Introduction and Summary of Works

In this project, we seek to investigate the general trend of different model performances on question answering with the addition of relevant contexts. At the heart of the problem, we wish to explore the potential for self-attention and retrieval-based question answering tasks. It is more beneficial to allow models to gain more relevant information from an expanded context, or is it important to prune contexts to only get what is necessary?

In this project, we utilize a Graph RNN to find retrieve related contexts for question context pairs and add new context onto the existing data. We train 3 separate models in the regular query and data with the additional data points: 1) Bi-directional Attention Flow (BDAF) which is a LSTM model that utilizes a bidirectional attention flow layer to allow for attention to flow from the context to the question and vice versa, and two Transformer models: the Refformer and the Bidirectional Encoder Representations from Transformers (BERT).

Background:

For our baseline dataset, we utilized the Stanford Question Answering Dataset introduced by Minsh, G. et al. contains contexts from Wikipedia questions for each of these contexts and corresponding answers. Utilizing a graph-based recurrent retriever as proposed by Ao et al. we expand the context of each question answer pair. Due to computation restraints we take a subset of the total data, 7,200 train and 1,450 validation.

Bidirectional Attention Flow (BDAF):

Bidirectional Attention Flow models are an extension of LSTM models that have the following architecture:

$$ \mathbf{S} \in \mathbb{R}^{N \times M} $$

$$ \mathbf{S}_{ij} = \mathbf{w}_l \mathbf{c}_i \mathbf{q}_j \mathbf{r} + \mathbf{q}_j \mathbf{r} \in \mathbb{R} $$

Context to Question attention is derived from the row-wise softmax of similarity matrix and Question to Context attention is derived from the column-wise softmax of the attention.

GRNN Retrieval:

To augment the data we used a GRNN retrieval system to append additional contextual information. The procedure was pioneered in the paper “Learning to Retrieve Reasoning Paths” by Wikipedia Graph. For Question Answering, and was originally used for open domain QA. The model utilizes multi-modal bidirectional self-attention, it allows for relating many different parts of a sequence to attend to a single element in the sequence.

Evaluation Metrics:

We used two quantitative metrics to evaluate the models, exact match and F1 score. Below is the equation for F1 score:

$$ F1 = \frac{TP}{TP + 0.5(FP + FN)} $$

Consider a span of words labeled as the ground truth and a span of words labeled as the predicted answer. Then TP = true positive is the number of words that are correctly identified in the span, FP = false positive is the number of words in the predicted span that are not in ground truth, and FN = false negative is the number of words in the ground truth that do not appear in the predicted span.

Additionally we keep track of exact match for each sample, where either a score of 1 if it’s both spans are the same or 0 otherwise.

Experiments

- **Task:**
  - Metrics: Content and question pair
  - Outlines: Classification for answer start and end indices in context
  - Evaluation: EM and F1 scores

- **Goal:**
  - Core: produce model through evaluating conditional sample outputs and comparing conditional evaluation metrics

- **Baseline model:**
  - BDAF first tuned on regular data, then fine-tuned on expanded data
  - BERT fine-tuned on expanded data
  - BERT fine-tuned on expanded data

- **Bi-attention model:**
  - BDAF first tuned on regular data, then fine-tuned on expanded data
  - BERT fine-tuned on expanded data
  - BERT fine-tuned on expanded data

- **Refformer:**
  - BDAF first tuned on regular data, then fine-tuned on expanded data
  - BERT fine-tuned on expanded data

- **Experimental Results:**
  - BDAF trained on expanded data
  - BERT fine-tuned on expanded data
  - Bi-attention trained on expanded data

Sample:

**Control model:**

- **Sequence:**
  - Input: Question
  - Output: Answer

- **Relation:**
  - Input: Question
  - Output: Answer

**Experimental model:**

- **Sequence:**
  - Input: Question
  - Output: Answer

- **Relation:**
  - Input: Question
  - Output: Answer

Conclusions

We see that generally across the board, the control models outperform the experimental models across for all the different. Although there are many factors that we are not accounting for, we have built a general intuition that the passage context to the most relative text is important for question answering tasks.

- **Takeways:**
  - We observe on the sample on the left, the added context makes sense, and relates to the correct answer multiple times in the context, the model was not able to pick up on the added data.
  - We see that this general trend does not change for all models, no matter the attention mechanism.

Future Exploration

- **Baseline model:**
  - Added context allows for different plausible solutions, etc.
  - Plans to expand expanded data generation

- **Bi-attention model:**
  - Due to computational restrictions on the data generation, we could only pre-process a small subset of the data.
  - We can see in our validation data that the models perform relatively quickly. In future work, we would like to see if the differences in the packed and unpacked remain obvious in longer training.

- **Refformer model:**
  - We utilized was trained on a very small corpus, it was only trained on War and Peace. We are interested in seeing how a larger Refformer model would perform.