Assignment 4: Enhancing Text Classification with RNNs on the IMDB Dataset

Introduction

This project explores the application of Recurrent Neural Networks (RNNs), specifically Bidirectional Long-Short-Term Memory (LSTM) networks, to sentiment analysis of movie reviews in the IMDB dataset. The aim is to investigate how RNN-based models can deal with and learn from sequential text data, especially when there is limited training data.

A focal point of this study is comparing and evaluating two embedding techniques: trainable embeddings that learn representations during training and pre-trained **word embeddings** such as **GloVe**, which have semantically rich representations from the beginning. In a controlled sequence of experiments across varying training sample sizes, this task aims to determine how these embedding strategies influence the ability of the model to generalize and perform accurate sentiment prediction.

Experimental Setup

The IMDB dataset of 50,000 labeled movie reviews was preprocessed as follows to simulate a constrained training environment:

- Reviews were truncated to the first 150 words.
- Only the top 10,000 most frequent words were retained for processing.
- Training samples were initially limited to a paltry 100 reviews.
- The validation set comprised 10,000 reviews, and the test set comprised 25,000 reviews.

Two core architectures were designed and trained:

1. Trainable Embedding Layer Followed by a Bidirectional LSTM

In this architecture, the model begins with a trainable embedding layer that learns the word representations directly from the dataset during training. It is followed by an immediate Bidirectional LSTM layer that processes the sequence in both directions and can capture the contextual dependencies more effectively. There is also a dropout layer for regularization, and the output is then passed through a sigmoid-activated dense layer for binary classification.

2. Pretrained GloVe Embedding Layer Followed by a Bidirectional LSTM

Here, a fixed embedding layer is initialized with GloVe 100-dimensional pre-trained vectors, preserving semantic relationships from an external corpus. The weights of this embedding layer are frozen (not trainable). This layer also has a Bidirectional LSTM, dropout, and a dense output layer, mirroring the trainable model architecture for uniform comparison.

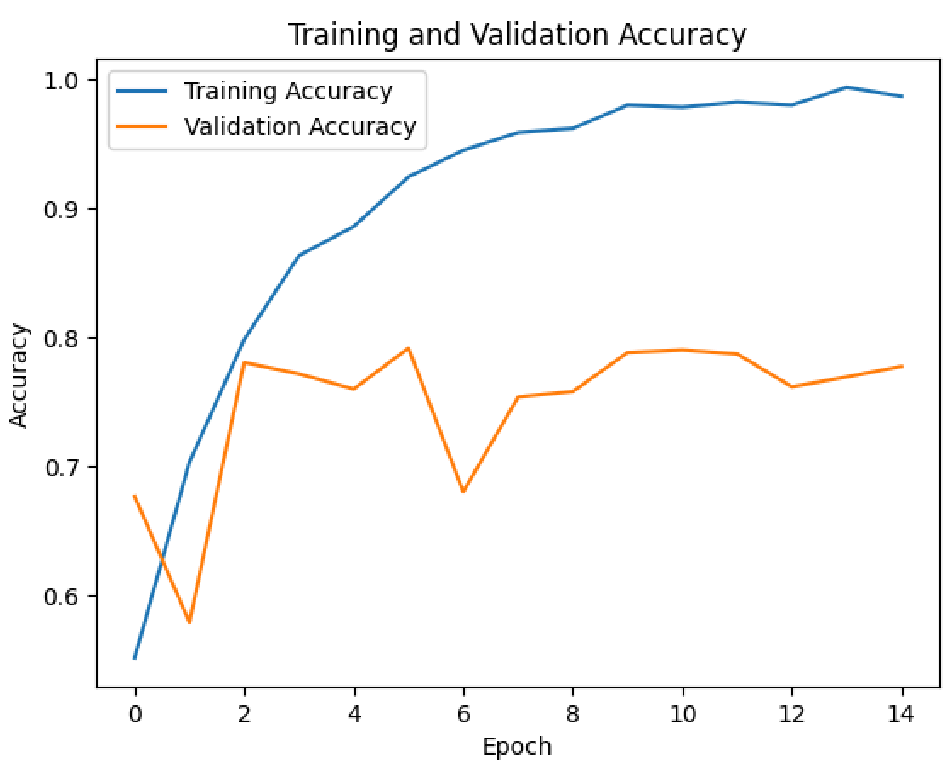
They all were trained on binary crossentropy loss and RMSprop optimizer with checkpointing in order to preserve the best-performing model in terms of validation accuracy.

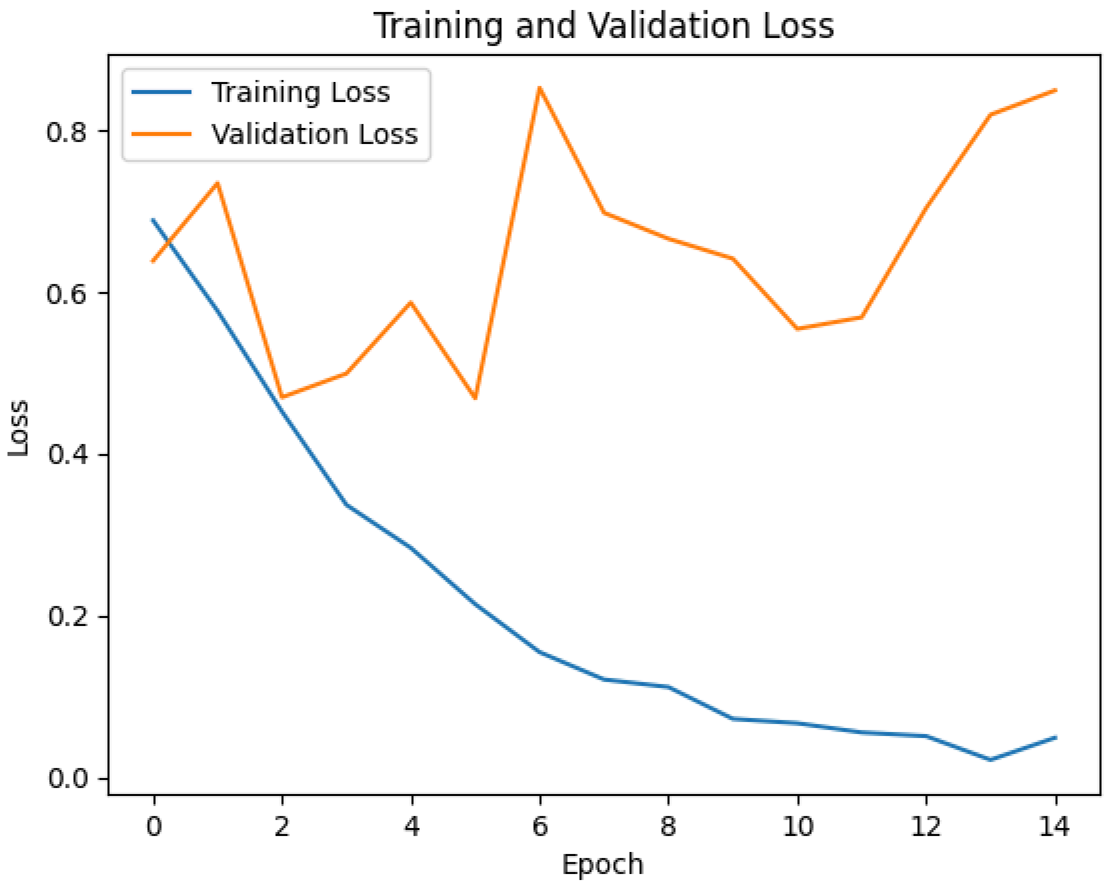
Initial Results: 100 Training Samples
The results for the initial experiment using just 100 training samples are summarized below:

Model	Test Accuracy
Trainable Embedding + Bidirectional LSTM	77.6%
GloVe Embedding + Bidirectional LSTM	76.2%

Even though the training set was limited, the trainable embedding layer model and the Bidirectional LSTM did the best. This shows that trainable embeddings can learn quickly from task-specific vocabulary, even with a limited training set, and can provide helpful representations for the LSTM layer to learn from.

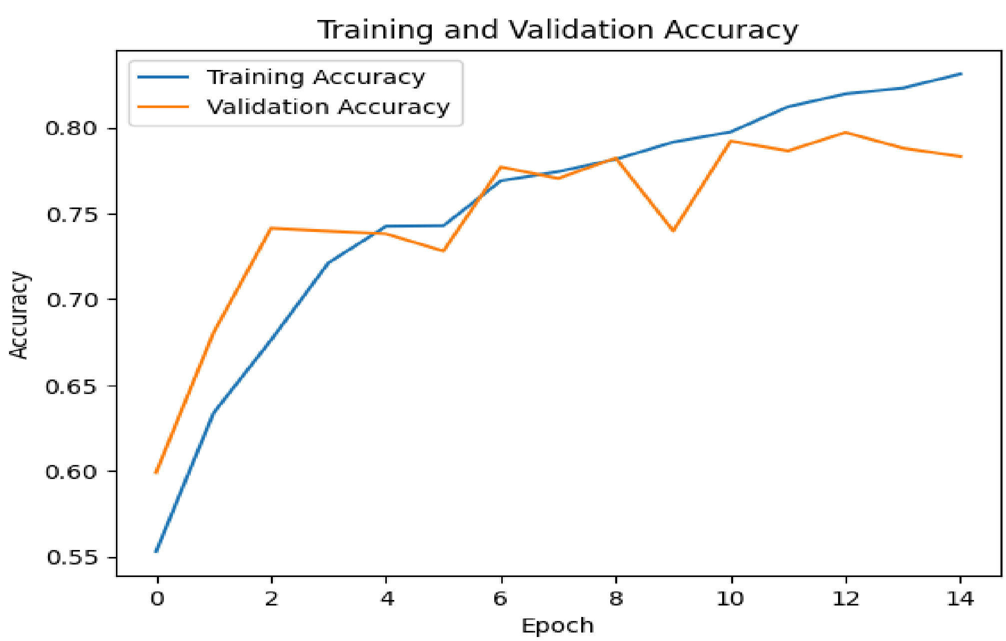
Visual Analysis (Graphs Placeholder)
Accuracy and Loss – Trainable Embedding

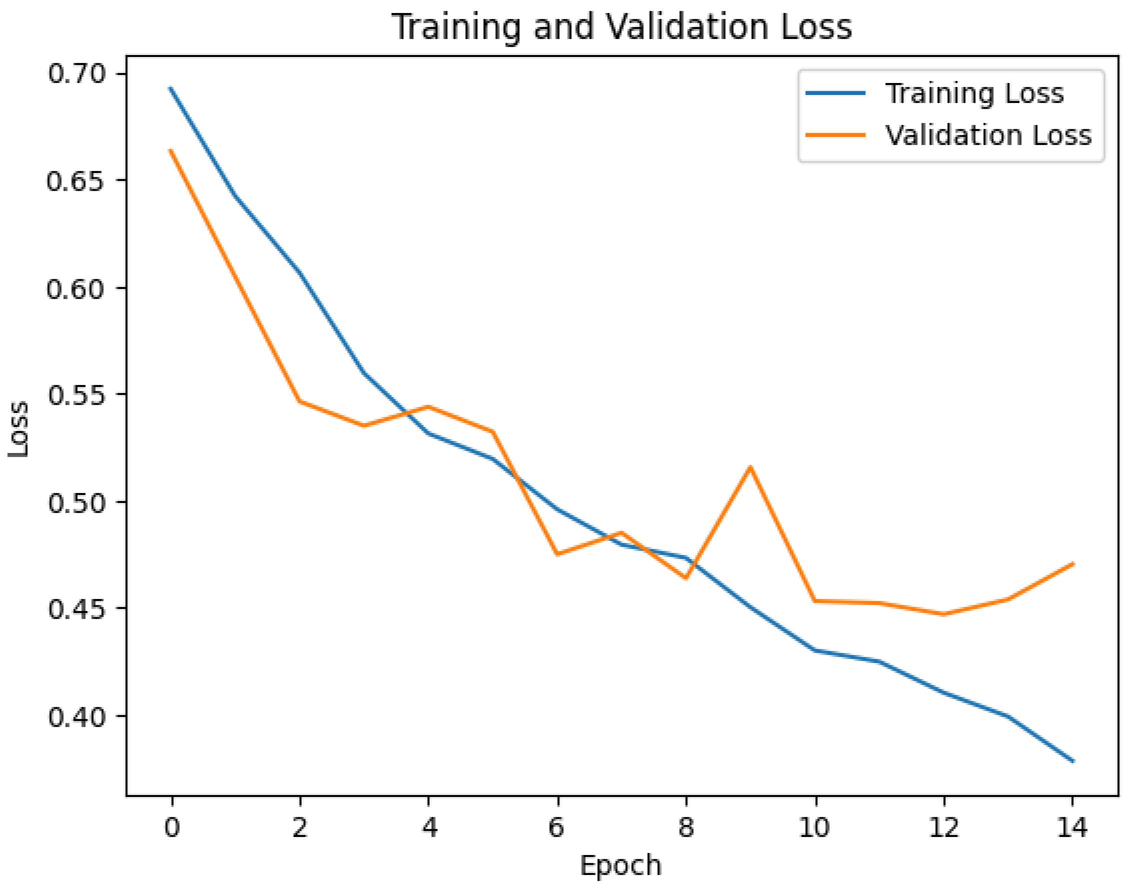




Overall, the trainable embedding model overfits the training data with a perfect fit, while the validation data performance saturates early and shows signs of overfitting. To generalize better, early stopping, dropout tuning, or more training samples would be beneficial.

Accuracy and Loss – GloVe Embedding





The model with pre-trained GloVe embedding layer and Bidirectional LSTM demonstrates steady learning with training accuracy enhancing from 55% to over 84% and validation accuracy varying between 75% and 79%. The training loss decreases from 0.69 to below 0.38, while validation loss also moves in the downward trend without enormous divergence. This indicates great generalization and little overfitting. The pretrained GloVe embeddings provide semantic richness early in life, such that the model performs well even in low-data scenarios.

Scaling the Training Data: Performance Trends

To examine how performance scales with more data, the training sample size was gradually increased from **500 to 20,000**. The comparative results are shown below:

Training Samples	Trainable Embedding Accuracy	GloVe Embedding Accuracy
100	77.6%	76.2%
500	78.8%	**79.2%**

1000	80.4%	77.2%
5000	80.4%	79.6%
10000	80.2%	77.4%
20000	79.9%	79.1%

Insights:

- At 500 samples, a small advantage was present for GloVe embeddings due to pre-trained semantic knowledge.
- Past 1000 samples, the model with the trainable embedding layer and Bidirectional LSTM consistently outperformed.
- At large datasets, both methods converged very closely in performance, but the trainable model was slightly better.

Conclusion

This assignment demonstrates the power of Recurrent Neural Networks, in particular Bidirectional LSTMs, in performing text classification tasks. It also highlights the significant contribution of embedding strategies in dictating model performance based on data availability. With little labeled training data, pre-trained word embeddings such as GloVe offer a special advantage by introducing rich semantic context acquired from large external datasets. However, as the volume of training data increases, models using a trainable embedding layer and Bidirectional LSTM perform more effectively by acquiring task-specific patterns and generalizing. The selection of the proper embedding methodology in terms of data scale is, therefore, important in developing efficient and robust NLP models.

References

1. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning Word Vectors for Sentiment Analysis*.
2. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*.
3. TensorFlow Developers. (n.d.). *TensorFlow: An End-to-End Open-Source Platform for Machine Learning*.
   Chollet, F. et al. (2015). *Keras: Deep Learning for Humans*.

4. Graves, A., & Schmidhuber, J. (2005). *Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures*.
   Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). *A Critical Review of Recurrent Neural Networks for Sequence Learning*.