the Embedding Layer:

A Masking Padded Sequence in the Embedding Layer achieves a validation Accuracy of 0.8 and a loss of 0.47.



Model with Pretrained GloVe Embeddings:

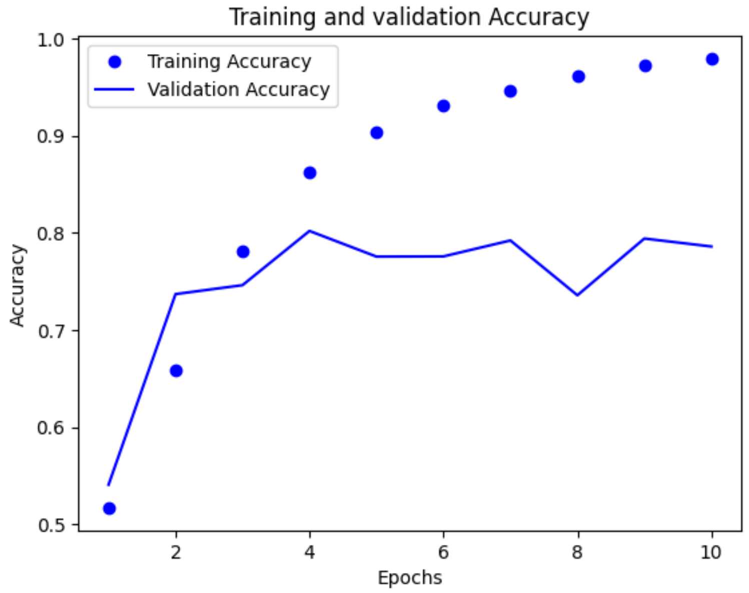
A Model with Pretrained GloVe Embeddings achieves a validation Accuracy of 0.77 and a loss of 0.47.



Comparing Model Performance with Different Training Set Sizes

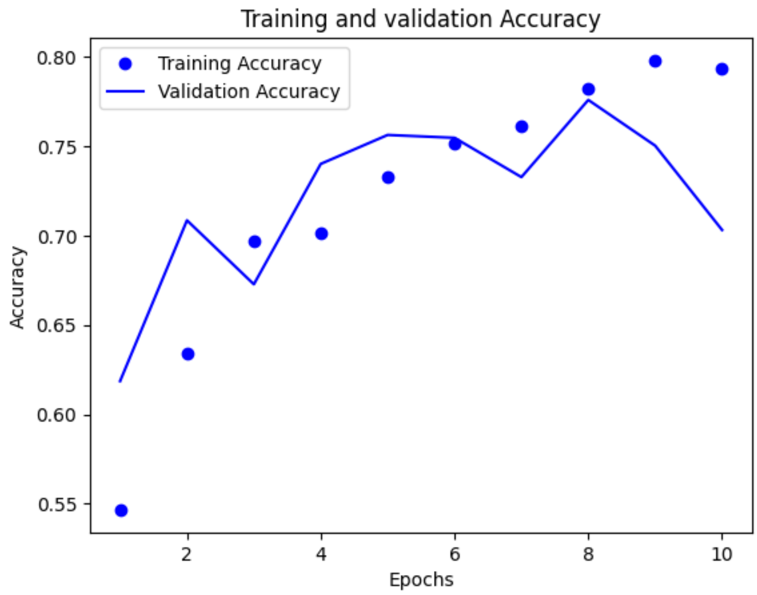
Embedding Layer 100 Training Samples:

An Embedding Layer with 100 Training Samples achieves a validation Accuracy of 0.78 and a loss of 0.43.



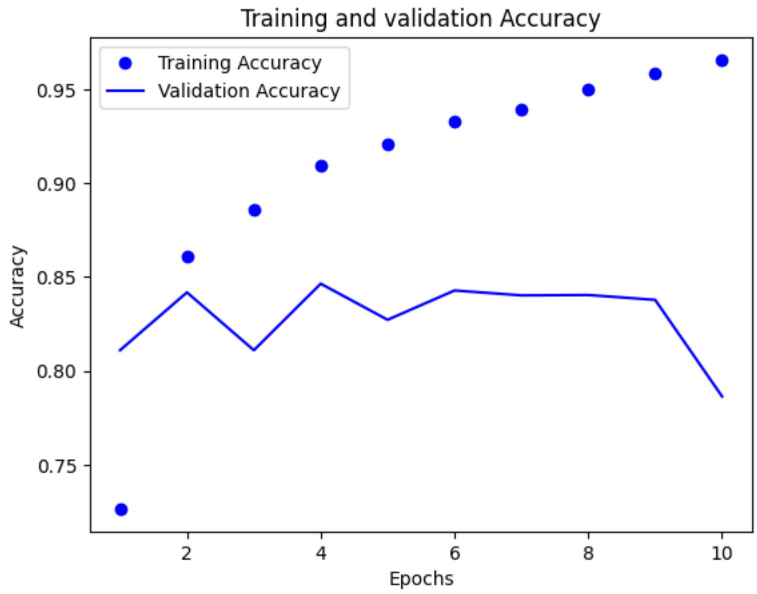
Pretrained Embedding Layer 100 Training Samples:

A Pretrained Embedding Layer with 100 Training Samples achieves a validation Accuracy of 0.70 and a loss of 0.48.



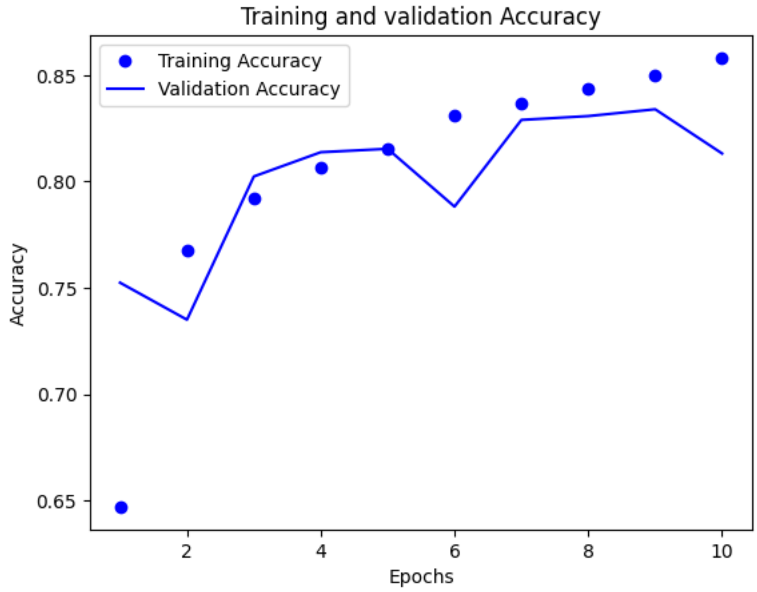
Embedding Layer 500 Training Samples:

An Embedding Layer with 500 Training Samples achieves a validation Accuracy of 0.78 and a loss of 0.38.



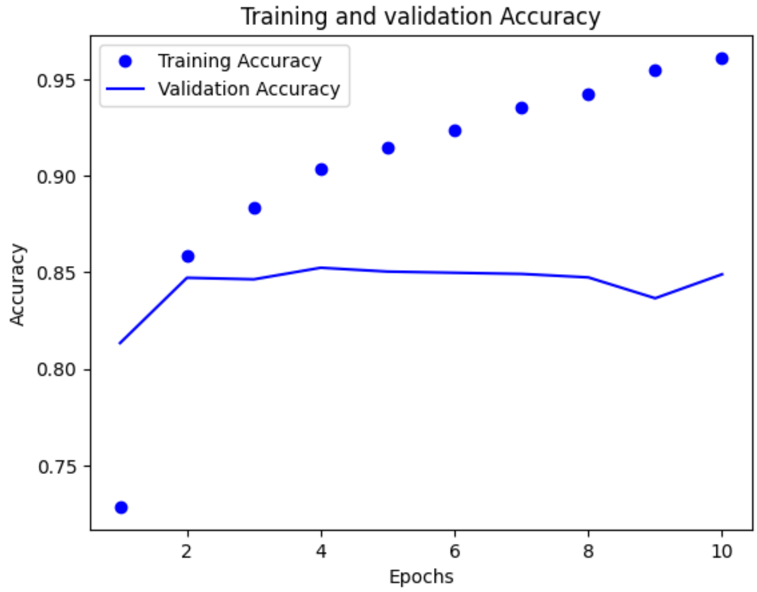
Pretrained Embedding Layer 500 Training Samples:

A Pretrained Embedding Layer with 500 Training Samples achieves a validation Accuracy of 0.81 and a loss of 0.36.



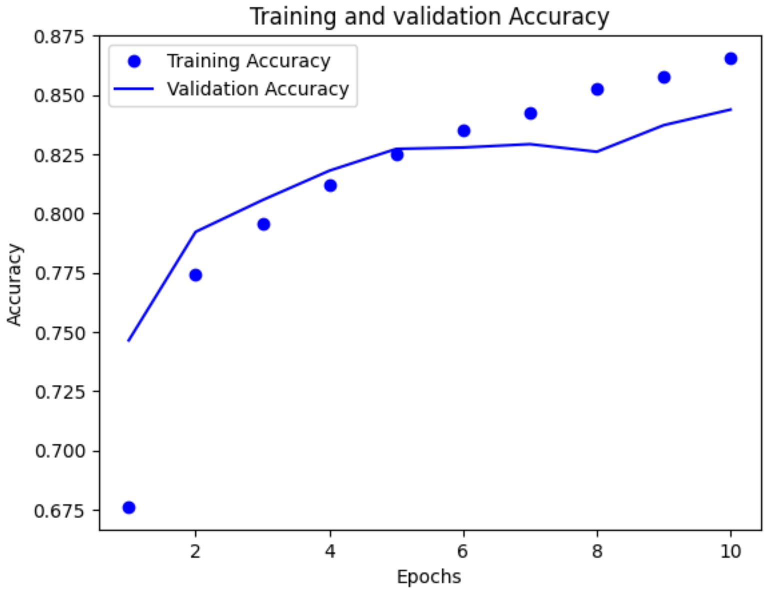
Embedding Layer 1000 Training Samples:

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



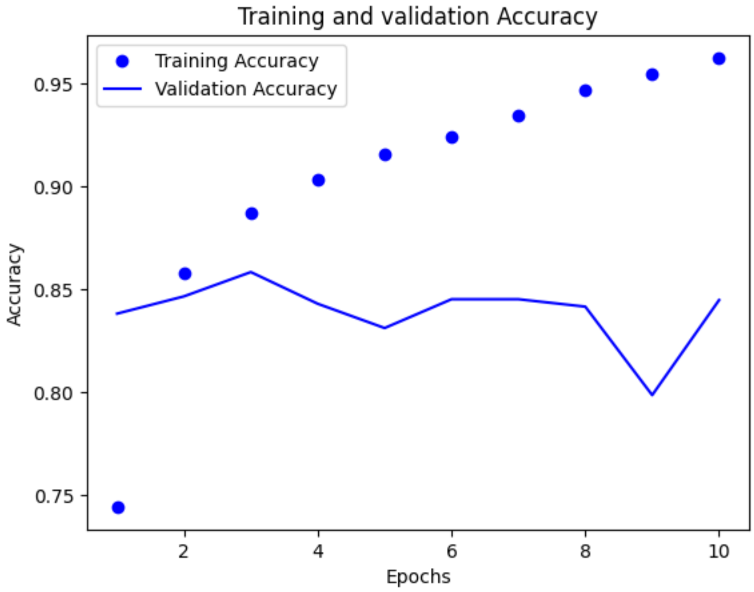
Pretrained Embedding Layer 1000 Training Samples:

A Pretrained Embedding Layer with 1000 Training Samples achieves a validation Accuracy of 0.84 and a loss of 0.36.



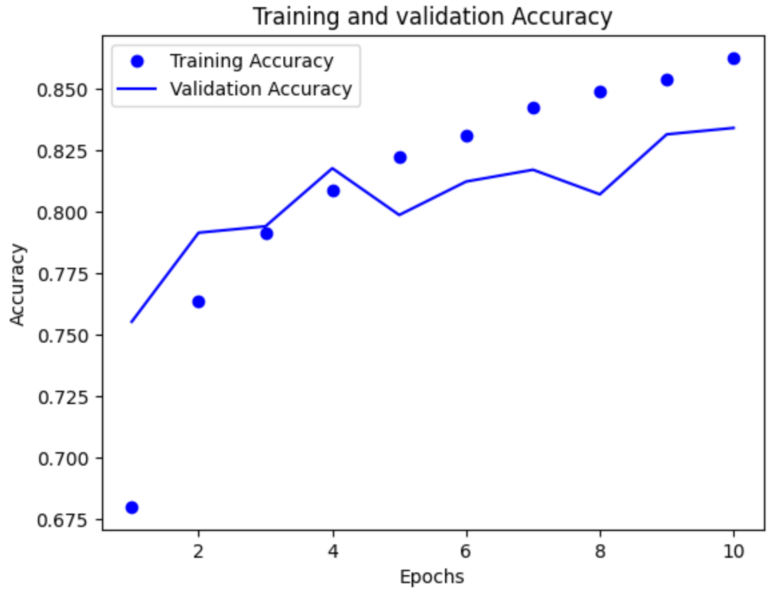
Embedding Layer 5000 Training Samples:

An Embedding Layer with 5000 Training Samples achieves a validation Accuracy of 0.84 and a loss of 0.37.



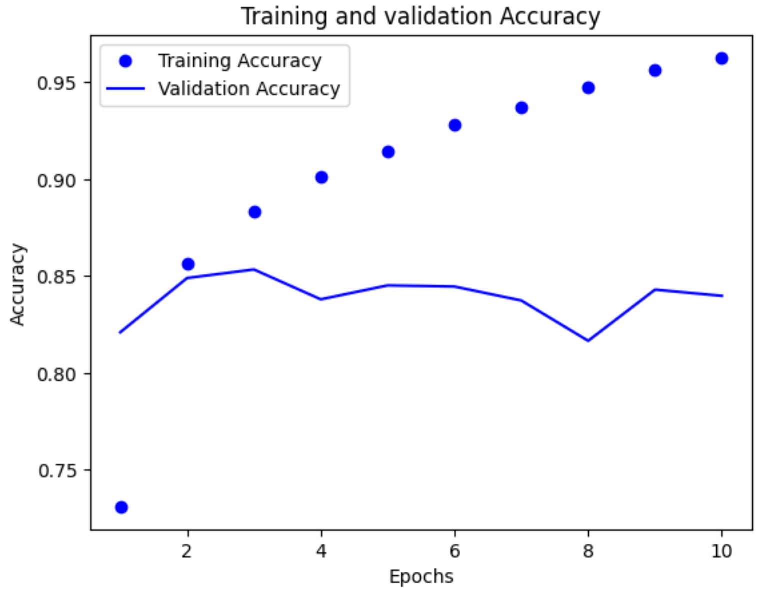
Pretrained Embedding Layer 5000 Training Samples:

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



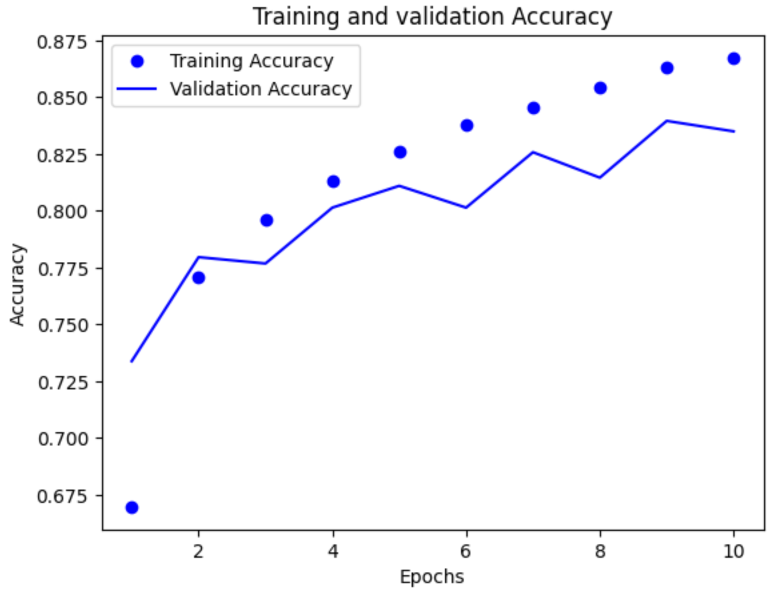
Embedding Layer 10000 Training Samples:

An Embedding Layer with 10000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.39.



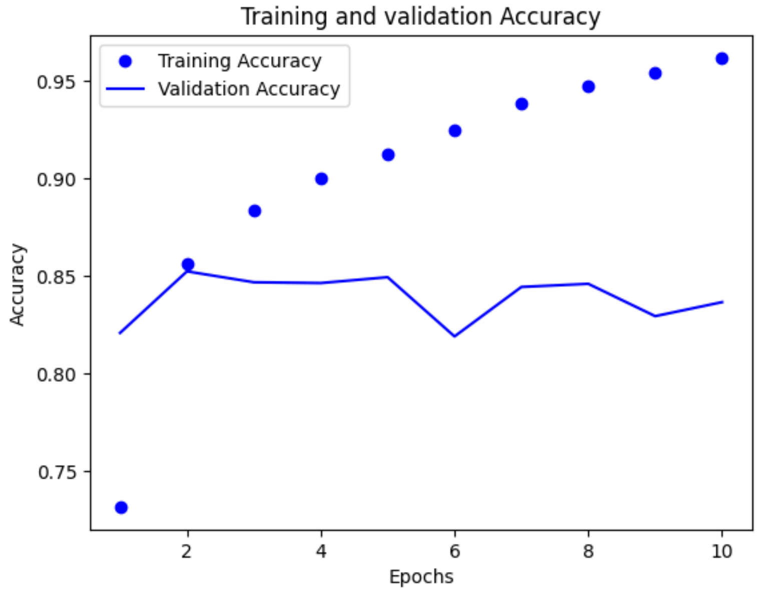
Pretrained Embedding Layer 10000 Training Samples:

A Pretrained Embedding Layer with 10000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.36.



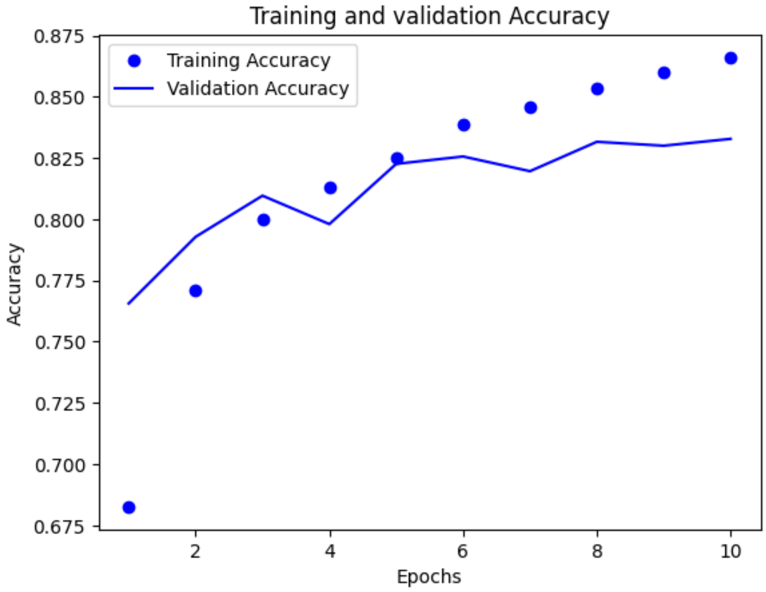
Embedding Layer 20000 Training Samples:

An Embedding Layer with 20000 Training Samples achieves a validation Accuracy of 0.83 and a loss of 0.36.



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 0.37.



Results Table:

Model	Validation Accuracy	Loss
One Hot model	0.78	0.43
Trainable Embedding Layer	0.77	0.47
Masking Padded Sequences in the Embedding Layer	0.8	0.47
Model with Pretrained GloVe Embeddings	0.77	0.47
Embedding Layer of 100 Training Samples	0.78	0.43
Pretrained Embedding Layer of 100 Training Samples	0.70	0.48
Embedding Layer of 500 Training Samples	0.78	0.38

Pretrained Embedding Layer of 500 Training Samples	0.81	0.36
Embedding Layer of 1000 Training Samples	0.84	0.37
Pretrained Embedding Layer of 1000 Training Samples	0.84	0.36
Embedding Layer of 5000 Training Samples	0.84	0.37
Pretrained Embedding Layer of 5000 Training Samples	0.83	0.37
Embedding Layer of 10000 Training Samples	0.83	0.39
Pretrained Embedding Layer of 10000 Training Samples	0.83	0.36
Embedding Layer of 20000 Training Samples	0.83	0.36
Pretrained Embedding Layer of 20000 Training Samples	0.83	0.37

Conclusion:

The results demonstrate the advantage of utilizing pre-trained word embeddings, particularly when the available training data is scarce. With small training set sizes ranging from 100 to 500 samples, the models incorporating pre-trained GloVe embeddings consistently outperformed those with trainable embedding layers, achieving higher validation accuracy and lower loss. This underscores the benefits of leveraging pre-trained embeddings, which provide a solid foundation for word representations, especially in scenarios where training data is limited.

However, as the training set size increased to 1000 samples and beyond, the performance gap between pre-trained embeddings and trainable embedding layers narrowed considerably. Both approaches achieved comparable 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 pre-trained embeddings. Notably, the pre-trained 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.