ATCS Practical

Despoina Touska - 15216608
despoina.touska@student.uva.nl

April 22, 2024

1 Introduction

The objective of this assignment is to learn general-purpose sentence representations through a Natural Language Inference task. For this scope, there were implemented four different neural models for sentence classification:

- Average Word Embeddings (baseline)
- Unidirectional Long Short-Term Memory (LSTM) networks
- Bidirectional LSTMs
- Bidirectional LSTMs with Max Pooling (a variant to capture the most important information)

The models were trained on a large dataset called Stanford Natural Language Inference (SNLI). For evaluation, Facebook's SentEval framework was used. This framework tests our models on entirely new tasks they haven't seen before. Our goal is to confirm the results of a previous study, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data" by Conneau et al. (2017).

2 Reproducibility of the paper results

This section provides a description of the results of the SNLI task. The results are visualized in the analysis.ipynb. Firstly, regarding the Average Word Embedding (AWE) model, the results depict that it performs significantly better than random guessing with a test accuracy of around 68%. This simple model is able to capture some form of semantic entailment/contradiction information.

Regarding the Unidirectional LSTM (LSTM) model, we see a significant boost in performance compared to AWE. The test accuracy increased to around 76%, which means that it can capture better the semantic entailment/contradiction information. This can be attributed to the fact that an LSTM uses a history of words and hidden states to create a complex combination of word embeddings that takes word order into account.

Interestingly, when upgrading to a Bidirectional LSTM (BiLSTM Last) that uses the last hidden states of the forward and backward passes concatenated as the sentence representations(BiLSTM) we see no improvement in dev and test accuracy, compared to the Unidirectional LSTM. The test accuracy remains the same, around 76.79%.

Regarding, Bidirectional LSTM (BiLSTM Max) that uses max-pooling over the word representations, we see another improvement in test accuracy by about 3-4% compared to the UniLSTM and BiLSTM models. The test accuracy reaches 83.34%.

In conclusion, the BiLSTM Max model is the best model with around 83% test accuracy on the SNLI task. These results align with the one in the paper.

3 SentEval Evaluation

Regarding the results of SentEval, the BiLSTM Max model is the best model compared to the other models. It achieves 82.3 % and 77.1 % in the metrics Micro and Macro respectively. Then surprisingly the AWE follows with 81.0% and 75.4% in Micro and Macro respectively. In the third and fourth position we can find the BiLSTM Last and Unidirectional LSTM respectively. So, the trend remains the same in all models but AWE, which achieves the 2nd place from the last position in the SNLI task.

4 Qualitative Analysis

In this section, two different examples of performance failure are going to be analyzed. The first example is: Premise - “Two men sitting in the sun”, Hypothesis - “Nobody is sitting in the shade”, Ground Truth - Neutral, Model Prediction - Contradiction. This particular example presents a challenge for models such as AWE due to the difficulty of capturing the negation within the hypothesis. The Unidirectional LSTM encoder might also struggle with this scenario since the negation is positioned at the beginning of the sentence. Moreover, the LSTM encoder is likely to prioritize the latter token of the sentences (”sun” and ”shade”), which potentially can lead to the creation of contradictory sentence representations. Similarly, the other two encoders may encounter difficulties because of the difficulty created by the negation within the hypothesis.

The second example is: Premise “A man is walking a dog”, Hypothesis - “No cat is outside” Ground Truth - Neutral, Model Prediction - Contradiction. In this example, the encoders will probably predict a contradiction, as they might interpret the words ”cat” and ”dog” similarly, potentially leading to a perceived contradiction between premise and hypothesis. The BiLSTM encoders (with and without max pooling) are likely to perform better by capturing the context of the entire sentence. Similarly, with the previous example, the negation may also pose some difficulties in the models.

The results are in the last table of the analysis.ipynb notebook, confirms the previous explanation, that the examples pose difficulties in the model to interpret. This is apparent if we observe that all the models predict contradiction for both examples.