Attention and CNN empowered BiDAF for SQuAD

Haojun Li

Department of Computer Science, Stanford University

Abstract

The SQuAD dataset introduced in [1] formulate the Question and Answer problem as a span prediction problem with adversarially written unanswerable questions. In this project I aim to

- Establish a better baseline by adding character-level embedding
- Creatively use Encoder Block inspired from [2]
- Experiment with a multi-headed BiDAF attention layer

I have shown that by combining CNN-based and RNN-based encoders, the model performs better than those encoders alone, and significantly better than the baseline model. On the other hand multi-headed BiDAF attention models perform worse due to overfitting and other fundamental model restrictions.

Data

The SQuAD dataset consists of entries of a context paragraph, a question, and an answer, where the answer can be found as an exact match to a span of words in the context. An example is shown below:

- 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**

CNN + RNN Encoder

I added a CNN Encoder Block inspired from [2], and concatenated the result from the RNN Encoder before feeding the whole thing into the BiDAF Attention Layer [3]. I use the same Encoder Block to encode both the context and query. Rest of the architecture is the same

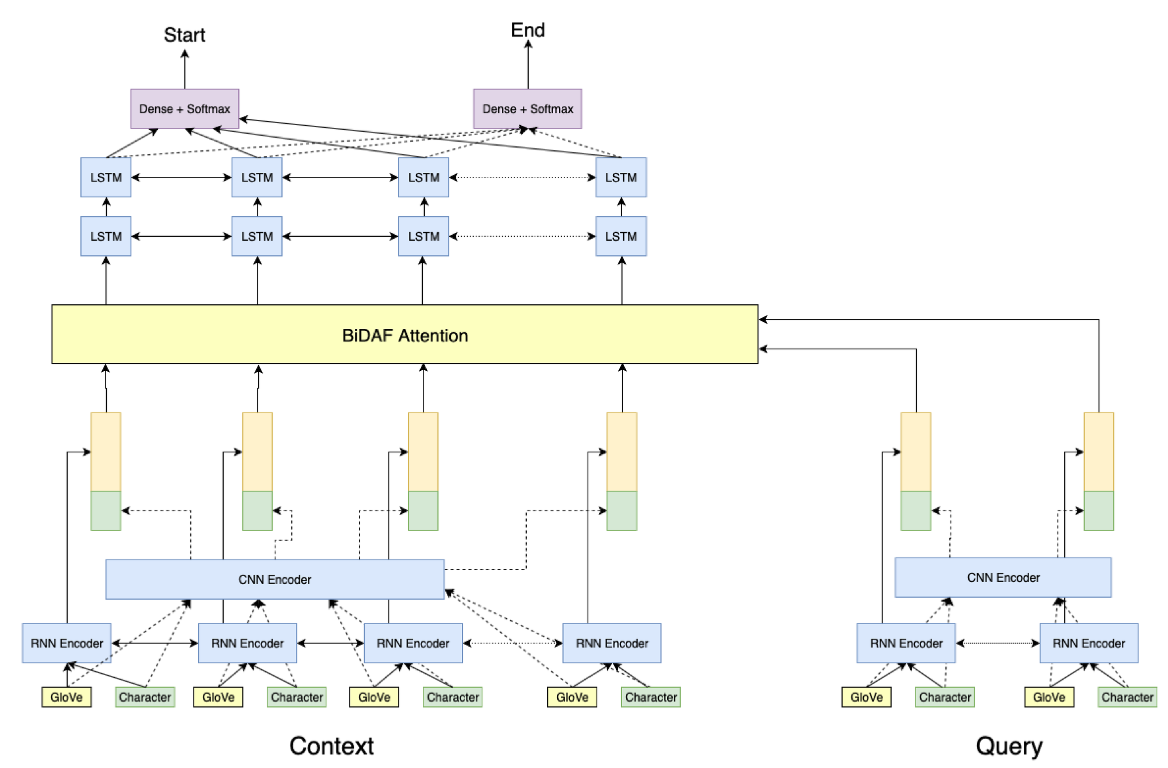

Figure 1: Attended CNN Encoder + RNN

Results & Analysis

EB is short for Encoder Block. We see that CNN + RNN model is better at predicting No Answer than baseline model, most likely benefit from more localized encoding from CNN. Still, it is unable to form long term dependencies, showing higher error rate for longer contexts. Multi-head BiDAF performs worse than basic character embedding as seen in table 1, due to overfitting and model restrictions (taking max gets rid of important negative and weak signals)

Model	Note	EM	F1
BiDAF Baseline		55	58
BiDAF-char	new baseline	61.18	64.2
Multi-BiDAF	feedforward	58.23	61.14
Multi-BiDAF	max pool	60.46	63.59
EB only		55.96	59.71
EB + RNN	fine-tune	64.779	67.87
EB + RNN	Test split results	**62.671**	**66.045**

Table 1: Results from experiments

Multi-headed BiDAF Attention

I added multiple attention heads to the BiDAF attention layer:

$$S_{ij}^{(k)} = w_{\text{sim}}^{(k)T}[c_i^{(k)}; q_j^{(k)}; c_i^{(k)} \circ q_j^{(k)}]$$

Then, I calculate multiple attention outputs

$$\bar{S}_{i,:}^{(k)} = \text{softmax}(S_{i,:}^{(k)}) \qquad a_i^{(k)} = \sum_{j=1}^{M} \bar{S}_{i,j}^{(k)} q_j^{(k)}$$

$$\bar{\bar{S}}_{:,j}^{(k)} = \text{softmax}(\bar{S}_{:,j}^{(k)}) \qquad b_i^{(k)} = \sum_{j=1}^{N} S_{i,j}^{\prime(k)} c_j^{(k)}$$

$$S^{\prime(k)} = \bar{S}^{(k)} \bar{\bar{S}}^{(k)T}$$

Combining them in 2 ways

1. Taking max: $c_{ij} = \max_k(c_{ij}^{(k)})$, $q_{ij} = \max_k(q_{ij}^{(k)})$, $a_{ij} = \max_k(a_{ij}^{(k)})$, $b_{ij} = \max_k(b_{ij}^{(k)})$.
2. Linear Feedforward: $a_i = W_a[a_i^{(1)}; a_i^{(2)}, ..., a_i^{(k)}] + bias_a$ and similarly for b_i, c_i, q_i.

Plots

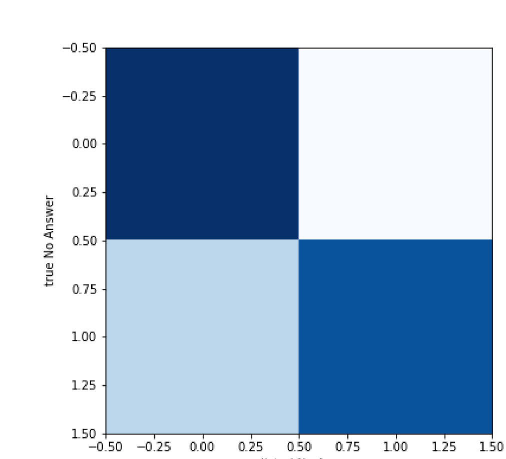

Figure 2: EB+RNN A/NA Confusion

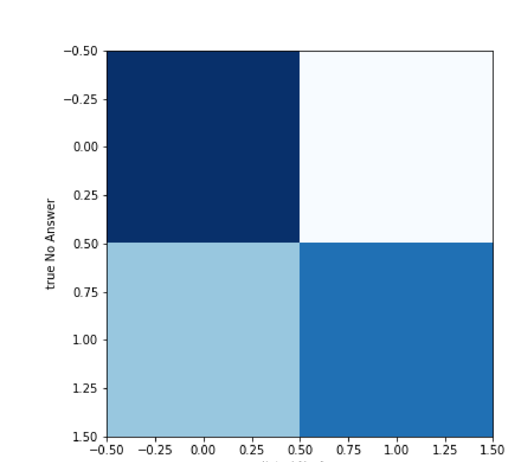

Figure 3: New Baseline A/NA Confusion

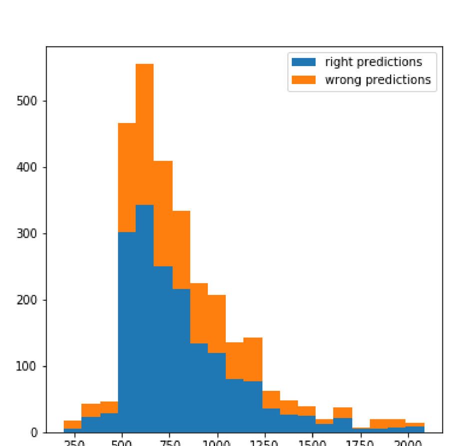

Figure 4: Histogram of Context Length

Conclusion

- Basic improvements such as character embedding and hyperparameter tuning significantly improves the baseline.
- Incorporating an Encoder Block to the Embedding Encoding layer inspired by QANet[2] is able to greatly improve the new baseline.
- Multi-headed BiDAF performs worse due to overfitting and fundamental model restrictions.

Future Works

A lot still need to be done if I have more time

- Use L2 regularization to solve overfitting
- Added self-attention after BiDAF as shown in [4]
- Use Transformer-XL as shown in [5] to improve long term dependency recognition

References

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018.

[2] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, abs/1804.09541, 2018.

[3] Ramon Tuason, Daniel Grazian, and Genki Kondo. Bidaf model for question answering. *Table III EVALUATION ON MRC MODELS (TEST SET). Search Zhidao All.*

[4] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198, 2017.

[5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019.