Report for Language Modelling

Srijan Chakraborty
2020115001

Introduction:

This report presents the results of a classification task carried out using a PyTorch LSTM model on the English-Atis dataset, located at ud-treebanks-v2.11/UD_English-Atis/en_atis-ud-{train, dev, test}.conllu., where the goal was to assign one or more labels to an input based on its characteristics. The model was trained on sentences labelled with different part-of-speech tags, and its objective was to predict these tags when given a sentence as input. The dataset was split into training and test sets, and the model's performance was evaluated based on its accuracy in predicting part-of-speech tags on the test set. The classification report provides a summary of the model's precision, recall, and accuracy for each dataset.

The outcomes reveal the strengths and weaknesses of the model, which can guide future performance improvements.

Results:

The following table shows the performance of a PyTorch LSTM model trained on annotated text data for a classification task, where the goal is to predict part-of-speech tags. The model's accuracy in predicting the tags on a separate test dataset was evaluated, and the results are presented in terms of precision, recall, and accuracy for each dataset.

The table presents the performance of a PyTorch LSTM model trained on annotated text data for a part-of-speech tagging classification task. The evaluation metrics used to measure the model's performance are precision, recall, and accuracy. Precision measures the percentage of true positives (correctly predicted labels) out of all positive predictions made by the model. Recall measures the percentage of true positives out of all actual positive labels in the dataset. Accuracy measures the percentage of correct predictions made by the model out of all predictions.

	Accuracy	Precision	Recall	F1
Train Set	0.8819	0.9263	0.9361	0.9361
Validation Set	0.8812	0.9267	0.9411	0.9231
Test Set	0.8897	0.9315	0.9415	0.9266

Inferences

Overall, the model's accuracy on the majority of classifications is 0.88, which is good. The model's precision, recall, and F1-score are all greater than 0.92 for all three datasets. This indicates that the model is correctly predicting the labels for all the tags. We must keep in mind that this prediction is after handling unknowns (any word that has occurred less than 2 times (0 or 1 time/s) in the dataset. Thus, we can be confident that our Bi-Directional LSTM, which we have included in our implementation, works well for the dataset that is given. We also see a high Recall percentage, and thus, we can be assured that our Bi-Directional LSTM gives fewer false negatives/positives. Because of the high F1 score, we can be rest assured of our correctness for our Precision as well as our Recall values, as a perfect F1 score of 1.0 means that the model has achieved both perfect precision and recall, which indicates that it has made no incorrect predictions. With a score of ~0.94, we are assured of the correctness of our values.