ADVANCED MACHINE LEARNING

(BA – 64061)

ASSIGNMENT – 4

TEXT AND SEQUENCE DATA

DIVYA CHANDRASEKARAN
ID: 811284790

INTRODUCTION

This study focuses on exploring word embedding methods for sentiment analysis using the IMDB dataset, which consists of 50,000 movie reviews evenly split between positive and negative sentiments. The dataset is divided into 25,000 reviews for training and 25,000 for testing. We truncated reviews to 150 words and limited training samples to 100, considering only the top 10,000 terms. The objective is to compare the performance of different models with varying training samples and embedding layers, all trained with a bidirectional LSTM architecture.

PROBLEM STATEMENT

The primary issue revolves around identifying the most effective method for predicting sentiment accurately within the IMDB dataset, specifically discerning whether a movie review expresses a positive or negative sentiment.

METHODOLOGY

DATASET:

The IMDB dataset comprises movie reviews classified as ***positive or negative*** based on sentiment. Preprocessing involves converting each review into word embeddings, representing each word with a fixed-size vector. The vocabulary is constrained to ***10,000 words***, and reviews are transformed into integer sequences, where each integer corresponds to a unique word. The integers are converted into tensors to prepare the data for neural network input, ensuring uniform length through padding.

In our research, we explored two methods of generating word embeddings for our IMDB review dataset: employing a pretrained word embedding layer based on the GloVe model and creating a custom-trained embedding layer. We utilized extensive text data to train the popular GloVe model, known for its ability to capture both syntactic and semantic relationships between words, making it widely favored for natural language processing tasks.

Specifically, ***we trained the 6B version of the GloVe model on a corpus consisting of Wikipedia data and Gigaword 5, containing 6 billion tokens and 400,000 words***. Our goal was to evaluate the effectiveness of different embedding strategies.

We assessed the accuracies of these models using varying training sample sizes, including 100, 500, 1000, and 10,000 samples. Initially, we developed a custom-trained embedding layer using the IMDB dataset and then tested the accuracy of each model on a separate testing set, trained with different sample sizes. Subsequently, we compared these accuracy results with a model utilizing a pre-trained word embedding layer, which was also evaluated across different sample sizes.

Initially, ***a basic sequence model*** was trained to establish a performance baseline, achieving high training accuracy but slightly lower validation accuracy, indicating potential overfitting to the validation set.

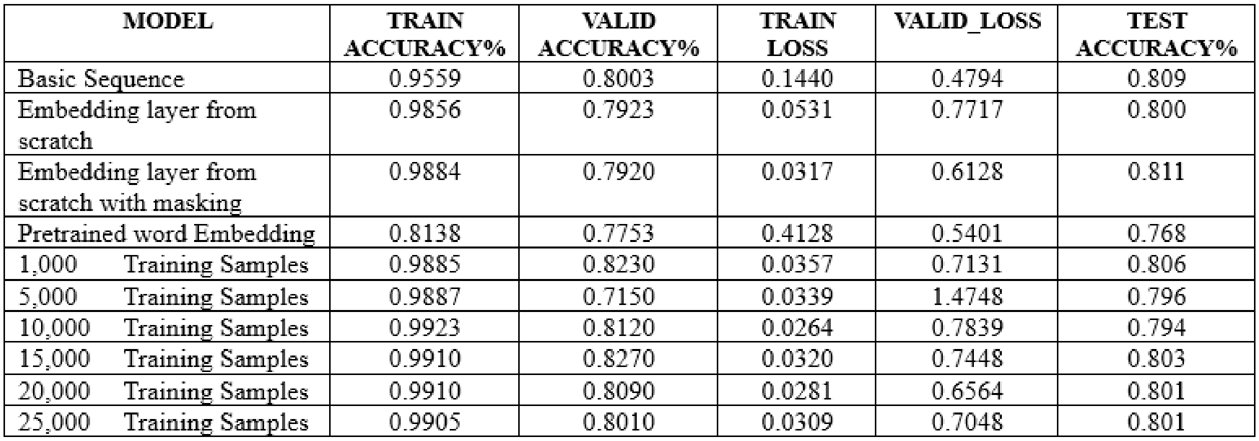
Following that, a ***model was trained from scratch*** using word embeddings without mask activation, resulting in higher training accuracy but lower validation accuracy, indicating overfitting to the training data. It was noted that activating masking could mitigate overfitting and enhance the model's ability to handle varying sequence lengths.

Subsequently, a model was trained by ***embedding the layer from scratch with masking*** with activation of the mask, surpassing the previous model in terms of validation accuracy, highlighting the importance of masking in word embeddings.

A model utilizing ***pre-trained GloVe word embeddings*** was also trained, but it exhibited poorer training accuracy compared to the previous models, suggesting that the pre-trained model didn't capture the dataset's nuances effectively. This underscores the need to explore alternative embeddings or fine-tune pre-trained models for better performance.

Lastly, different training sample sizes were tested to determine the optimal size for training the embedding layer. It was found that using 1000 training samples resulted in high training and validation accuracy while maintaining low training and validation loss.

TABLE



MODEL	TRAIN ACCURACY%	VALID ACCURACY%	TRAIN LOSS	VALID_LOSS	TEST ACCURACY%
Basic Sequence	0.9559	0.8003	0.1440	0.4794	0.809
Embedding layer from scratch	0.9856	0.7923	0.0531	0.7717	0.800
Embedding layer from scratch with masking	0.9884	0.7920	0.0317	0.6128	0.811
Pretrained word Embedding	0.8138	0.7753	0.4128	0.5401	0.768
1,000 Training Samples	0.9885	0.8230	0.0357	0.7131	0.806
5,000 Training Samples	0.9887	0.7150	0.0339	1.4748	0.796
10,000 Training Samples	0.9923	0.8120	0.0264	0.7839	0.794
15,000 Training Samples	0.9910	0.8270	0.0320	0.7448	0.803
20,000 Training Samples	0.9910	0.8090	0.0281	0.6564	0.801
25,000 Training Samples	0.9905	0.8010	0.0309	0.7048	0.801

Based on the data analysis, the "Basic Sequence" model achieved the highest test accuracy at 0.809, while the "1000 Training samples" model, despite using significantly fewer training examples, also performed well with a test accuracy of 0.806. When selecting the most suitable model for a specific task, it's crucial to consider the trade-offs between model performance and the resources required for training. If computational resources and time are limited, the "1000 Training samples" model presents a viable alternative. Conversely, if higher precision is needed and more resources are available, the "Basic Sequence" model may be preferable. Additionally, it was observed that employing a larger training sample size (10,000) consistently yielded the best results across all models.

CONCLUSION

In conclusion, the study highlights the significance of constructing word embeddings from scratch to enhance model training accuracy. It emphasizes the importance of employing appropriate regularization methods like masking to mitigate overfitting. The performance of pre-trained models may vary across different datasets, underscoring the need for exploring diverse embeddings or refining them for better results. The study also indicates that the training sample size plays a crucial role in the performance of the embedding layer. Word embedding remains a fundamental aspect of natural language processing, influenced by factors such as dataset size, regularization techniques, and pre-trained models.

The insights gained from this research can contribute to enhancing model performance not only in sentiment analysis but also in other NLP applications. Notably, the study identifies 1000 as the optimal training sample size for training the embedding layer. Overall, this project provided a comprehensive understanding of word embedding techniques and their application in sentiment analysis.