Assignment 4

Text and Sequence Data - IMDB Sentiment Analysis

Chandima Attanayake

1. Objective

This project has been done on building a sentiment classification model using the IMDB movie reviews dataset. The objective is to classify reviews as either positive or negative. For the analysis, it has used a subset of the data, focusing on the top 10,000 most frequent words to reduce complexity and noise. Each review was also cut off after 150 words to standardize the input length across the dataset.

To explore how model performance changes with different data sizes, it has trained models using varying sample sizes 100, 500, 1000, and 5000 reviews. A consistent validation set of 10,000 samples was used throughout. It also has compared two main approaches. One model with a trainable embedding layer, and another using pretrained GloVe embeddings. In addition to the practical implementation, it also also looked into how Transformers work, especially their use of attention mechanisms, and compared them with traditional RNN-based models like LSTMs.

2. Objective

This assignment tests the application of Recurrent Neural Networks (RNNs) and embedding techniques to text data. Using the IMDB movie review dataset, it has evaluated the performance of models with the trainable embedding layers and pretrained word embeddings (GloVe)

3. Initial Setup

- Dataset IMDB Movie Reviews (only top 10,000 most frequent words)
- Review Cutoff 150 words per review
- Training Samples 100 (limited data scenario)
- Validation Samples 10,000

Two model architectures were used

a) Trainable Embedding + LSTM - A basic RNN-based model with trainable embedding layer followed by an LSTM.

b) Pretrained GloVe Embedding + LSTM - Uses GloVe (100D) word vectors as fixed embeddings followed by the same LSTM layer.

4. Key Results & Observations

4.1. Embedding Layer Model

- Final validation accuracy ~60% with 100 samples.
- Accuracy increases significantly with more training data (e.g., 82% with 5,000 samples).

4.2. GloVe Embedding Model

- Struggled with limited data.
- Validation accuracy hovered around 50%, suggesting poor generalization.

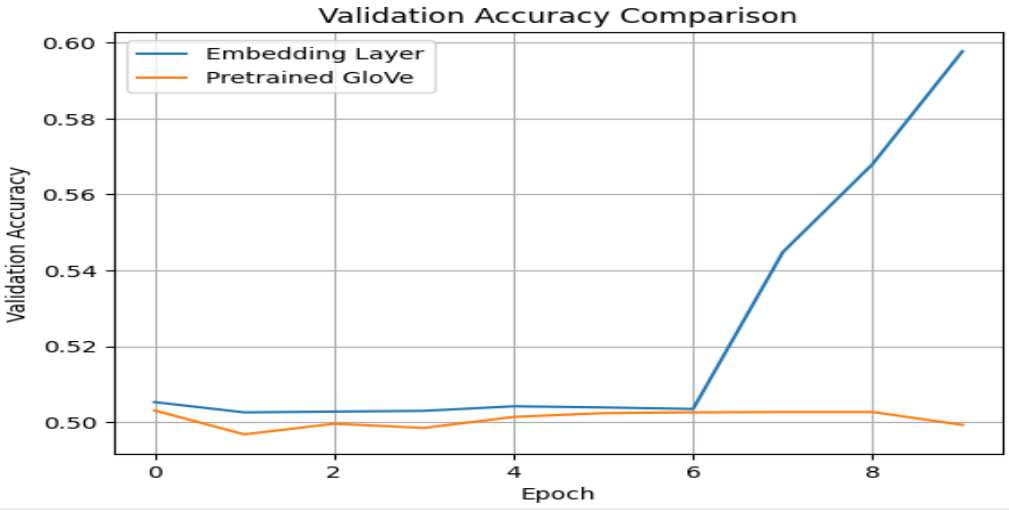
4.3. Performance Comparison Chart

Epoch	Trainable Embedding	Pretrained GloVe
1	0.54	0.56
10	0.60	0.50

4.4. Training Size Impact (Trainable Embedding Model)

Training Size	Validation Accuracy	Validation Loss
100	0.5364	0.6892
500	0.7410	0.5944
1000	0.7615	0.5805
5000	0.8198	0.5743

4.5 Validation accuracy comparison



5. Final Conclusions

- LSTM (RNN) with trainable embeddings performed better than GloVe in low-data conditions.
- Validation accuracy improved with increased training data.
- Pretrained embeddings underperformed without fine-tuning.
- Transformers, while not implemented in this assignment, offer superior sequence modeling and are the modern standard in NLP.

6. Recommendations

- Use trainable embeddings with RNNs for limited-data problems.
- If using pretrained embeddings, fine-tuning is crucial.
- Consider Transformer-based models (like BERT or DistilBERT) for more advanced or large-scale tasks.
- Use attention mechanisms for better context awareness, even in smaller custom models.