Evaluating emrQA through Google's open source pre-trained language model, BERT

Dixit Patel

1 Introduction

The goal of this independent study was to research how Google's newly released pre-trained model can be used, through fine-tuning, for evaluating emrQA [1]. emrQA is a question answer corpus of Electronic Media Records, generated automatically from a QA generation framework.

Through this framework, we created QA corpus for 5 different domains obtained from i2b2 namely, heart disease risk, medications, obesity, relations, smoking. Although the QA corpus were generated, only the heart disease risk QA had the actual start and end positions for questions posed in each paragraph context. We identify that only this corpus can be used to understand performance of BERT. Essentially, we started by using the BERT-Base Cased pre-trained model to evaluate the quality of the QA pairs generated.

BERT-Base has 12 Transformers, 768 Hidden layers and 12 Attention heads. This model can be run locally due to it's lower memory requirements. Google BERT provided a high f1 measure on the SQUAD task. The EMR QA pairs generated are highly domain specific. Upon analysing the results, we note that this task is significantly different than the SQUAD task in which BERT achieved higher than human performance. We tried running the BERT model using the run_squad.py script provided in the github repo of BERT, on the emrQA dataset. There were several problems we faced including the answer type present in the emrQA not matching the SQUAD. We were able to get the data points into TFRecord format required to run on TPU, but the script requires several more modifications. The original script already does a lot of complex pre-processing on the SQUAD dataset before training on TPU's. We will work towards completing this in our future independent study.

2 References

- emrQA: A Large Corpus for Question Answering on Electronic Medical Records (https://arxiv.org/pdf/1809.00732.pdf)

- Google BERT

  (https://arxiv.org/abs/1810.04805)