Assignment 4

Troy Toth

Purpose:

The purpose of this assignment is to apply Recurrent Neural Networks (RNNs) or Transformer-based models to text and sequence data in order to evaluate how different sequence modeling architectures perform under data-limited conditions. This exercise also aims to demonstrate methods for improving performance when the amount of labeled data is small. For this, we will again be utilizing the IMDB database with a few modifications stated below.

Data Preparation:

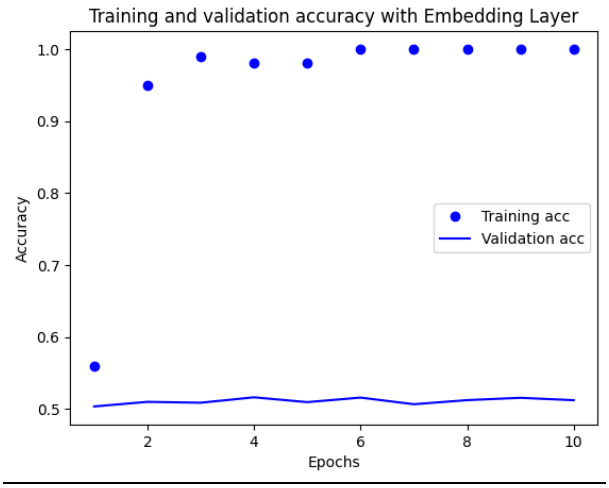
- Dataset: IMDB movie review sentiment dataset (25,000 training, 25,000 test).
- All reviews truncated or padded to 150 words.
- Only top 10,000 most frequent words retained.
- **Training Set:** Restricted to 100 samples to simulate extreme low-data conditions.
- **Validation Set:** 10,000 samples.

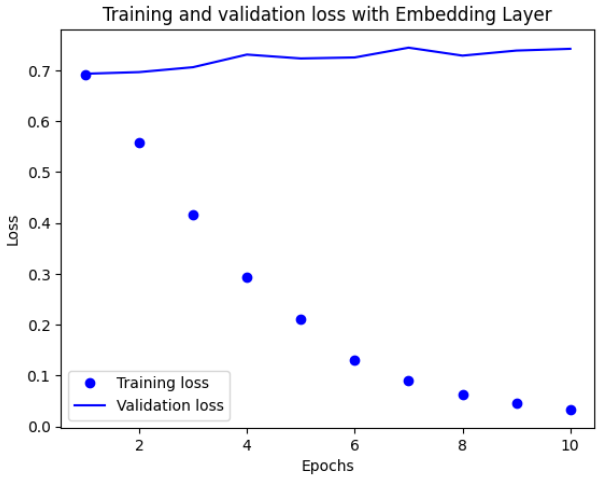
Models:

Two embedding approaches were compared:

- Trainable embedding layer
- Pretrained embedding (GloVe 100d)
    - Embedding weights loaded and either frozen or fine-tuned.

Results:





- Pretrained embedded Initial Model: 50.98%
- Embedded Refined Model: 82.32%

Conclusion:

The comparison between pretrained word embeddings and a trainable embedding layer shows a clear performance difference. The initial model using pretrained embeddings achieved an accuracy of 50.98%, indicating that although pretrained vectors provide general semantic knowledge, the model struggled to adapt effectively to the specific IMDB sentiment classification task under these conditions. In contrast, the refined model using a trainable embedding layer reached a significantly higher accuracy of 82.32%. This suggests that when provided with sufficient training adjustments (e.g., optimized hyperparameters, refined architecture, or increased training data), the model benefits more from learning domain-specific embeddings rather than relying solely on generic pretrained vectors. Overall, the results indicate that custom-trained embeddings ultimately offered better task-specific performance, demonstrating the value of learning representations directly from the IMDB dataset.