NLU project exercise lab: 10

Mateo Rodriguez (239076)

University of Trento

md.rodriguezromero@studenti.unitn.it

The objective of the lab was to compare the performance of 4 different models, some based on LSTMs and one based on BERT. The dataset used was the ATIS dataset which contains labeled slots and intents and the models were trained to predict both at the same time. Slot and intent prediction performance was the highest with the BERT based model.

1. Introduction

- For the first part of the lab, the three models used were: a base model (ModelIAS) composed of an LSTM followed by two fully connected layers, the same model but now with a bidirectional LSTM, and the last modification was adding dropout to this bidirectional model. The task in hand is slot and intent classification using the ATIS dataset.
- For the second part a pre-trained BERT model was used along side hidden layers to predict both slots and intents at the same time.

To maintain the variables as similar as possible the loss used for all the models was the same: a sum of the cross entropy for intents and for slots. The models in the first part used an embedding size of 300 and a hidden size of 200. Batch size of 128 was used in all the trainings, maxing out available computing power.

2. Implementation details

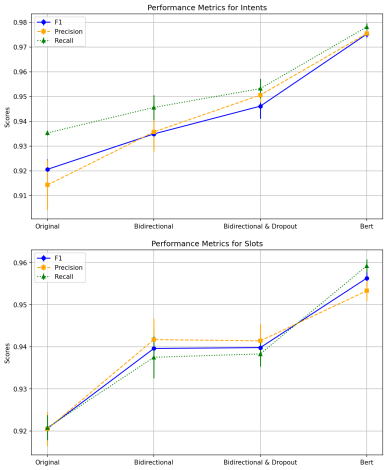
The 4 models implemented in both parts of the lab are relatively straight forward. The original model (and the 2 variations tested) are based on LSTMs to process the text which has been already outperformed in various NLP tasks by transformer architectures like BERT. All models were trained using Adam. Regarding evaluation it was decided to use F1, precision and recall since they evaluate performance in different ways, taking into account the ability to identify positive instances, avoid false positives, and handle imbalances in the dataset.

To be able to make the BERT model train for two tasks at once it was necessary to use two different linear layers following the outputs of BERT. The first linear layer takes as input the last hidden state of BERT's CLS and is used to obtain the intent predictions. The second linear layer uses instead the whole last hidden layer. The input for both branches passes through a dropout of 0.3 before being sent to each pipeline.

Slot prediction required more processing so it was necessary to add extra modifications to this branch. After various tests the final configuration I ended up with was 1 linear layer for the intents predictions and 3 linear layers for the slots prediction, in these slots branch layers I added batch normalization and Mish as an activation function to add non-linearity, alongside dropouts of 0.4 in between.

3. Results

As mentioned before the metrics used to evaluate the models were precision, recall and F1 score. The following figure shows the scores of each of the model variations tested.



As evidenced in the figures, as the complexity of the model increases the metrics for the intent predictions increase almost linearly, which is to be expected. LSTMs seem to reach a cap in slot prediction, going from the bidirectional configuration to the bidirectional plus dropout only variance is reduced.

As expected and hinted at the beginning of this report, BERT outperforms all the other models by a significant margin specially in slot prediction. But to achieve this performance a small prepossessing had to be applied: since BERT utilizes the WordPiece tokenizer, a word can be tokenized as sub parts of itself, impacting the prediction of slots which are dependent on the position of words in the sentence. Without this fix BERT cannot surpass the performance of LSTMs in the task of predicting slots, in fact it performs worse. To work around the tokenization problem it was necessary to basically remove tokens that started with ##, effectively keeping only the first part of a word in case it was subdivided, as mentioned in [1].

Specific scores can be found annexed below.

4. References

[1] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," *arXiv preprint arXiv:1902.10909*, 2019. [Online]. Available: https://arxiv.org/abs/1902.10909

Model	F1	Precision	Recall
Original	S: 0.9268 $\pm$ 0.004 I: 0.9198 $\pm$ 0.003	S: 0.9263 $\pm$ 0.005 I: 0.9139 $\pm$ 0.008	S: 0.9273 $\pm$ 0.004 I: 0.9359 $\pm$ 0.002
Bidirectional	S: 0.9448 $\pm$ 0.001 I: 0.9475 $\pm$ 0.006	S: 0.9466 $\pm$ 0.002 I: 0.9507 $\pm$ 0.006	S: 0.9429 $\pm$ 0.003 I: 0.9550 $\pm$ 0.005
Bidirectional & Dropout	S: 0.9456 $\pm$ 0.003 I: 0.9575 $\pm$ 0.004	S: 0.9459 $\pm$ 0.004 I: 0.9595 $\pm$ 0.004	S: 0.9453 $\pm$ 0.004 I: 0.9624 $\pm$ 0.004
Bert	S: 0.9563 $\pm$ 0.002 I: 0.9752 $\pm$ 0.0013	S: 0.9533 $\pm$ 0.0024 I: 0.9754 $\pm$ 0.0011	S: 0.9592 $\pm$ 0.0016 I: 0.9781 $\pm$ 0.0015

Table 1: *Results of the different models*