Problem

Problem Statement: Common pre-trained word embeddings do not capture key information on contextual relationships which are essential for encapsulating accurate word and phrase meaning. This leads some NLP models to have a difficult time picking up on contextual nuances, especially for long sequences or sequences with uncommon words. I aim to transform the GLUE input embeddings into more context-aware and positionally attuned inputs, as well as add a trained character-level embedding layer to enhance conditioning on the internal structure of words, ultimately enhancing the BERT/GQA system.

Goal: In this project, I implemented a learned positional encoding layer, a pre-model scaled DP self-attention scheme, as well as a character-level embedding layer for both the queries and context to make up for the lack of context awareness in the default model.

Existing Approaches & Q&A approaches that inspired my project: ERTN applies the transformer model architecture to GQA; GQAlet makes use of stacked encoder blocks with convolution, attention, and feedforward layers and a special context-query attention layer.

Data/Task

My project involves improving upon the BERT/GQA model using the techniques described above: I will train on the SQuAD training set of over 130,000 examples of simple question-context pairs like the one below, and test on a condensed version of the official dev set. The official test will be conducted on the complete SQuAD test set, which is hidden.

Evaluation metric: I will use two metrics to evaluate my model’s performance: the Exact Match (EM) score, which is a binary measure of whether the output matches the ground-truth answer exactly, and the F1 score, a less strict measure that is the harmonic mean of precision and recall.

Analysis

From the F1 and EM graphs, it is clear that the final learned positional encoding and character level embedding implementation performs better than the baseline and all other implementations at all levels of difficulty.

Conclusions

We see that by adding implementations that emphasize contextual and positional awareness to the context/query embeddings, as well as incorporating key information on internal word structures through character-level embeddings, we are able to significantly improve upon the baseline BERT model. I also found that too much attention can be detrimental and confuse our model, as the SDP attention layer after the positional encoding actually worsened overall performance.

I didn’t have time to get to it, but future work may involve stopping training earlier, as many of my experiments appeared to have peaked within epoch 20-25 range. This would allow me to see if preventing overfitting makes slight differences in overall performance. Furthermore, I also would like to delve deeper into the cell behavior of the positional encoding/self attention layer and why its progression during training and overall worse performance was such an outlier from the rest of the experiments.