Proposal: Ensembling with BERT and its variations for Question Answering

Team members: (Group 5) Jakub Wasylkowski and Akash Govindarajula and (Group 14) Unnati Parekh and Nimesh Tripathi

Question Answering models serve as a vital component of many modern virtual assistants, chatbots, and automated customer service. Recent advances in this field have pushed their effectiveness to near-human levels of error. One of the more successful models used in this task is Bidirectional Encoder Representations from Transformers (BERT). We propose an experimental overview of combining BERT with other popular models using different ensembling techniques to create more powerful word embedding models capable of closing the human error gap for the task of Question Answering.

The datasets used for Question Answering experiments frame the task as reading comprehension where finding the answer to a question about a given paragraph/ document involves finding a span in the paragraph/ document. Thus, the aim for our models is to highlight this “span” of a text containing the answer–this is represented as simply predicting which token marks the start of the answer, and which token marks the end.

Dataset

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, which consists of passages from Wikipedia and questions posed by crowd workers, the answer to which is a text span from the particular context passage. For this project, we will be using SQuAD 2.0, which consists of 150,000 short contexts with question-answer pairs and consists of over 50,000 unanswerable questions.

Model description

Reading Comprehension is a very interesting topic of research in the field of Natural Language Processing. Most of the proposed high-performing models use neural attention mechanisms to combine the representations for the context and the question. Drawing insights from the previous work, we aim to leverage the performances of the BiDAF (Bi-Directional Attention Flow) model and variations of BERT models on the task of Question Answering by using different ensembling techniques.

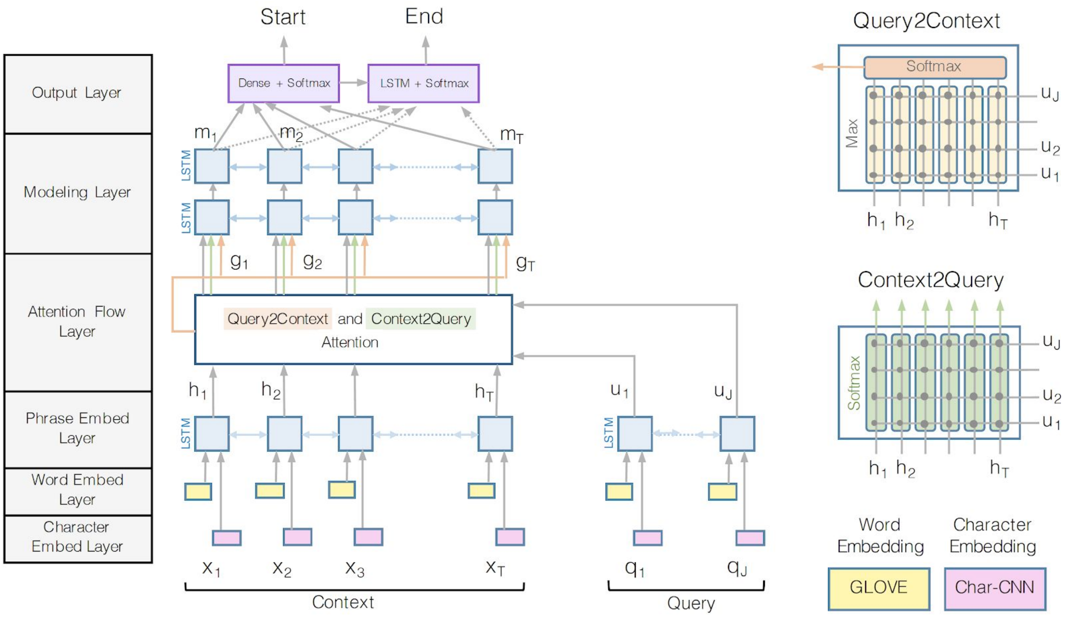

Figure 1: A graphical representation of the BiDAF multi-stage architecture from Seo et., 2017

The BiDAF network is a multi-level architecture that includes character-level, word-level, and contextual embeddings and uses bi-directional attention to obtain query-aware context representations. We plan to consider BiDAF model scores as well as a BiDAF model with BERT's last layer output as contextual word embeddings.

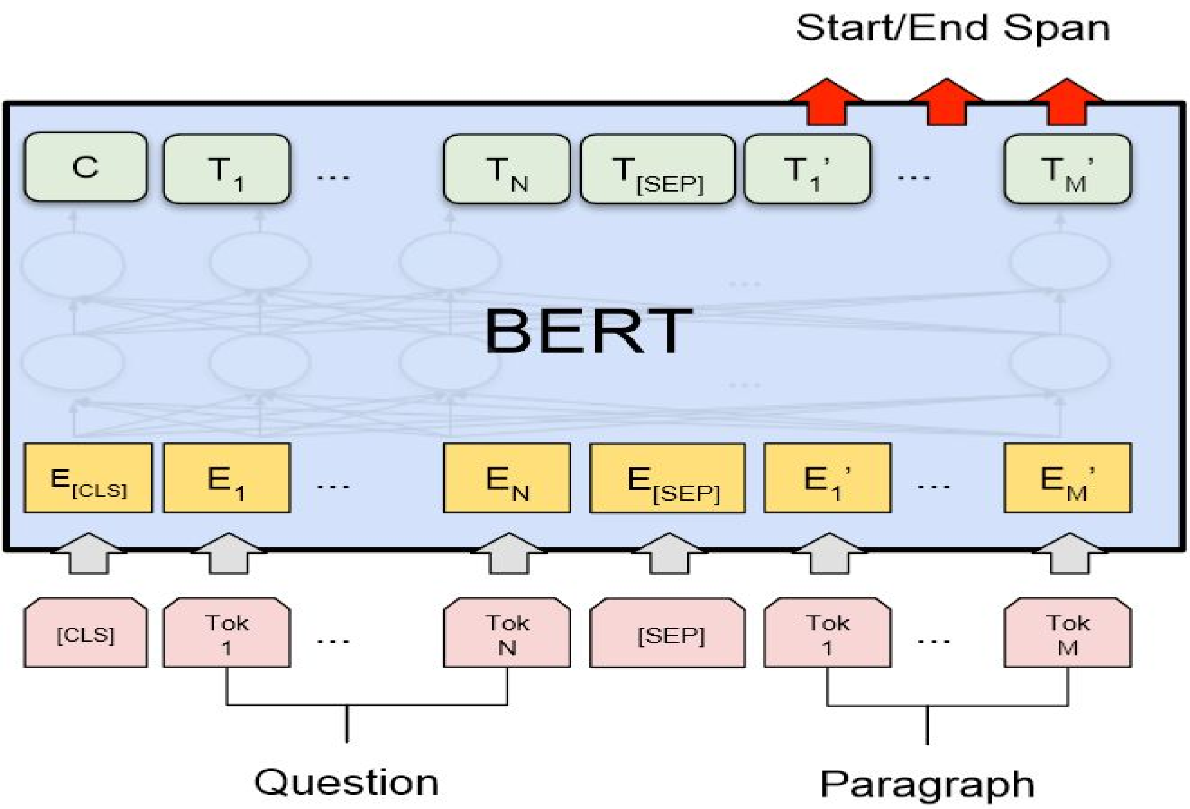

Figure 2: A graphical representation of the BERT architecture for Question Answering task

BERT architecture consists of a multi-layer Transformer encoder with a bidirectional self-attention mechanism. The encoder is pre-trained by jointly conditioning on both left and right context in all layers and this provides significant performance for downstream tasks like question answering. For our experiments, we will be fine-tuning the BERT-Base model on our data set.

We also aim to use the predictions from the above models and their variations and use different ensemble techniques to combine the predictions and analyze the effect of such methods.

Evaluation metrics

The evaluation metrics we will be using for this task are the Exact Match (EM) and F1 score. EM is a binary measure of whether the system output exactly matches the ground truth answer. F1 score is a harmonic mean of precision and recall.

As a baseline for our experiments, we will be considering the BiDAF model scores and the human performance scores.