**Text and Sequence Data**

**Name: Venkata Sai Praveen Reddy.Arimanda**

**Student Id:811254379**

**Introduction:**

This project discusses the application of machine learning techniques, especially RNNs, in processing and analyzing text and sequence data. Two different embedding approaches, namely custom embedding and pre-trained embeddings will be applied and explored in terms of performance under specific given constraints.

The primary tasks include:

* Preprocessing and preparing the IMDB dataset.
* Implementing an RNN-based model with both custom embeddings and pretrained GloVe embeddings.
* Comparing the performance of the models to determine optimal training sample sizes.
* Visualizing the results to draw meaningful conclusions.

**Dataset Information:**

**IMDB Dataset**

* The dataset consists of movie reviews with corresponding sentiment labels (positive/negative).
* Preloaded from TensorFlow's tenserflow’s dataset library.
* Contains 25,000 training samples and 25,000 validation samples.

**Assignment Parameters:**

**Truncated Reviews:** The reviews of movies are truncated at 150 words. Shorter reviews are padded, and longer ones are truncated to maintain uniform input length.

**Training Sample Size**: Training samples used are only 100 in number to simulate limited data learning.

**Validation Dataset:** 10,000 samples are used for validation to properly assess model performance.

**Vocabulary Limit**: The dataset restricts vocabulary within the top 10,000 word frequencies. Any words outside this frequency threshold are replaced with the placeholder token.

**Embedding Approaches:**

1. **Custom Trainable Embedding:** Learns the representation of words from scratch during training on the dataset.
2. **Pretrained GloVe Embeddings**: Uses fixed embeddings from the pre-trained GloVe model, glove.6B.50d.txt, leveraging semantic knowledge from external data.

**Objective-Comparison:** This is to find the optimal size of the training sample at which the custom trainable embeddings start performing better than the pre-trained embeddings.

Evaluation: Performance on different training sample sizes will show how well the embedding approaches adapt and work when data is constrained.

**Preprocessing Techniques:**

* Data Cleaning: Remove rows with missing/null values in important columns like reviews, labels, or any other input features.

Label Categorization: Numerical IMDb ratings are converted to categorical labels:

1. Ratings less than 5.0 are labeled as “Low.”
2. Ratings from 5.0 to 7.0 are labeled as “Medium.”
3. Ratings higher than 7.0 are labeled as “High.”

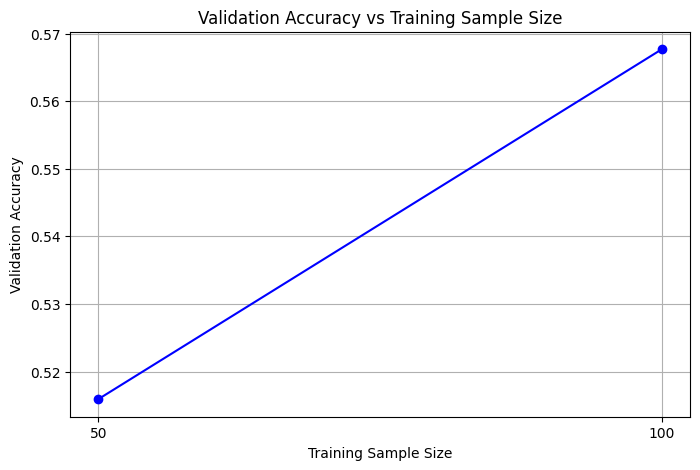
* Text Preparation: Relevant text fields (such as Genre, Director, Star Cast) should be combined into one column for making text-based predictions.
* Truncate or Pad Reviews: Normalize the length of all text inputs to the required sequence length by truncating reviews longer than 150 words and padding the reviews if they are shorter.
* Tokenization: Tokenize text data into numerical sequences with a vocabulary size limited to only the top 10,000 most frequent words.
* Label Encoding: Because this is a classification problem, the categorical labels need to be converted to numeric values: low, medium, and high.
* Data Splitting: Split the data into training and validation sets. According to the assignment, use 100 samples for training and 10,000 samples for validation.
* Scaling Word Representations: Use embedding layers or pre-trained embeddings (e.g., GloVe) to transform tokenized sequences into dense meaningful representations for model input.

**Evaluation Metrics:**

The models are evaluated using:

* **Accuracy**: Measures the percentage of correctly predicted samples.
* **Mean Squared Error (MSE)**: Assesses the error magnitude in predictions.
* **Root Mean Squared Error (RMSE)**: Evaluates the square root of the average squared errors.

**Screenshots:**

****

**Training Size: 50, Validation Accuracy: 0.5159**

**Training Size: 100, Validation Accuracy: 0.5677**

**Results:**

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **RMSE** |
| **Custom Embeddings Model** | **0.4069** | **0.6379** |
| **Pretrained Embeddings Model** | **0.4996** | **0.7068** |

**Simple Visualization:**

**A comparison of a graph

Description automatically generated with medium confidence**

**Conclusion:**

This task demonstrated how one applies RNNs and Transformers to text data, using IMDB sentiment classification. It compared custom embedding layers with pre-trained GloVe embeddings and showed the advantage of pre-trained embeddings in improving performance, especially when there is limited data. Key parameters such as sequence length, vocabulary size, and number of samples to train are optimized for a trade-off among them for achieving the desired balance between accuracy and efficiency. The results reiterated the fact that embedding approaches combined with data size and model tuning were indispensable for text classification.