Assignment 4: Text and sequence

This assignment uses the IMDB dataset to look at text and sequence data in the context of sentiment analysis. The primary focus is understanding the effects of varying sample sizes and exploring the impact of different embedding layers. First, the IMDb data sets must limit the reviews to 150 words and restrict the training data to 100 samples. there is a validation set of 10,000 samples, only the top 10,000 words are considered for this analysis.

In our analysis, we look at a few different models. Model One is configured with an embedding layer featuring a dimensionality of 32, while Model Two adopts a simpler architecture without an embedding layer. The primary focus is to objectively discern the significance of embedding layers in text sequences.

Shifting to the training dynamics, both models undergo a 30-epoch training process with a batch size of 32. The inclusion of the ModelCheckpoint callback ensures the retention of the best-performing models during training. This approach is trying to be efficient and capture optimal model states for evaluation.

The core analysis revolves around comparing models with 100, 5000, and 10000 training samples. Key metrics such as training accuracy, validation accuracy, and validation loss are scrutinized to understand the models' behavior with varying dataset sizes.

Model 1:

The first model, features an embedding layer with a dimensionality of 32, training dynamics over 30 epochs. The training accuracy steadily increased from an initial 45% to a perfect 100%. However, the validation accuracy plateaued at 52.09%, suggesting an issue of overfitting.

Model 2:

In contrast, the second model, which omitted the embedding layer, displayed efficient training dynamics. Starting with an accuracy of 45% in the first epoch, it reached 100% accuracy on both the training and validation sets by the final epoch, indicating strong generalization performance.

Model 3:

The third model, also without an embedding layer, exhibited similar training trends. It started with an accuracy of 45%, reaching 100% accuracy on the training set and 52.34% on the validation set by the end of the 30-epoch training process.

After looking at the performance of these three models it is evident that model 1 despite achieving perfect training accuracy struggles to generalize to the validation set which could be due to overfitting. on the other hand models 2 and 3 both reach 100% accuracy on training and validation sets. The choice of inclusion or exclusion of embedded layers appears to influence training performance and generalization.

This assignment explores sentiment analysis using textual sequences. The initial steps involved data preparation, limiting reviews to 150 words, and employing a training set of 100 samples. The comparative study of two models the first incorporating an embedding layer of dimensionality 32, while the second opted for a simpler architecture without an embedding layer. The training dynamics involved a 30-epoch process with a batch size of 32, utilizing the ModelCheckpoint callback to preserve the best-performing models. Fine-tuning hyperparameters exploring diverse architectures and experimenting with alternative embedded layers are all potential options for further investigation.