Exercise 4

Refael Shaked Greenfeld 305030868,

Danit Yshaayahu 312434269

# Which paper you chose to implement

Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention by Yang Liu, Chengjie Sun, Lei Lin and Xiaolong Wang

# Why you chose that particular one.

First, we tried to find a model without many parameters because less parameters would decrease the computational complexity and reduce running time (which is important when you have limited computational resource.

Also, we wished to experience with Attention mechanism because we didn’t have the opportunity to do so in the course. We believe it is fundamental skill we wished to acquire.

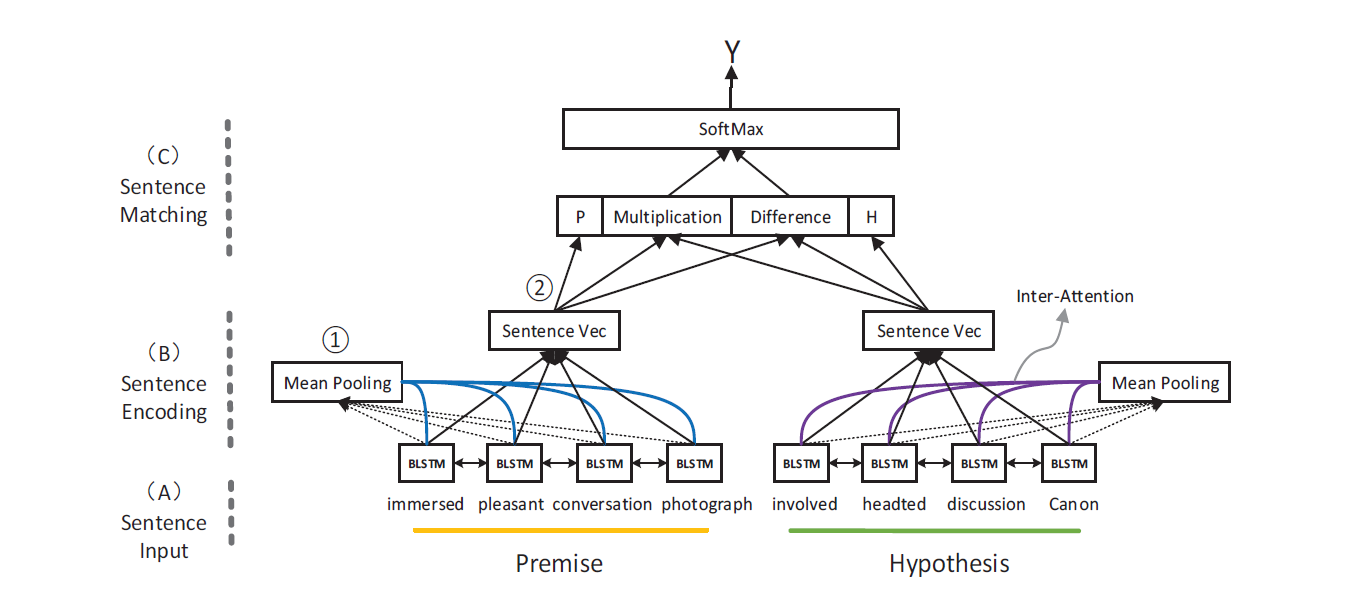
Finally we felt we can put in to action things we learned in the course such as Bi directional LSTM that we had the chance to experience with on the course assignment and here we could do on some different data with less guidance.

# What was the result reported in the paper?

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| Accuracy | 84.5 | 84.2 |

# What method was used in the paper?

The paper used a method called “Inner-Attention” on top of a mean pooling of a Bi-LSTM. The inner-Attention supposed to emphasize the important words in each sentence and afterwards concatenate the results of the two plus the difference and their multiplication, When the Network is trained in a Siamese network form (same network compute the hypothesis and the premises sentences) – not mentioned if only the LSTM is shared or also . The architecture described above looks as follow:



The inner attention equation () is:

When:

1. is the output of the words form the Bi-LSTM
2. is the mean pooled vector from the Bi-LSTM output.
3. is the unite matrix[[1]](#footnote-1)
4. (all parameters of the model)

We weren’t sure if the inner attention was shared or not (maybe it is not necessary here, mentioned in the next question)

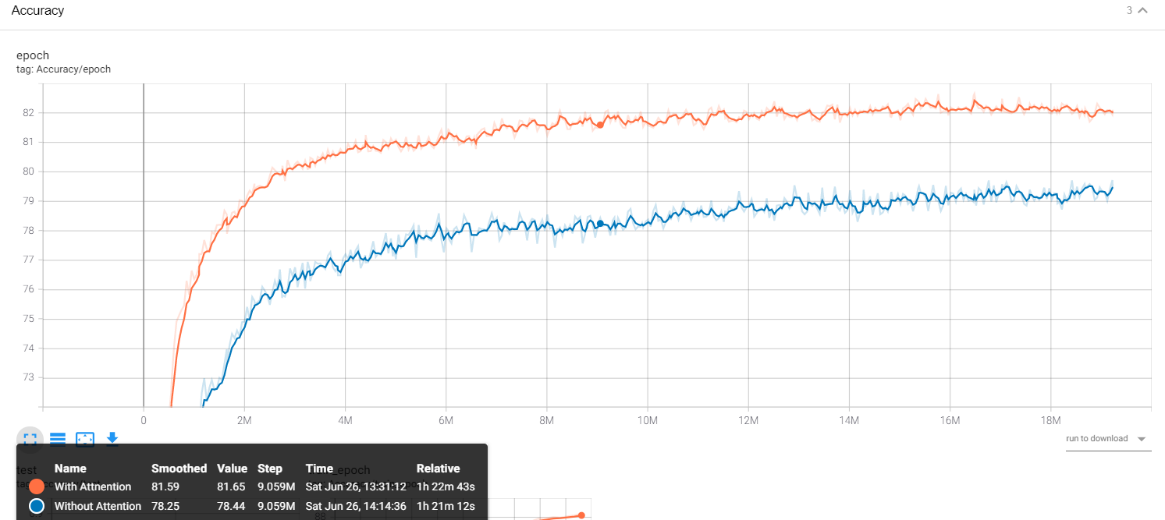
# Did your code manage to replicate this result?

NO but least hope I can rewrite this shit

# What was your performance on that dataset (how does your report compare to theirs)?

# What was involved in replicating the result?

First, in our vanilla implantation there were some unclear points.

We tried to run the net with and without the attention and we saw a significant increase in both the dev accuracy and in the test accuracy, so we conclude that our implementation of attention was indeed correct. The accuracy graph for test set on two runs with the same exact parameters when one is with the inner-attention and the other is without:

The attention added 2.71 points to the test score (81.54 vs 78.83)

We also tried to check whether the attention supposed to be shared across the premise and hypothesis sentences, but we didn’t notice any significant difference.

Play with the dropout – they did mention the p of the dropout but it isn't clear if they used it only in the end of the net or maybe in more places…?

Number of layers in the bi-lstm

Different glove vectors (although they did mention the exact embedding matrix)

Play with the optimizer

Used mini batch RMProp – but didn’t share all of its parameters, we used also AdamW with different learning rates.

Also, the paper mention that the last layer is a non-linear projection to the number of classes. The authors didn’t give the exact details about this function and in the beginning we tried to use a linear layer with Relu() activation function (for the non-linearity). But later on, after some thoughts, came in to conclusion than Relu() which is linear for every value > 0 is less appropriate for our purpose and we replaced it with Tanh()

What worked straightforward out of the box? what didn't work?

Are there any improvements to the algorithm you can think about?

1. As far as we understood, we tried to reach out to the writer but they didn’t answer us and we consulted some experts (and literature) who advise us so. [↑](#footnote-ref-1)