Assignment 4: Applying RNN to Text and sequence data.

This script designs an RNN architecture aimed at evaluating the sentiment conveyed in movie reviews. It harnesses pre-existing GloVe embeddings and undergoes training on a subset of the IMDb dataset, subsequently gauging its effectiveness on a test set. Additionally, it delves into the impact of dataset size variations on the model's performance, systematically assessing performance metrics as the dataset's magnitude fluctuates.

This entails training the model on different portions of the training data, gradually increasing the sample size. Performance metrics are then logged and scrutinized for each subset to ascertain how dataset size influences the model's ultimate performance. Regarding RNN and GloVe, they're briefly summarized as follows: RNN aids in processing sequential data, while GloVe provides pre-trained word embeddings, offering a way to represent words numerically in a model.

RNN

RNN excels in handling sequential data, a feature widely applied across diverse fields such as speech recognition, natural language processing, video analysis, and time series forecasting. The effectiveness of RNN hinges on factors like the quality and quantity of training data, choice of hyperparameters, and model complexity. These elements collectively shape the performance of RNN models.

GloVe

GloVe, an abbreviation for Global Vectors for Word Representation, functions as a tool for creating word embeddings, representing words numerically in a compact, multi-dimensional space. These embeddings, powered by GloVe, provide substantial assistance across a spectrum of natural language processing (NLP) applications including text classification, sentiment analysis, and machine translation. They play a vital role in understanding the semantic connections between words, which is pivotal in NLP tasks.

The IMDb dataset is initially loaded, with only the first 100 samples utilized for training at the outset. Subsequently, various training sets of sizes 800, 1500, 2000, and 2500 are employed. Pre-trained GloVe embeddings are extracted and constrained to a maximum of 10,000 words, while the reviews are capped at a maximum length of 150 characters.

The Sequential API from Keras is employed, defining an LSTM model. The initial layer is an embedding layer, associating each word in the reviews with its corresponding embedding via the pre-trained embedding matrix. Subsequently, an LSTM layer follows, featuring 32 units and a dropout rate of 0.5 to mitigate overfitting. A dense layer with a sigmoid activation function is utilized to output the probability of positive or negative reviews.

For training, the RMSprop optimizer and binary cross-entropy loss function are utilized. The model undergoes training on the training dataset for 10 epochs, with a batch size of 32, and validation is conducted on 10,000 samples from the test dataset.

The table illustrates the Test Accuracy of the Initial Basic Sequence Model, Embedding Layer Constructed from Scratch, embedding layer with Mask Enabled, and Pretrained Embedding across varying training sample sizes.

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

The findings underscore the significance of both pretrained embeddings and dataset size in sentiment analysis tasks. The performance of the RNN model is notably influenced by the quantity and quality of the training data, as well as the selection of hyperparameters. Models utilizing pre-trained GloVe embeddings exhibit the potential for achieving high accuracy in sentiment analysis tasks.