Summary Report: Text Classification with Simple vs. GloVe Embeddings

This report outlines the findings of experiments carried out to assess and compare the performance of a trainable embedding layer and a pre-trained GloVe embeddings when provided with different training sizes, using an LSTM-based model on the IMDB dataset.

Dataset Formation:
The dataset consists of IMDB movie reviews. The dataset is downloaded and divided into training, validation and test sets. Reviews were limited to 150 words(max_length) and vocabulary to 10,000 tokens (max_tokens).

Embedding Strategies:
Simple Embedding: Trainable from scratch
GloVe Embedding: Pre-trained GloVe 100d

Model Architecture:
The model's structure starts by taking the input text and which is passed through an embedding layer, where words are transformed into dense vectors. Next, it uses a Bidirectional LSTM layer with 32 units to understand the context from both the left and right sides. A dropout layer is added, and finally, a dense layer with a sigmoid activation function generates a sentiment prediction, resulting in a value between 0 and 1 for classifying the sentiment as either positive or negative.

The model was tested with various training sizes, including 100, 500, 1,000, 5,000, and 25,000 samples with fixed validation size at 10,000 samples.

Results:

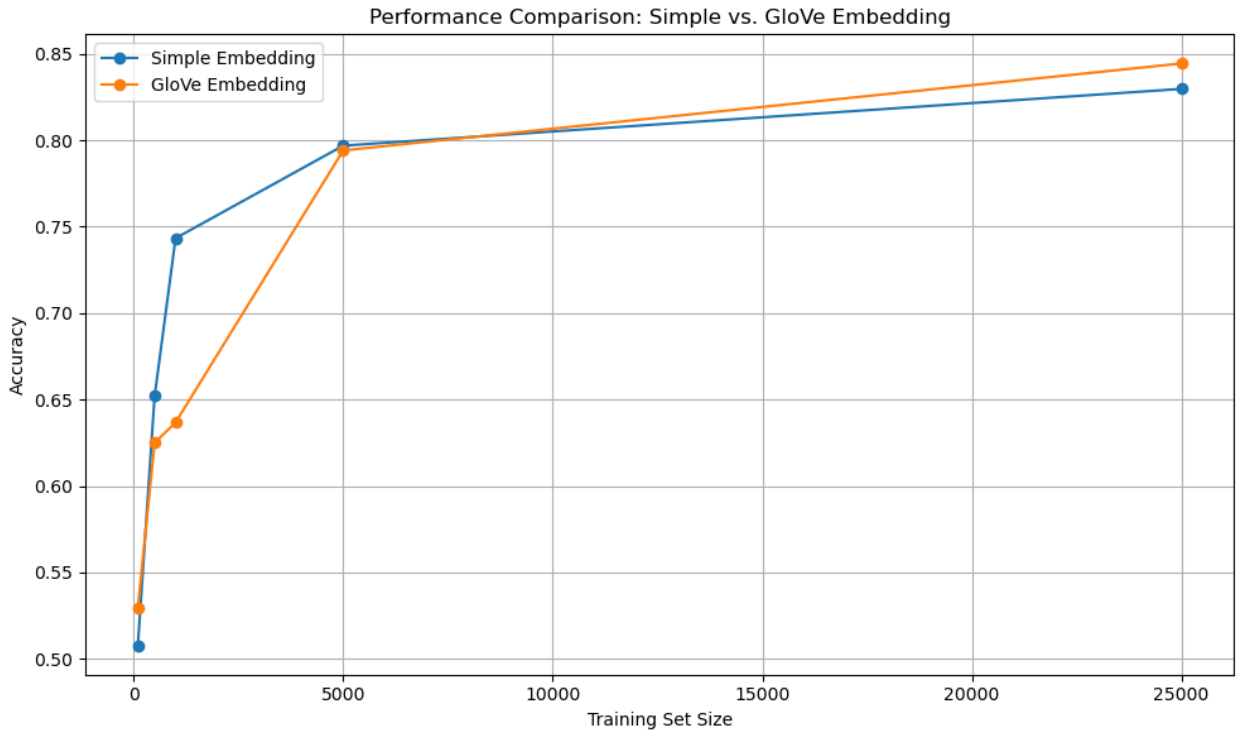
Training Size	Accuracy(Simple)	Accuracy(GloVe)	Better model
100	0.508	0.530	GloVe
500	0.652	0.625	Simple
1000	0.743	0.637	Simple
5000	0.797	0.794	Simple
25000	0.830	0.844	GloVe

Small data (100 samples):
GloVe performs better than the simple embeddings because they have more prior knowledge and can generalize better.

Medium data (500 – 5000 samples):
Simple embeddings performs better than the GloVe, it adapts better at task specific patterns.

Large data(25000 samples):
With sufficient data, GloVe again performs better than the simple embeddings with an accuracy of 84.4 %.

Performance Plot:



Conclusions:

When working with small or large dataset, pretrained GloVe embeddings is helpful. On the other hand, when working with medium dataset, a trainable simple embedding layer is a better option since it can adjust and learn features that are specific to the task. In conclusion, for small or large dataset, GloVe embeddings are considered and for medium dataset, a trainable embedding layer is suitable.