Question Answering Classification With Expanded Datasets

Eric Feng, Rain Juhl

Department of Computer Science

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 importance of choosing contexts when collecting data for a question answering task. Is it is more beneficial to allow models to pick up larger amounts of potentially 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 on the regular squad data and data with the additional data points: a **Bidirectional Attention Flow (BiDAF)** 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 vise versa, and two **Transformer** models: the **Reformer** and the Bidirectional Encoder Representations from Transformers (**BERT**)

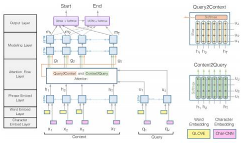
Background:

Dataset:

For our baseline dataset, we utilized the Stanford Question Answering Dataset introduced by Minaee, S. et al contains contexts from Wikipedia, questions for each of these contexts and corresponding answers. Utilizing a graph based recurrent retriever as proposed by Asai et al, we expand the context of each question answer pair. Due to computation restraints we take a subset of the total data, 7200 train and 1400 validation.

Bidirectional Attention Flow (BiDAF):

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



The core component of this model is the bidirectional attention flow layer. During the training, the model utilizes the hidden states of contexts and questions and calculates and learns the similarity matrix:

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

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T [\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

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.

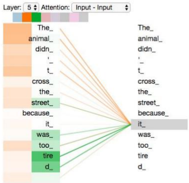
GRNN Retrieval:

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 Over Wikipedia Grah For Question Answering and was originally used for open domain QA. The GRNN uses a graph to model Wikipedia's link structure and then find high probability reasoning paths.

Technical Methods

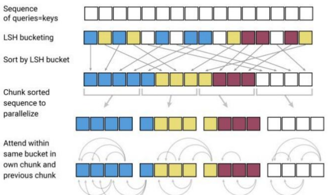
Bidirectional Encoder Representations from Transformers:

BERT is a transformer model proposed by Jacob Devlin et al. This model utilizes multi headed bidirectional self attention. It allows for relating many different parts of a sequence to attend to a single element in the sequence.



Reformer:

The Reformer is a transformer model that utilizes both Locality Sensitive Hashing (LSH) as well as reversible residual layer to make a more computationally efficient transformer.


Sequence of queries=keys
LSH bucketing
Sort by LSH bucket
Chunk sorted sequence to parallelize
Attend within same bucket in own chunk and previous chunk

The figure above highlights what LSH. A series of keys get hashed such that the buckets are likely to contain a similar amount of queries. The chunks are then sorted and attention is performed within the chunk. Attention calculation in this case is O(L log L)

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 or true positive is the number of words correctly identified in the span. FP or false positive is the number of words in the predicted span that are not in ground truth. Finally, FN or False Negative represents the number of words in the ground truth that do not appear in the the prediction span.

Additionally we kept track of exact match for each sample, where either a score of 1 is given if both spans are the same or 0 otherwise.

Experiments

Task:

- Inputs: Context and question pair
- Outputs: Classification for answer start and end indices in context
- Evaluation: EM and F1 scores
- Goal: contrast models through evaluating conditional sample outputs and comparing conditional evaluation metrics

Control runs:

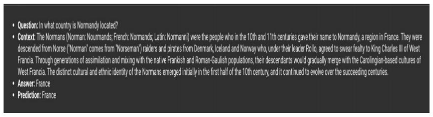
- BiDAF trained on regular squad data,
- DistilBERT fine tuned on regular data
- Reformer fine tuned on regular data

Experimental Runs:

- BiDAF trained on expanded data.
- DistilBERT fine tuned on expanded data
- Incorporates class labels through CBN in residual blocks

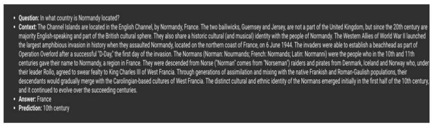
Sample:

Control model:



- **Question:** In what country is Normandy located?
- **Context:** The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
- **Answer:** France
- **Prediction:** France

Experimental model:

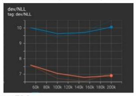


- **Question:** In what country is Normandy located?
- **Context:** The Channel Islands are located in the English Channel, by Normandy, France. The two bailiwicks, Guernsey and Jersey, are not a part of the United Kingdom, but since the 20th century are majority English-speaking and part of the British cultural sphere. They also share a historic cultural (and musical) identity with the people of Normandy. The Western Allies of World War II launched the largest amphibious invasion in history when they assaulted Normandy, located on the northern coast of France, on 6 June 1944. The invaders were able to establish a beachhead as part of Operation Overlord after a successful "D-Day," the first day of the invasion. The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
- **Answer:** France
- **Prediction:** 10th century

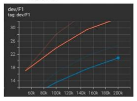
Evaluation

Baseline BiDAF Experiments

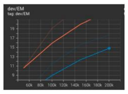
Validation Negative Log Likelihood



Dev F1



Dev EM

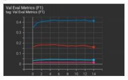


🟧 Control Data

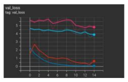
🟦 Experiment Data

Transformer Experiments :

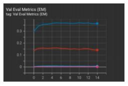
Validation F1



Validation Loss



Validation EM



🟦 BERT with Regular Data

🟥 BERT with Augmented Data

🟦 Reformer with Regular Data

🟪 Reformer with Augmented Data

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 see can build a general intuition that, the pruning of contexts to the most relative text is important for question answering tasks.

Takeaways:

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

Future Exploration

- Added context allows for different plausible solutions, IE: France, northern coast of france. Would be interested in Semi Supervised learning in order to label plausible new answers presented in new context.
- Due to computational restrictions on the data generation, we could only preprocess a small subset of the data. We can see in our validation data that the models overfit relatively quickly. In future work, we would like to see if the differences in the padded and unpadded remain obvious in longer training.
- The Reformer 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 reformer model would perform.