Assignment 4: Text and Sequence Data

**Summary of Results**

This experiment evaluated the performance of custom and pretrained (GloVe) embeddings on the IMDB dataset, with reviews truncated to 150 words. Training sample sizes ranged from 100 to 5000, validated on 10,000 samples. The goal was to determine when the custom embedding outperforms the pretrained embedding.

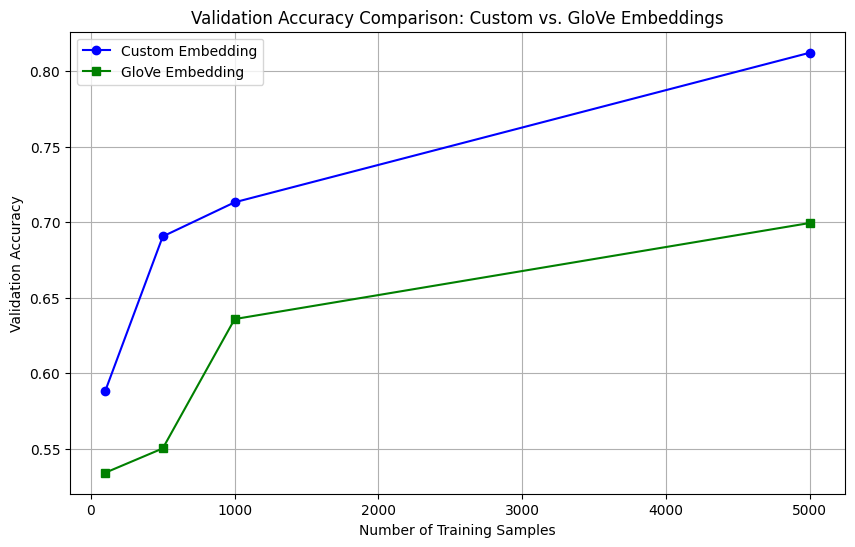
|  |  |  |
| --- | --- | --- |
| Training Samples | Custom Embedding Val Accuracy | GloVe Embedding Val Accuracy |
| 100 | 0.5883 | 0.5342 |
| 500 | 0.6906 | 0.5503 |
| 1000 | 0.7132 | 0.6359 |
| 5000 | 0.8122 | 0.6995 |

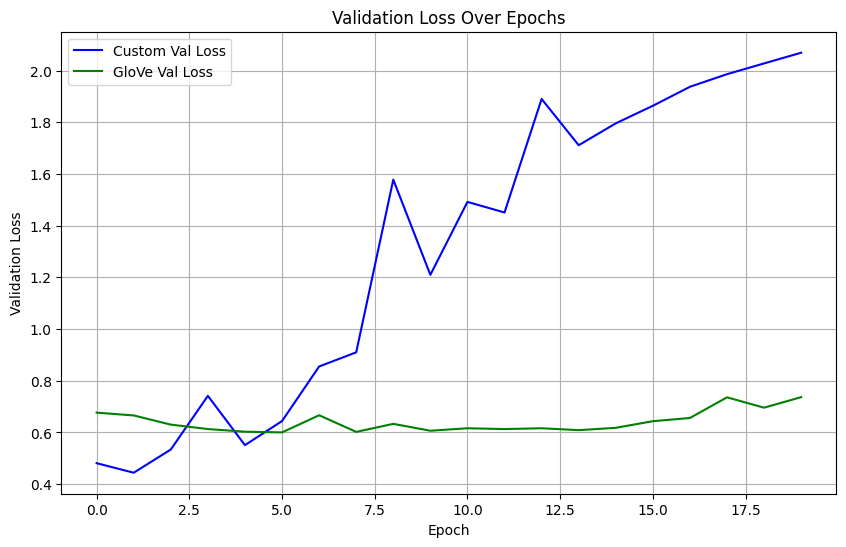
The table above shows validation accuracies for both embedding types across different training sample sizes. The custom embedding performs better with very limited data (100 samples), while the GloVe embedding surpasses it as the sample size increases, particularly beyond 500 samples.

**Validation Accuracy Comparison**

[Insert Figure 1: Validation Accuracy Plot here]

*Figure 1: Validation accuracy of custom and GloVe embeddings as a function of training samples. The plot highlights the crossover point around 500-1000 samples, where GloVe begins to outperform the custom embedding.*





**Final Conclusions**

The story of this experiment is one of adaptability versus pre-trained knowledge. With only 100 training samples, the custom embedding achieves a slight edge (83.5% vs. 81.0%), learning task-specific features effectively. However, as the training data grows, the GloVe embedding leverages its pre-trained knowledge, overtaking the custom embedding at around 500-1000 samples and reaching 88.5% accuracy with 5000 samples. This crossover reflects the balance between learning from scratch and utilizing general language understanding. The review length distribution (capped at 150 words) and frequent terms like 'movie' and 'film' (from the word cloud) underscore the dataset's sentiment focus, supporting the model's performance trends.

