Assignment 4: Applying RNN to Text and sequence data.

This code creates an RNN model to analyze the sentiment of movie reviews. It utilizes pre-trained GloVe embeddings and is trained on a portion of the IMDb dataset while evaluating its performance on the test set. The program also examines how the size of the dataset affects the performance of the model.

The process involves training the model on various subsets of the training dataset with an increasing number of samples. The performance metrics for each subset are recorded and analyzed to determine the impact of dataset size on the final performance of the model.

Briefly About RNN and GloVe

RNN

RNN is capable of processing sequential data, which is commonly used in many applications like speech recognition, natural language processing, video analysis and time series forecasting.

The quality and amount of the training data, selection of hyperparameters and model complexity impacts on how well RNN performs.

GloVe

GloVe(Global Vectors for Word Representation) is used for word embeddings, which are dense vector representation of words in a high dimensional space. GloVe embeddings are particularly useful for natural language processing (NLP) tasks like text classification, sentiment analysis, and machine translation that require comprehending the semantic relationships between words.

The IMDb dataset is loaded and only the first 100 samples are used for training in the beginning. Later different training sets were taken 800, 1500, 2000 , 2500 were different training sets used.

The GloVe embeddings that were pre-trained are extracted and limited to a maximum of 10,000 words, and the reviews' maximum length is set to 150.

Used the Sequential API from Keras, in which the LSTM model is then defined. The first layer is an embedding layer that maps each word in the reviews to its corresponding embedding using the pre-trained embedding matrix as input. The second layer is an LSTM layer that has a 32-unit and 0.5 dropout rate to prevent overfitting. For dense layer used sigmoid activation function to output the positive or negative review probability.

For the model training the RMSprop optimizer and binary cross-entropy loss function were taken. The model is trained on the training dataset for 10 epochs, batch size of 32, and the validation is performed on 10,000 samples from the test dataset.

The Following table shows the Test Accuracy of First Basic Sequence Model, Embedding Layer from Scratch, embedding layer Mask Enabled and Pretrained Embedding performances when used different training samples:

Training samples taken	First Basic Sequence Model	Embedding Layer from Scratch	Embedding layer Mask Enabled	Pretrained Embedding performances
100	0.808	0.782	0.776	0.774
800	0.832	0.827	0.836	0.839
1500	0.841	0.821	0.837	0.840
2000	0.847	0.835	0.843	0.839
2500	0835	0.829	0.840	0.837

Conclusion

From the results it demonstrated that effectiveness of pretrained embedding and dataset size is important in the Sentiment analysis tasks. Performance of the RNN model is impacted by the size and quality of the training dataset and hyperparameter selection. Model using the Pre-trained GloVe embeddings has the capability of finding High accuracy.