BiDAF with Self-Attention on SQuAD 2.0

Zachary Chen

Department of Computer Science

Introduction

- SQuAD 2.0 [1] is a question-answering (QA) dataset that consists of question and paragraph pairs of which half the questions can be answered using a chunk of text from the corresponding paragraph.
- Evaluating on SQuAD 2.0, I improved upon a simplified BiDAF model [2] baseline by implementing word-character concatenated embeddings as well as integrating self-attention into the BiDAF attention flow layer.

Example

Question: What is the name of Harvard's primary recreational sports facility?

Shortened Context: Harvard has several athletic facilities, such as the Lavietes Pavilion, a multi-purpose arena and home to the Harvard basketball teams. The Malkin Athletic Center, known as the "MAC", serves both as the university's primary recreation facility and as a satellite location for several varsity sports.

Answer: Malkin Athletic Center

Methods

- **Character Embeddings:** I concatenated the baseline word embeddings with character embeddings that had been fed through a CNN and ReLU function.
- **Context2Context Self Attention:** I modified the baseline Attention Flow Layer by implementing self attention using the context embeddings, then concatenating the resulting attention output to the baseline output of the Attention Flow Layer.
- **Finetuning:** I experimented with a few different hyperparameters during the training process, such as dropout probability and kernel size of the character embedding CNN.

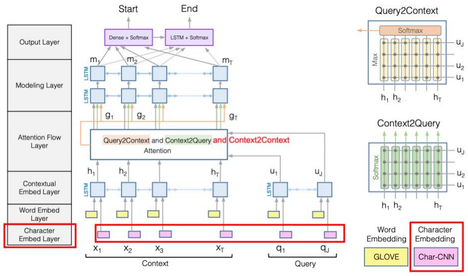
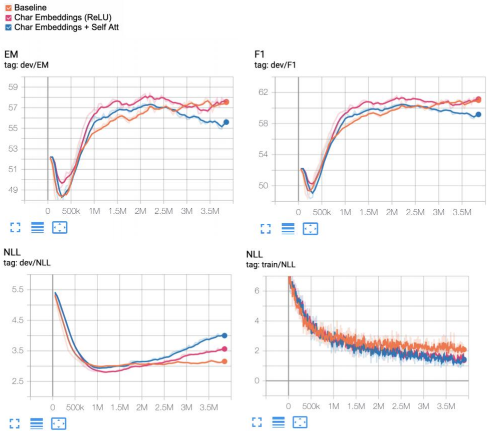


Figure 1. BiDAF Model [2] with my changes in red

Experiments/Results

Evaluating on SQuAD 2.0 using an Exact Match (EM) and F1-score metric, I found considerable improvement over the baseline by implementing the character embeddings. However, my integrated self attention seemed to perform at around the level of the baseline, perhaps indicating some small error in implementation.



Model	Dev EM	Dev F1
Baseline	57.755	61.186
Char Embed	57.822	61.428
Char Embed w/ ReLU	**58.343**	**61.560**
Char Embed w/ ReLU, drop=0.15	58.041	61.675
Char Embed w/ ReLU, Self Att	57.402	60.614

Table 1. Model Performances On Validation Set

Analysis

Question: What can the exhaust steam not fully do when the exhaust event is insufficiently long?

Shortened Context: However, as one and the same valve usually controls both steam flows, a short cutoff at admission adversely affects the exhaust and compression periods which should ideally always be kept fairly constant; if the exhaust event is too brief, the totality of the exhaust steam cannot evacuate the cylinder, choking it and giving excessive compression ("kick back").

Answer: evacuate the cylinder

Char Embed w/ ReLU Prediction: N/A

While the best model found is relatively successful on SQuAD 2.0, there are still some cases that fail. In the example above, where "too brief" and "insufficiently long" serve as synonyms, and words like "fully" don't appear in the context, our best model can't find the answer.

Conclusions

In this project, we explored character embeddings, self attention, and finetuning on a BiDAF baseline, and were able to achieve improvements over the baseline on SQuAD 2.0. I hope to further explore different methods of self-attention and aim for more improvement in the future.

References

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang.
Know what you don't know: Unanswerable questions for SQuAD.
In *Association for Computational Linguistics (ACL)*, 2018.

[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.
Bi-directional attention flow for machine comprehension.
In *International Conference on Learning Representations (ICLR)*, 2017.