SENTIMENT ANALYSIS ON IMDB REVIEWS

Assignment 4

Priyanka Jonnala

Harini Padmaja Solleti

Objective:

The primary objective of the IMDB dataset is binary classification, specifically categorizing movie reviews as positive or negative. With a total of 50,000 reviews, the analysis focuses on evaluating the top 10,000 words. The dataset's training samples are restricted to varying sizes: 100, 5000, 1000, and 100,000, while validation is conducted on a consistent 10,000 samples. After data preparation, the reviews are processed through both an embedding layer and a pre-trained embedding model. Various strategies are then employed to gauge the performance of these models across different training sample sizes.

Data Preprocessing:

- Each review undergoes preprocessing to convert it into a sequence of word embeddings, where each word is represented by a fixed-size vector. This process is applied to a maximum of 10,000 samples. Additionally, instead of using strings of words directly, a series of numerical representations, each corresponding to a unique word, is generated from the reviews. However, the neural network's input cannot directly accept this list of numbers without further processing.
- To generate tensors from the integer lists, we need to ensure that each sample has the same length. This means we must pad or truncate each review to a fixed length using dummy words or numbers. This ensures consistency in the length of each sample, allowing us to create tensors with dimensions (samples, word indices) using integer data types.

Procedure:

In this work, I investigated two different approaches to word embedding generation for this IMDB dataset:

1. Custom-trained embedding layer.

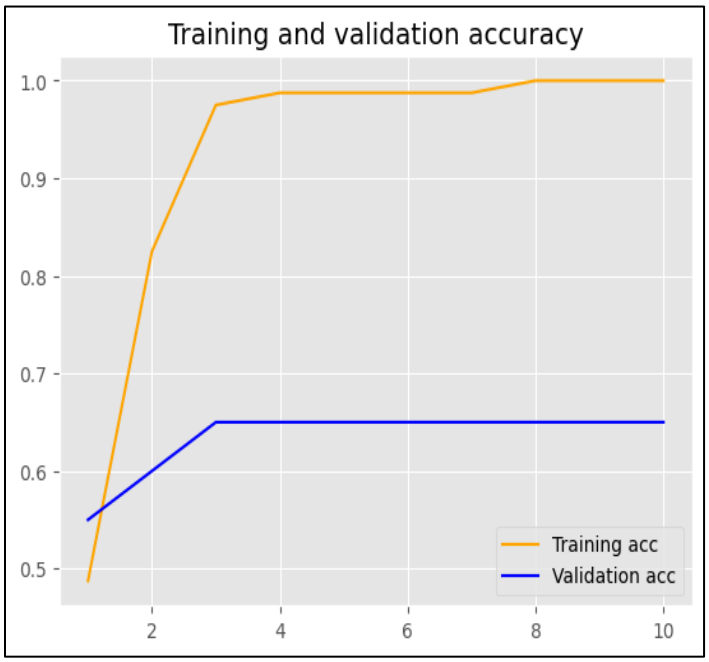
2. pre-trained word embedding layer using the GloVe model.

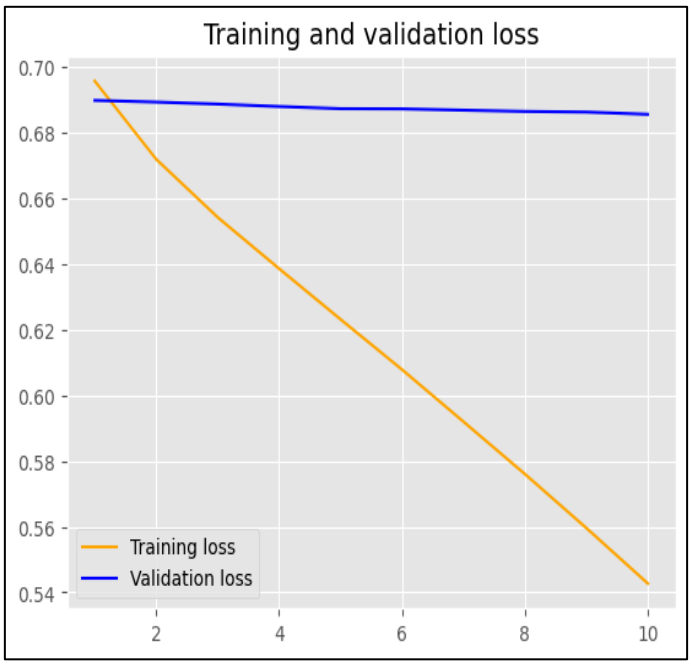
The GloVe pre-trained word embedding model, widely employed in our study, undergoes training on extensive textual data.

- In the IMDB review dataset analysis, I employed two distinct embedding strategies: one utilizing a custom-trained embedding layer and the other integrating a pre-trained word embedding layer. The objective was to evaluate the effectiveness of these strategies. I conducted a comparative analysis of the accuracy achieved by both models across varying training sample sizes, including 100, 5000, 1000, and 10,000.
- Initially, I constructed a custom-trained embedding layer using the IMDB review dataset. Subsequently, each model underwent training with diverse dataset sample sizes, followed by accuracy assessment using a testing set. Subsequently, I compared these accuracy metrics with those of a model that underwent similar testing across different sample sizes, incorporating a pre-trained word embedding layer.

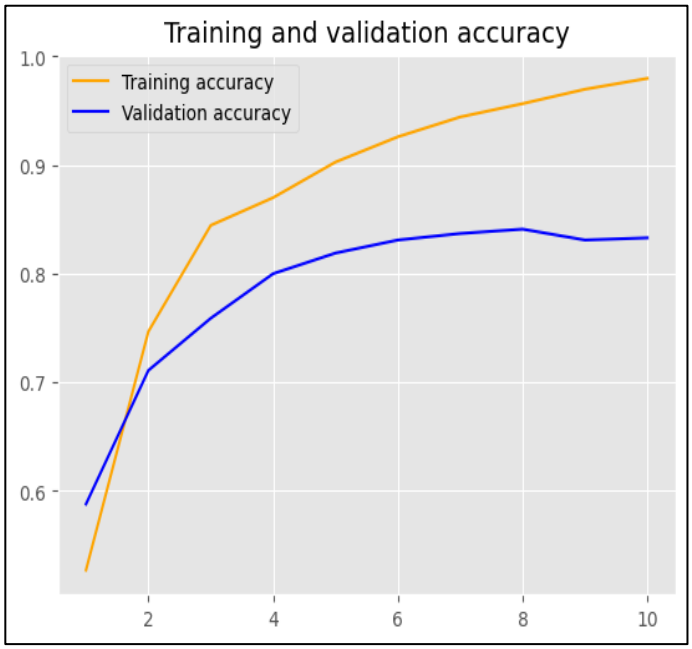
CUSTOM TRAINED EMBEDDING LAYER:

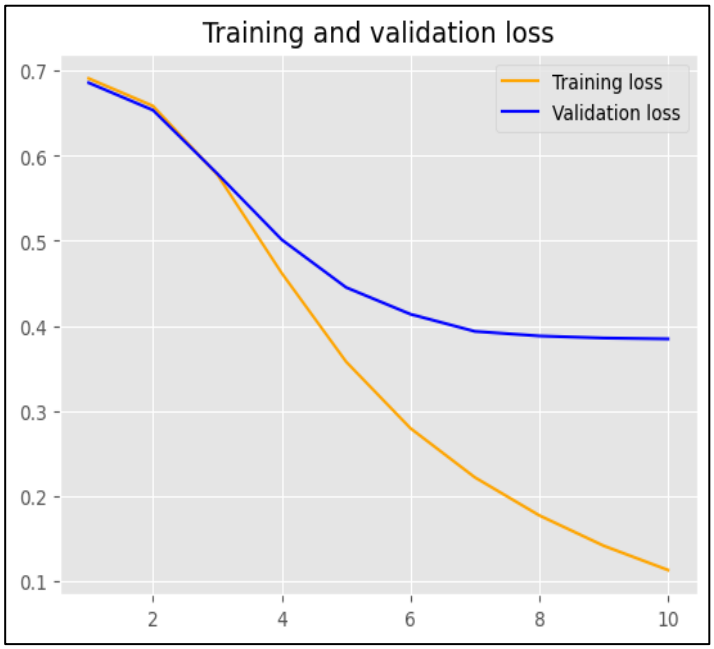
1. Custom-trained word embedding layer with training sample size = 100



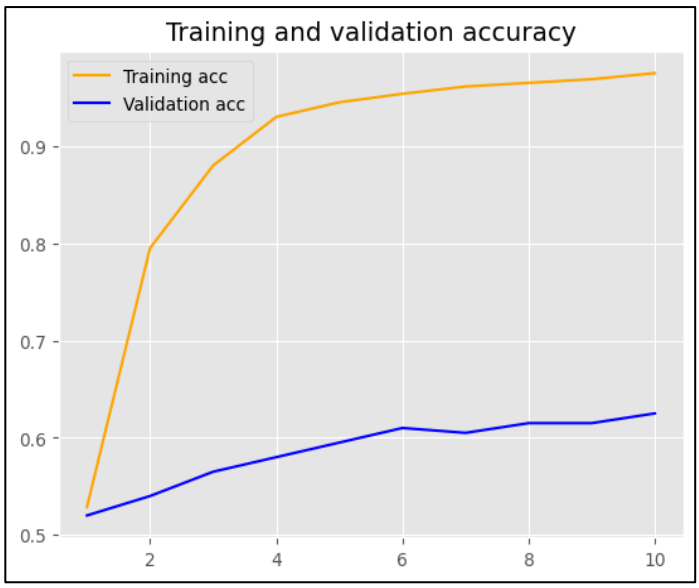


2. Custom-trained word embedding layer with training sample size = 5000

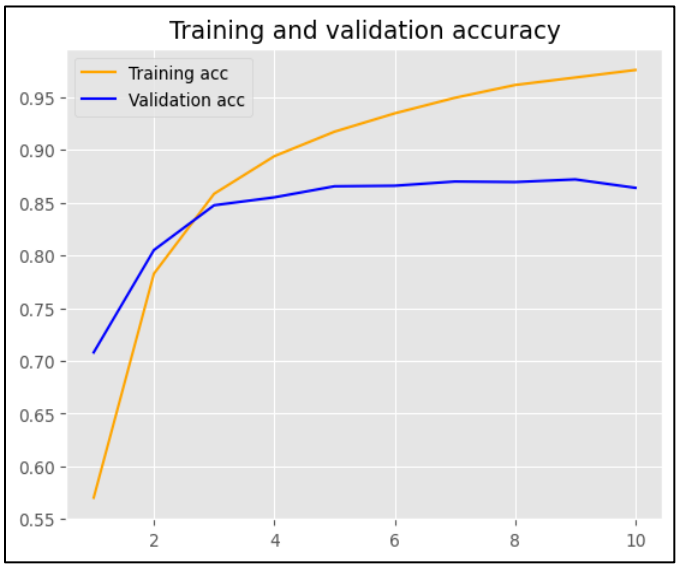


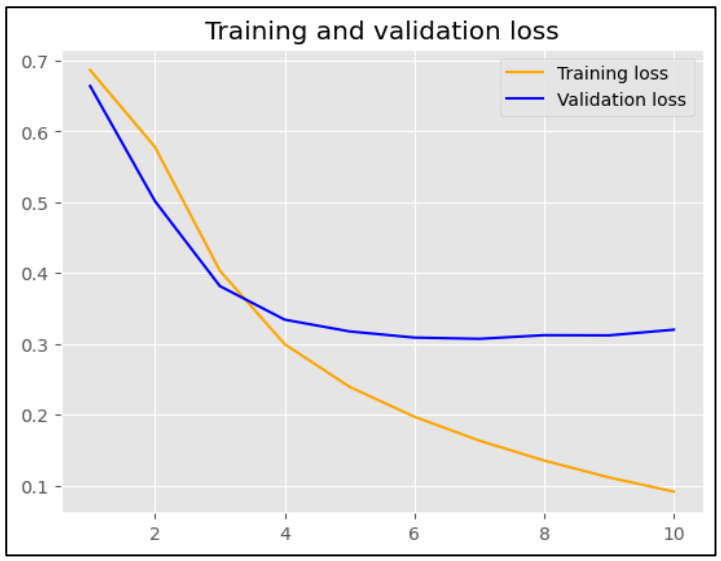


3. Custom-trained word embedding layer with training sample size = 1000



4. Custom-trained word embedding layer with training sample size = 10000

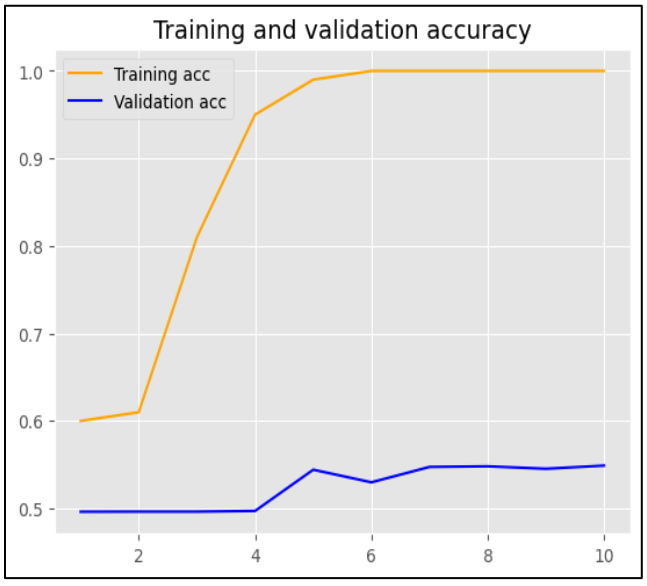


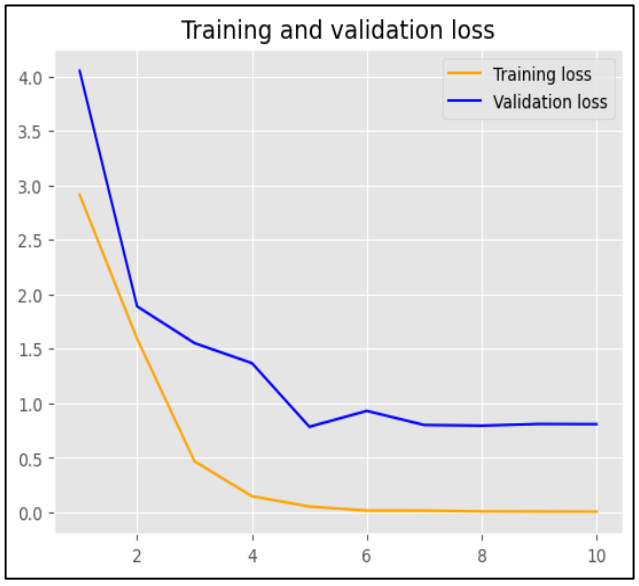


- The training sample size varies from 100 to 10,000.
- As the training sample size increases, both test accuracy and test loss generally improve.
- For instance, with a sample size of 100, the test accuracy is around 50%, but with a sample size of 10,000, the test accuracy increases to around 85.77%.
- Similarly, the test loss decreases as the sample size increases, indicating better model performance.

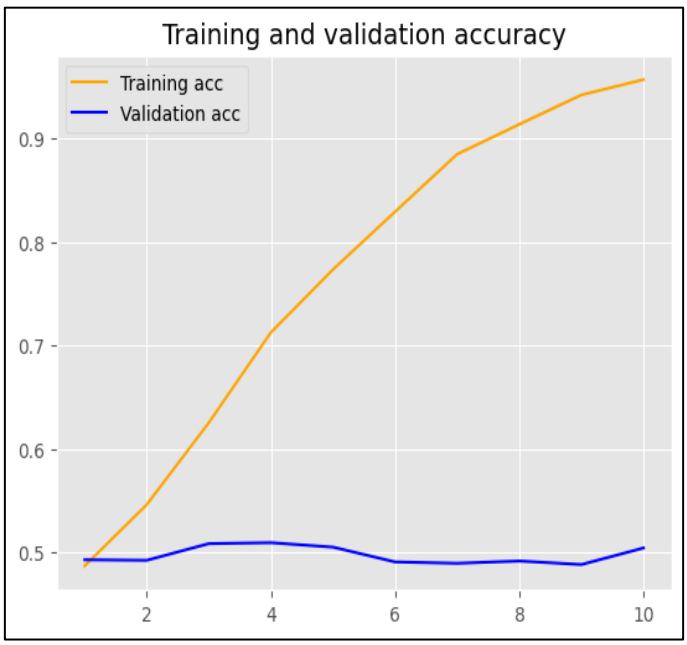
PRE-TRAINED WORD EMBEDDING LAYER:

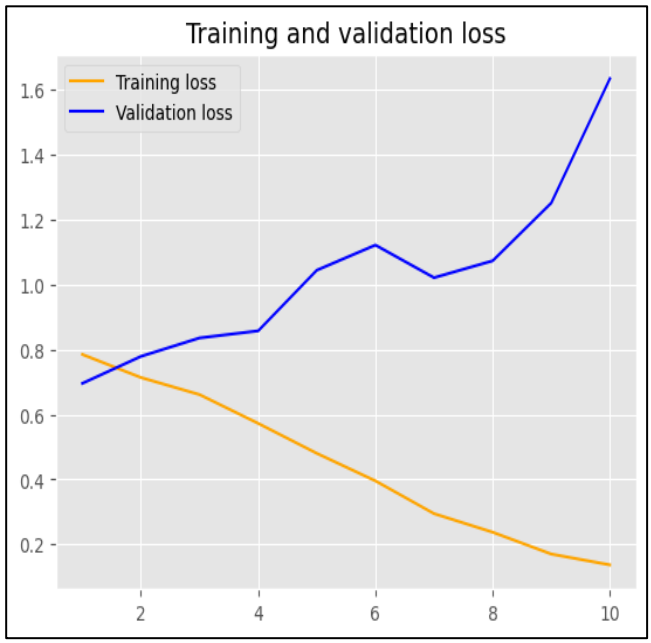
1. Pre-trained word embedding layer with training sample size = 100



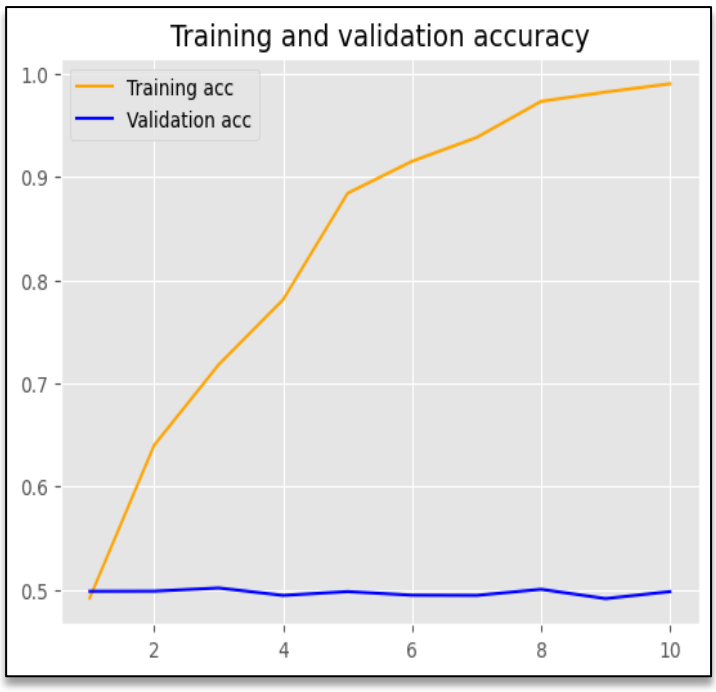


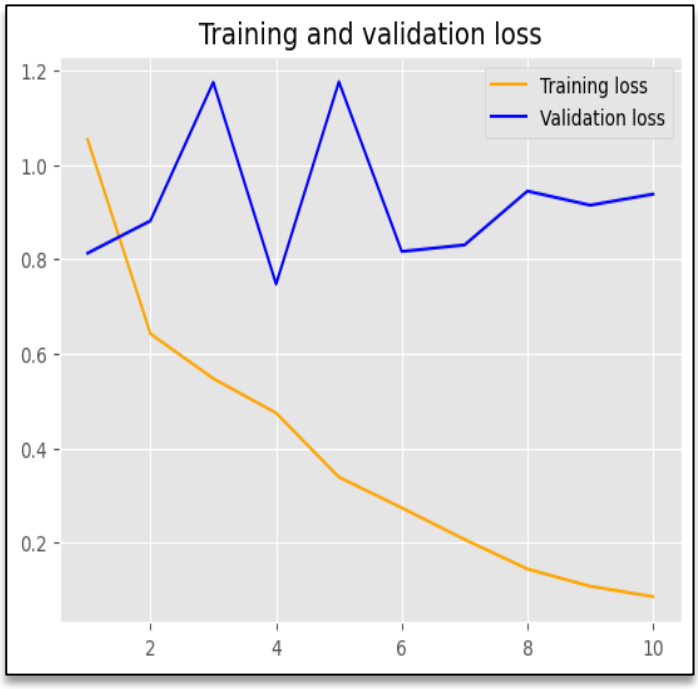
2. Pre-trained word embedding layer with training sample size = 5000



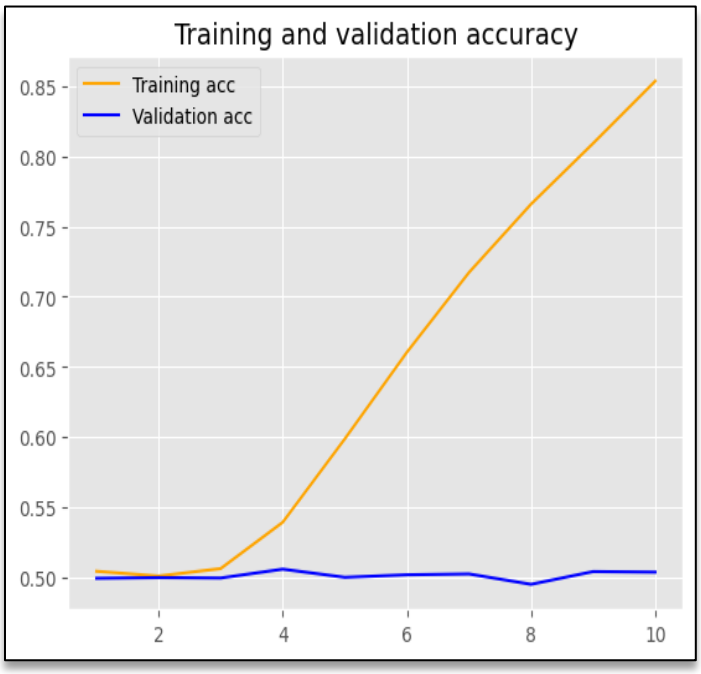


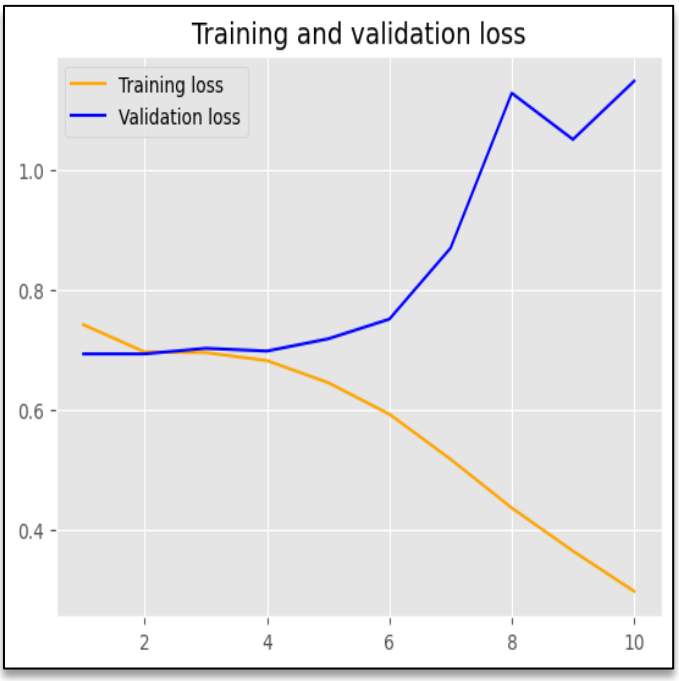
3. Pre-trained word embedding layer with training sample size = 1000





4. Pre-trained word embedding layer with training sample size = 10000





- GloVe (Global Vectors for Word Representation) is a popular method for pre-training word embeddings using global statistics of the corpus.
- Like the custom-trained embeddings, the training sample size varies from 100 to 10,000.
- However, the test accuracy and test loss don't show significant improvements with increasing sample size.
- For instance, even with a sample size of 10,000, the test accuracy remains relatively low, around 49.66%.
- The test loss also remains high compared to the custom-trained embeddings, indicating poorer performance.

Results:

Embedding Technique	Training Sample Size	Test Accuracy(%)	Test Loss
Custom-trained embedding layer	100	49.92	0.69
Custom-trained embedding layer	5000	83.82	0.36
Custom-trained embedding layer	1000	57.87	0.67
Custom-trained embedding layer	10000	85.77	0.33
Pre-trained word embedding (GloVe)	100	50.38	0.85
Pre-trained word embedding (GloVe)	5000	50.16	1.63
Pre-trained word embedding (GloVe)	1000	50.92	0.90
Pre-trained word embedding (GloVe)	10000	49.66	1.14

Conclusion:

In conclusion, the custom-trained embedding layers seem to outperform the pre-trained GloVe embeddings in terms of test accuracy and test loss, especially as the training sample size increases. This suggests that for this task, training embeddings specifically for the task at hand leads to better results compared to using pre-trained embeddings like GloVe.