LIGN 167
Project Proposal
Sepehr Ardebilianfard
Adam Lareza

Question Answering Model for Squad

Hypothesis:

We will create a language question and answering model using Squad 2.0 dataset which consists of 100,000 questions and 50,000 unanswerable questions. We will be implementing a BiDAF (Bidirectional Attention Flow) model as our baseline. Using *Bi-Directional Attention & Self Attention for SQUAD*[1] as our inspiration, We will experiment with self-attention to improve our performance and compare our results to SQUAD dataset leader board.

Approach:

Based on the [figure1], model include a Embedding layer where questions and context are represented using Glo Ve and encoded using GRU. Once input is encoded, it is passed on to bidirectional attention layer before getting concatenated and passed on to a double LSTM layer. The bidirectional layer computes similarity matrix for context and questions hidden state. Given similarity vector a row-wise softmax is computed as Context to Question attention and a column-wise softmax will provide us with Question to Context attention hence the bidirectional nature of the model.

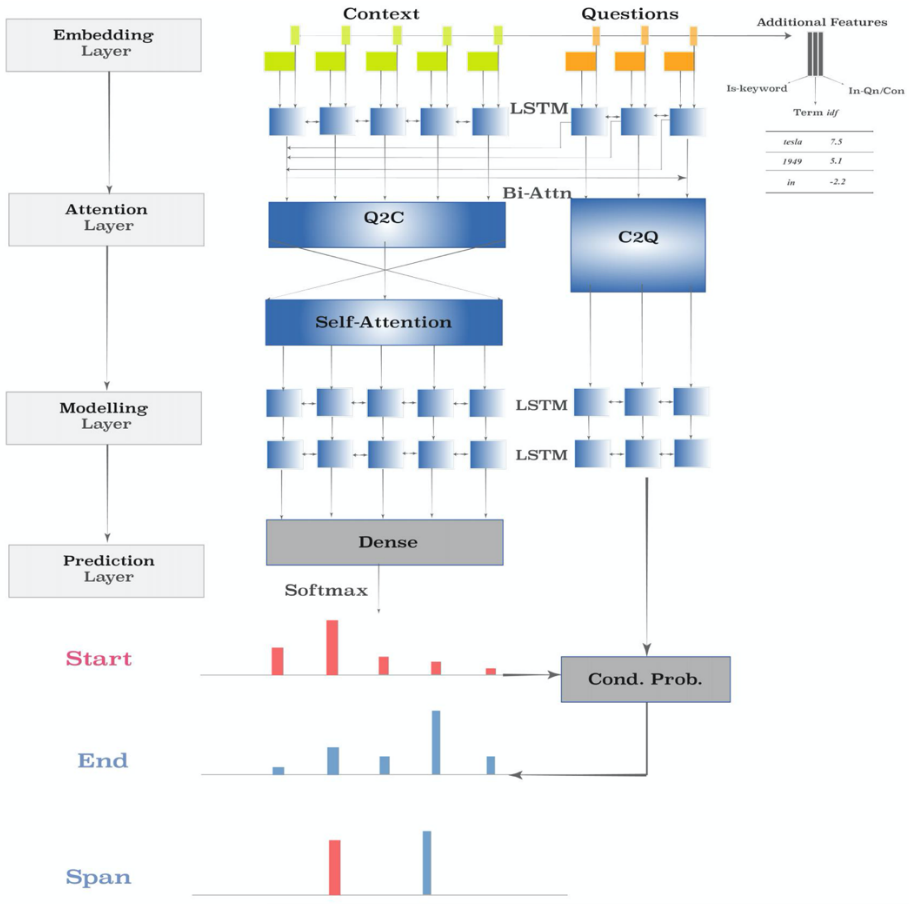

Figure 1

Product of both bidirectional matrices will be passed on to a double LSTM where a LSTM is used for forward direction and other for reverse direction to encode the sequence. Data is passed on the output layer where softmax is calculated using output of reverse LSTM to produce the probability vector of where index start of the answer should be. Similarly the output of forward LSTM is used to predict the end index probability for the answer. Loss will be calculated using [Figure2] and evaluated for its performance by using metrics in [Figure 3].

$$\text{loss} = -\log \boldsymbol{p}_{\text{start}}(i) - \log \boldsymbol{p}_{\text{end}}(j)$$

Figure 2

Metrics:

$$EM = N_{exact\_match}/N_{total\_predictions}$$

$$precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

Figure 3

Experimentation:
From *Bi-Directional Attention & Self Attention for SQUAD* [1] we will be experimenting with self attention [Figure3] instead of bi-directional attention and a combination of both where encoded data is given to both attention modules and their results are concatenated before it is given to prediction layer. After ensuring baseline model meets performance metrics of current paper, we will begin to experiment with combining attention modules to improve performance compared to SQUAD leader board.

Background:

- Bi-Directional Attention & Self Attention for SQUAD
    - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6909173.pdf
- Bi-Directional Attention Flow for Machine Comprehension
    - https://arxiv.org/pdf/1611.01603.pdf
- Gated-Attention Readers for Text Comprehension
    - https://arxiv.org/pdf/1606.01549.pdf
-

Data Sets:

- SQuAD 2.0
    - https://rajpurkar.github.io/SQuAD-explorer/