Improving BiDAF Heuristically and with Self-Attention for SQuAD 2.0

Nathaniel Goenawan (nathgoh@stanford.edu)
Christopher Wong (cwong7@stanford.edu)
CS224N: Natural Language Processing with Deep Learning

Problem

Motivation:

- Question answering (QA) is an important end-user task for NLP and IR.
- QA systems propose complex architectures; do the aforementioned architecture justifies their empirical results?

Goal:

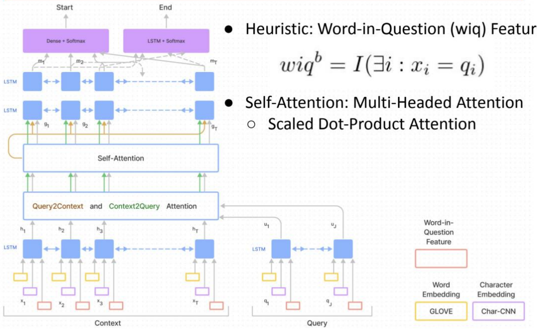
- Create a relatively simple QA system to answers questions correctly from the SQuAD 2.0 dataset.
- Extend a BiDAF baseline be employing an additional input feature (word-in-question) and self-attention.

Data

- **Question:** What president eliminated the Christian position in the curriculum?
- **Context:** Charles W. Eliot, president 1869–1909, eliminated the favored position of Christianity from the curriculum while opening it to student self-direction. While Eliot was the most crucial figure in the secularization of American higher education, he was motivated not by a desire to secularize education, but by Transcendentalist Unitarian convictions. Derived from William Ellery Channing and Ralph Waldo Emerson, these convictions were focused on the dignity and worth of human nature, the right and ability of each person to perceive truth, and the indwelling God in each person.
- **Answer:** Charles W. Eliot

- SQuAD 2.0: composed of question, context, answer triples
  - Contains answerable and unanswerable questions
  - Train: 129,941 example, Dev: 6078 examples, 5915 examples

Approach

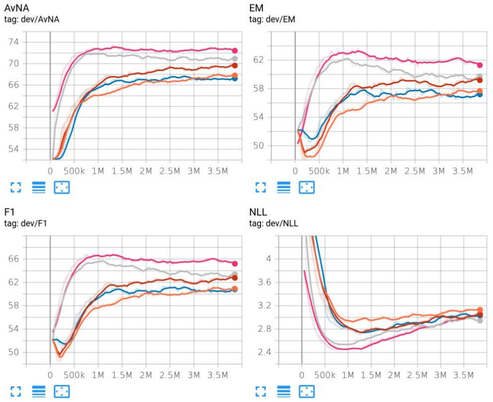


- Heuristic: Word-in-Question (wiq) Feature

$$wiq^b = I(\exists i : x_i = q_i)$$

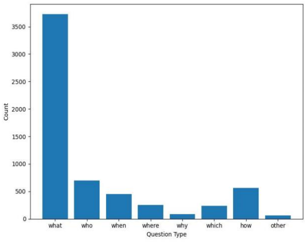
- Self-Attention: Multi-Headed Attention
  - Scaled Dot-Product Attention

Results + Analysis



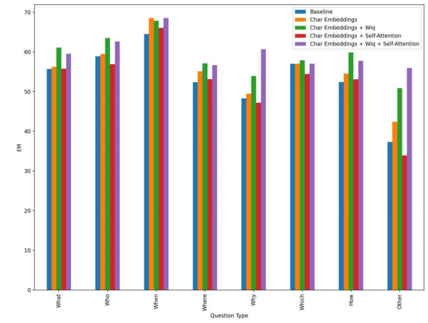
Model	Dev EM	Dev F1	Test EM	Test F1	Total Train Time
Baseline	58.175	61.440	-	-	3 hr 9 min
Baseline + Char Embeds	59.452	63.076	**59.104**	**62.673**	4 hr 4 min
Char Embeds + Wiq	**63.519**	**66.926**	56.585	59.464	10 hr 18 min
Self-Att + Char Embeds	58.14	61.37	-	-	8 hr 47 min
Self-Att + Char Embeds + Wiq	62.275	65.772	54.269	56.846	8 hr 48 min

- All proposed models performed better than baseline on dev set.
- On test set, neither word-in-question feature or self-attention outperformed character-level embedding.
- Self-attention and word-in-question takes over twice as long to train versus character-level embedding.



- Dev set heavily skewed towards 'what' question types
- Very small sample sizes for 'why' and 'other' question types.

Analysis Cont'd



- The sizeable improvements in 'why' and 'other' question types performance could be because of lack of training and dev data.
  - High variance, possibly misleading estimation of model's performance.
- Most models overall showed improvement over baseline on dev set.

Conclusion

- Character-level embeddings slightly improved on the performance of the baseline at the cost of a minimal increase in training time.
- Self-attention performed minimally worse than baseline.
  - Implementation potentially unsuited for QA systems.
- Holistically, baseline + char embeddings is the best model.
- Wiq feature has some potential merit given dev set performance.
  - Can explore further into, additional wiq feature types?

References

1. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension, 2018
2. Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Fastqa: A simple and efficient neural architecture for question answering. CoRR, abs/1703.04816, 2017.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.
4. Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Association for Computational Linguistics (ACL), 2018