Comparison of Custom and GloVe Embedding Models on IMDB Dataset

1. Introduction

Purpose: To compare the performance of a custom embedding layer with a pretrained GloVe embedding on a limited subset of the IMDB dataset.
Methodology: Two models were trained on 100 samples of the IMDB dataset, one using a custom embedding layer and the other using GloVe embeddings. Both models were evaluated on a validation set of 10,000 samples.

2. Model Specifications

Both models used a similar architecture: an embedding layer followed by an LSTM layer and a dense output layer. The custom model used an embedding layer trained from scratch, while the GloVe model utilized pretrained embeddings.

3. Training Results

Over 10 epochs, both models showed an increase in accuracy on the training set. The custom embedding model exhibited a larger increase in accuracy, from 56% to 82%. The GloVe model showed a more moderate increase, from 52% to 62%.

4. Validation Results

On the validation set, the custom embedding model achieved a higher accuracy (54.52%) compared to the GloVe model (51.48%). However, both models showed only around 50% accuracy for most of the training epochs, indicating challenges in generalization.

5. Discussion and Conclusions

The custom embedding model slightly outperformed the GloVe model in this context. The limited number of training samples (100) likely affected the performance of both models. Both models showed signs of overfitting, as indicated by the higher training accuracy compared to the validation accuracy.

6. Recommendations

Increasing the number of training samples could improve the model's ability to generalize. Further tuning of the models, such as adjusting the LSTM units or the learning rate, may yield better results. Experimentation with different architectures or additional layers might also be beneficial.

7. Summary Table

Metric	Custom Embedding Model	GloVe Embedding Model
Final Training Accuracy	82.00%	62.00%
Final Validation Accuracy	54.52%	51.48%
Loss on Validation Set	0.7051	0.6972