ACTS - Assignment 1

Rodrigo Alejandro Chavez Mulsa

In this assignment, I managed to replicate the results from the paper Supervised Learning of Universal Sentence Representations from Natural Inference Data.

To do this I trained the proposed 4 models from the assignment, an average word embedding, an LSTM, a BILSTM, and a BLSTM with Max Pooling. In the following table, we can see the scores that the models achieved when training in the SNLI data. In order to get as close as possible to the authors, I used a fully connected MLP with 512 hidden units as a head without nonlinearities.

Model	NLI-dev	NLI-test	Transf-micro	Transf-macro
AWE	0.6173	0.6283	82.573	79.129
LSTM	0.791	0.7834	79.894	78.337
BILSTM	0.7935	0.7948	83.36	82.185
BILSTM-MAX	0.834	0.8333	87.075	84.95

Error Analysis

For the analysis, I focused on some examples where depending on the tokenization the predictions are different. This is the case of the example: `"A man is typing on a machine used for stenography."/"The man is'nt operating a stenograph."` where if we change “isn’t” to “is not” the models correctly predict it as a contradiction.

Overall the results show me that the AWE performs the worst, the LSTM and BILSTM show not much of a difference, and the BILSTM-MaxPooling performs better by increasing the amount of correct entailment and neutral predictions but it does not improve over the contradiction examples. Bellow is a table with the predictions per class from the dev set.

Model	Acc Entailment	Acc Contradiction	Acc Neutral
AWE	0.43	0.75	0.62
LSTM	0.84	0.80	0.73
BILSTM	0.81	0.80	0.75
BILSTM-MAX	**0.87**	0.79	**0.81**

Finally, I would like to note that even though the models ran for multiple hours in RTX GPUS the best models were selected based on the smallest evaluation loss which happened after a few epochs (approx. 3) for the LSTM based models while for the AWE the model did improve over a longer run.