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

Introduction

This assignment explores the application of Recurrent Neural Networks (RNNs) and bidirectional Long-Short-Term Memory (BiLSTM) models to sentiment classification using the IMDB movie review dataset. It aims to identify the performance of such sequence models when applied in low-data situations and how different embedding methodologies impact model accuracy and generalization.

We investigate the comparative performance of using two types of word embeddings: task-specific training embeddings learned from the training data, and pre-trained embeddings such as GloVe that provide fine-grained semantic content extracted from large external corpora. The research investigates the extent to which the approaches hold up when training data are in short supply and more data become available.

Data Preparation and Design Constraints

The IMDB 50,000 labeled movie reviews dataset was preprocessed to simulate real-world constraints:

- Reviews were truncated to the first 150 words.
- Only the 10,000 most frequent words were retained.
- The number of training instances was varied from 100 to 20,000.
- A fixed validation set of 10,000 samples and a test set of 25,000 reviews were used.

The text was vectorized using TensorFlow's TextVectorization layer. All models were trained using binary cross-entropy loss and the RMSprop optimizer, with dropout regularization and checkpointing on validation accuracy.

Model Architectures

1. Trainable Embedding Layer + Bidirectional LSTM

It uses a trainable embedding layer that learns task-specific word representations during training. This is followed by a Bidirectional LSTM layer to learn forward and backward contextual dependencies. There is a dropout layer to avoid overfitting and a final dense layer with sigmoid activation to produce a binary output.

2. pre-trained GloVe (Word Embedding)Layer + Bidirectional LSTM
This model utilizes a frozen embedding layer that is initialized using 100-dimensional pre-trained GloVe vectors. This embedding layer does not change during training. The rest of the architecture copies the original model to ensure a fair comparison.

Initial Evaluation (100 Training Samples)

Model	**Test Accuracy**
Trainable Embedding + Bidirectional LSTM	78.2%
GloVe Embedding +Bidirectional LSTM	76.7%

Even with just 100 training samples, the model with a trainable embedding layer and a Bidirectional LSTM performed slightly higher at test accuracy. This suggests that task-specific embeddings can work well even in low-data settings.

Scaling Model Performance with Training Data

Performance was evaluated across increasing training sizes from 100 to 20,000 samples. The results are summarized below:

Training Samples	Trainable Embedding Accuracy	GloVe Embedding Accuracy
100	78.2%	76.7%
500	79.0%	79.6%
1,000	79.0%	77.9%
5,000	79.2%	79.2%
10,000	79.3%	78.1%
20,000	79.4%	78.7%

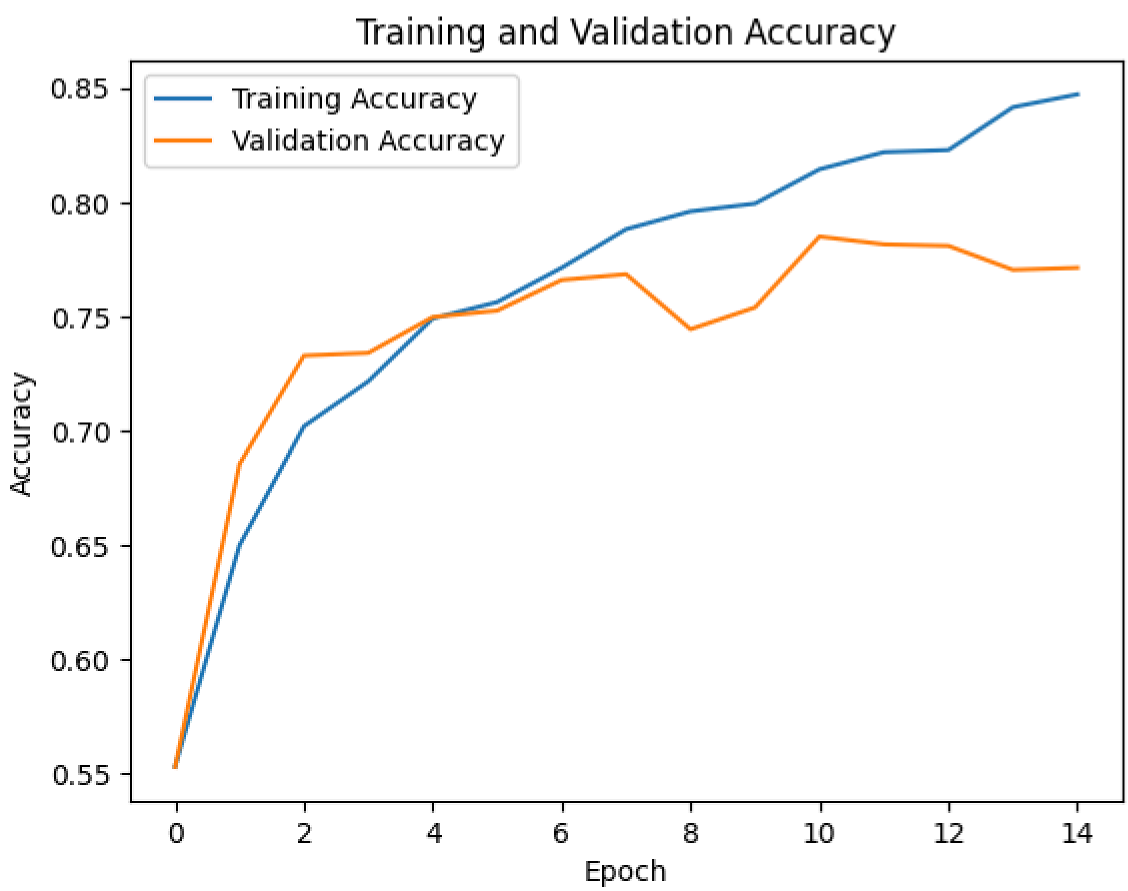
Key Observations

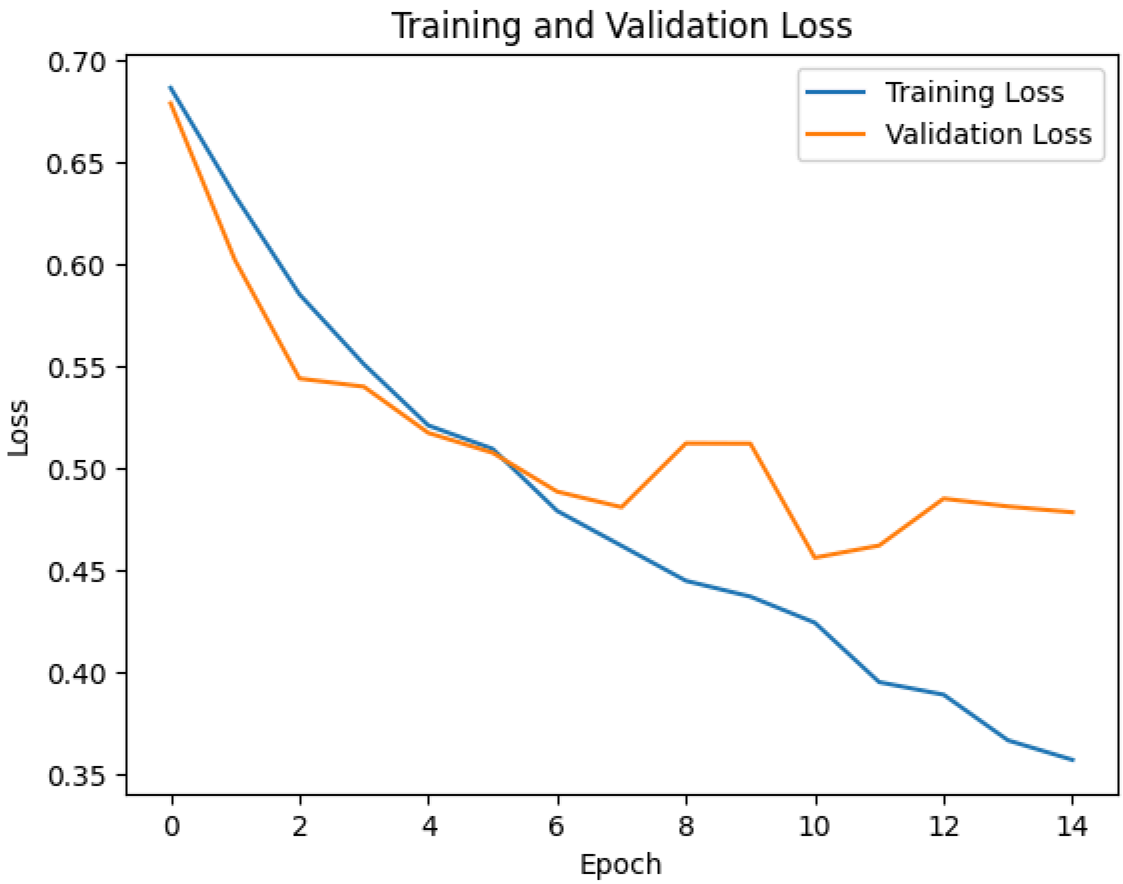
- At 500 samples, the GloVe pre-trained model performs better due to its semantic richness.

- Above 1,000 samples, the trainable embedding model consistently gives more accuracy.
- At larger training sizes, both models converge but with a slight edge for the trainable model.

Model Learning Behavior

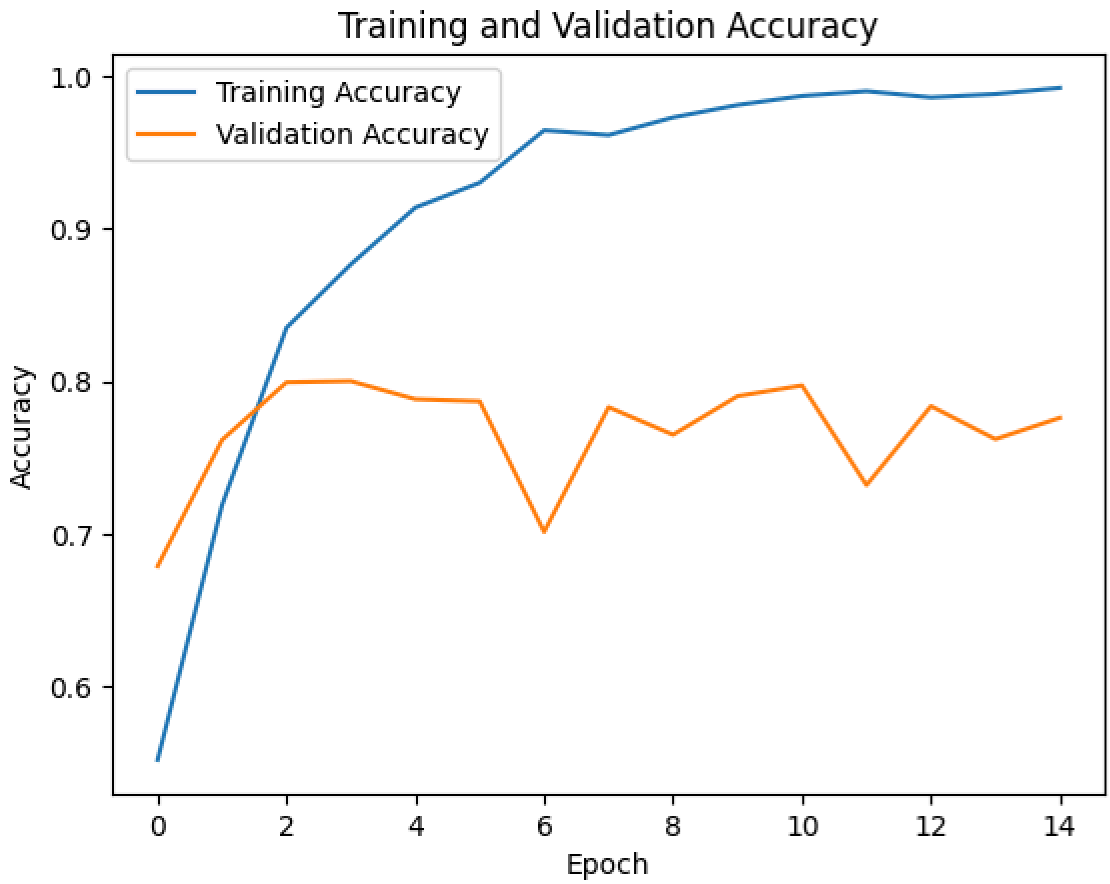
Trainable Embedding Observations

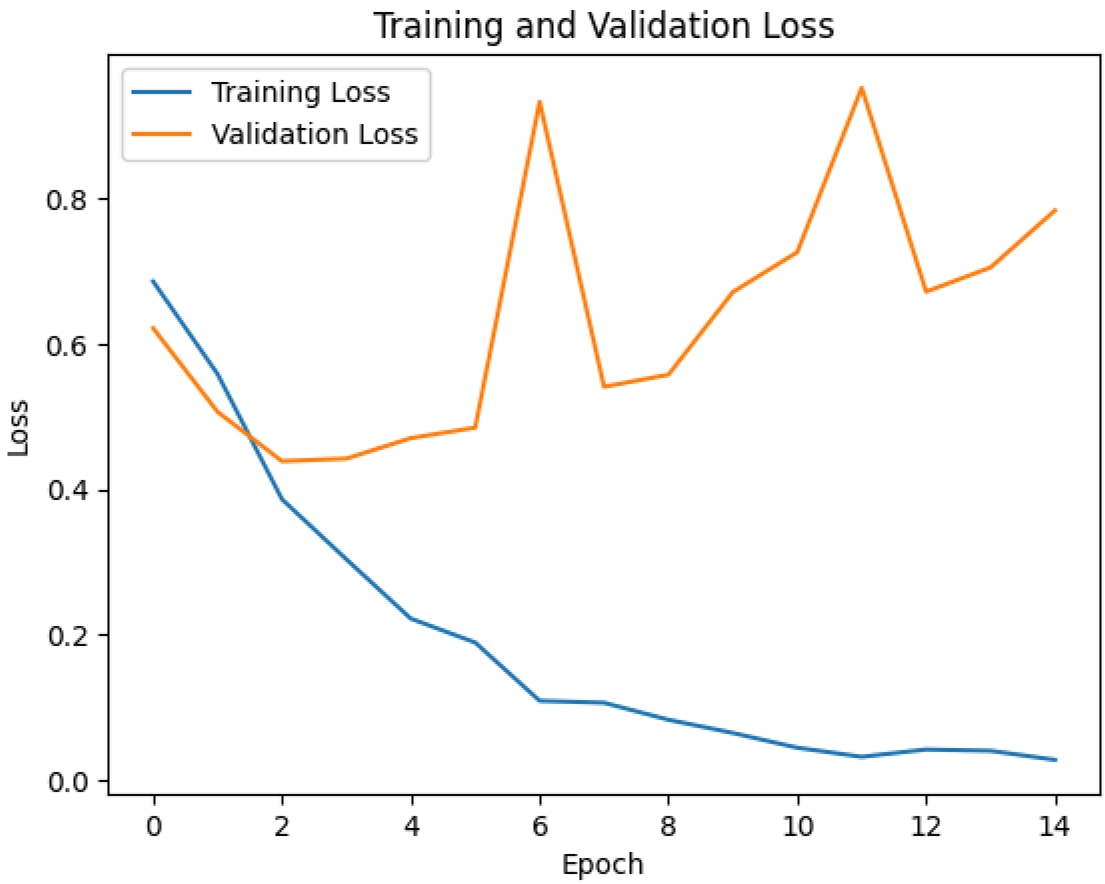




The graphs show a sharp increase in training accuracy to 85.3% while validation accuracy flattens out at 77.1%. Training loss decreases consistently to 0.35, while validation loss oscillates after a few epochs with a peak of around 0.48. This is an indication that the model is overfitting, learning well on the training data but not generalizing as well to validation data. Early stopping or regularization techniques can be employed to improve performance in such cases.

GloVe Embedding Observations


Training and Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy
Epoch
1.0
0.9
0.8
0.7
0.6
0
2
4
6
8
10
12
14



The GloVe model demonstrates steady improvement in training accuracy, from 76.9% to 85.3%. Validation accuracy plateaus between 75% and 78%, with consistent generalization. Training loss declines from 0.50 to 0.35, with validation loss running smoothly between 0.46 and 0.51 without any significant overfitting. These patterns solidify the claim that pre-trained embeddings offer stable, interpretable learning even in sparse-data settings.

Conclusion

This assignment neatly demonstrates the strength of Bidirectional LSTM architectures for sentiment classification and the critical role of embedding strategy in model performance determination. The key findings are as follows:

- Pretrained GloVe embeddings perform strongly in low-resource training data conditions by exploiting rich semantic knowledge gained from large external corpora.
- With additional labeled data, trainable embeddings become more and more superior to pretrained embeddings by discovering task-specific linguistic patterns and customizing representations to the dataset at hand.

- The trainable embedding combined with a BiLSTM structure is more scalable and resilient to large-scale training, with enhanced accuracy and flexibility.
- Lastly, the use of pre-trained versus trainable embeddings needs to be guided by the availability and quality of data—with pre-trained embeddings offering the most viable alternative under low-resource conditions, while trainable embeddings would be more helpful where large volumes of data are present.

References

1. Maas, A. L., et al. (2011). *Learning Word Vectors for Sentiment Analysis*
2. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*
3. Graves, A., & Schmidhuber, J. (2005). *Bidirectional LSTM Architectures*
4. Lipton, Z. C., et al. (2015). *A Critical Review of RNNs for Sequence Learning*
5. TensorFlow Developers. *TensorFlow Platform*
6. Chollet, F., et al. *Keras API Documentation*