## Assignment 4

Text and Sequence

## -Veera Manikanta Vallabhaneni

-Divya Gunde

**Summary:**

In this project, we aimed to predict the sentiment of movie reviews (positive or negative) using the IMDB dataset. The dataset consists of 50,000 reviews, but we focused on the top 10,000 words. We experimented with training sample sizes of 100, 2500, 5000, and 100,000, validating on 10,000 samples, and limiting reviews to 150 words. After preprocessing the data, we compared the performance of two embedding strategies.

**Problem:**

The primary objective is to determine which method better predicts the sentiment of movie reviews in the IMDB dataset.

**Technique:**

**Dataset Description:** The IMDB dataset provides sentiment classifications (positive/negative) for movie reviews.

**Preprocessing:** We created word embeddings and number sequences from reviews to facilitate input into the neural network. This involved representing each word as a fixed-size vector, limiting the vocabulary to 10,000 words, converting integers into tensors, and ensuring consistent review length through padding.

**Approach**: We compared a pre-trained word embedding layer (using GloVe) with a custom-trained embed layer. The GloVe model was trained on a vast dataset containing 400,000 words and 6 billion tokens.

# Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Embed Technique | Training Sample Size | Training Accuracy (%) | Test loss |
| Custom-trained embed layer | 100 | 100 | 0.693 |
| Custom-trained embed layer | 2500 | 99.9 | 0.487 |
| Custom-trained embed layer | 5000 | 97.7 | 0.360 |
| Custom-trained embed layer | 10000 | 98.1 | 0.336 |
| Pretrained word embed (GloVe) | 100 | 100 | 0.801 |
| Pretrained word embed (GloVe) | 2500 | 99.9 | 0.021 |
| Pretrained word embed (GloVe) | 5000 | 100 | 0.660 |
| Pretrained word embed (GloVe) | 10000 | 99.8 | 0.019 |

**CUSTOM-TRAINED EMBED LAYER**

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

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

2. Custom-trained embed layer with training sample size = 2500

**A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated**

3. Custom-trained embed layer with training sample size = 5000

**A graph with blue dots

Description automatically generatedA graph of training and validation

Description automatically generated**

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

A graph of training and validation accuracy

Description automatically generatedA graph of training and vallation loss

Description automatically generated

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

**A graph with blue dots

Description automatically generatedA graph with blue lines and dots

Description automatically generated**

1. **Pre-trained word embed layer with training sample size = 2500**

**A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated**

1. **Pre-trained word embed layer with training sample size = 5000**

**A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated**

1. Pre-trained word embed layer with training sample size = 1000

**A graph with blue dots

Description automatically generated**A graph with blue dots

Description automatically generated

**Conclusion:**

The size of the dataset required accuracy, and available computing power all play an important role in which custom-trained or pre-trained embedding layers are selected. Custom embeddings perform much better with bigger datasets, but pre-trained embeddings are still a good choice when data is scarce. To conclude, the choice must be in line with the demands of the work at hand.