Stanford University

Improved BiDAF with Self-Attention

Mingchen Li, Gendong Zhang, Zixuan Zhou

Electrical Engineering

Introduction

Machine reading and question answering (Q&A) is essential for evaluating how well computer systems understand human languages.

We investigated an improved version of BiDAF model. We combined character embedding, self-attention and average-attention layers to a BiDAF model using GRU network to improve the accuracy of the baseline model. We experimentally prove that adding our learnable weighted average-attention layer is beneficial based on the significant improvement of model performance and negligible extra computational cost.

SQuAD 2.0

- More than 100000 question-answer pairs on more than 500 articles, and more than 50000 unanswerable questions.
- The original SQuAD dataset has three splits, train, dev and test, with the first two publicly accessible and the last one held privately.
- The original dev set was divided into two, one for dev and another for test.
- We also analyzed the amount of questions that start with different key words ("how", "what", "why", "which", "who", "where", "when"), as shown in Figure 1.
- "what" dominates both training set and dev set.

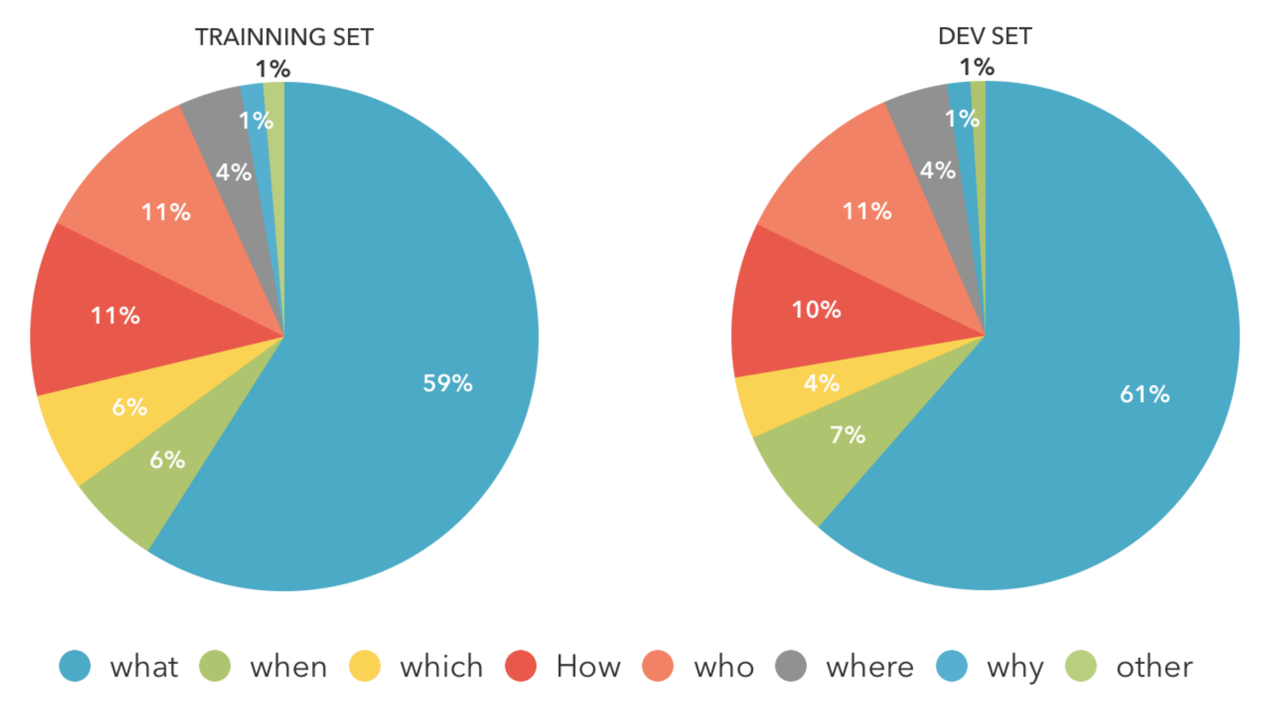

Figure 1: Dataset composition on different question types

Model

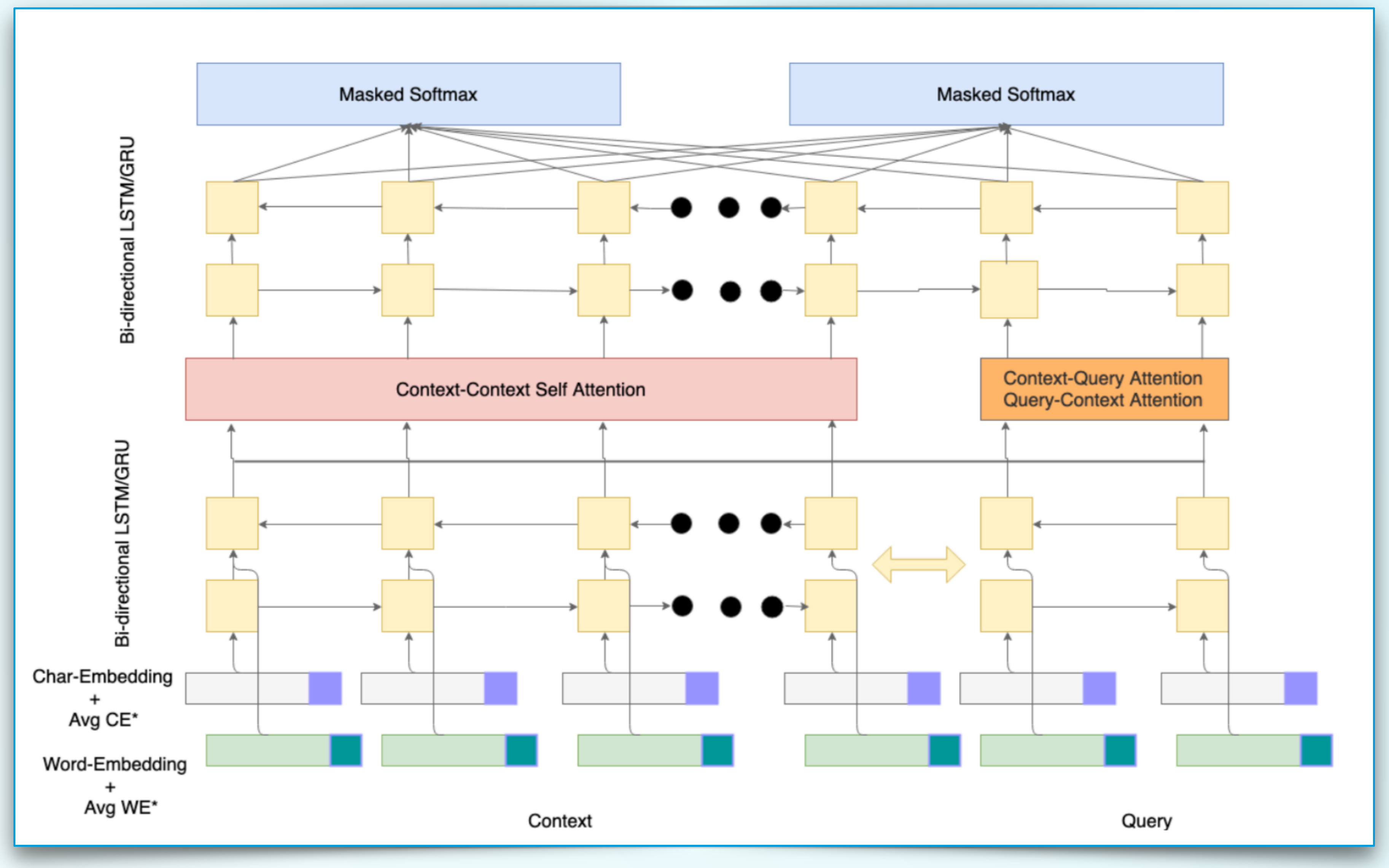

Figure 2: An overview of our model architecture. CE stands for char-embedding. WE stands for word-embedding.

Example

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.

Question:
In what country is Normandy located?

Answer:
France

Methods and Results

- Adding **character-level embedding** to the baseline, an increase in both F1 and EM score has been observed.
- Learnable **weighted average-attention** was then added to the embedding layer to further boost the performance.
- To decrease the training time, we leveraged **GRU** to accelerate the training process.
- After adding **self-attention layer**, the training time increased to 18 hours
- **Adam** were chosen to stabilize the training
- The performance of two-layer RNN model is better than the one-layer model during the first few iterations, and then becomes worse for the rest of the time. (overfitting)

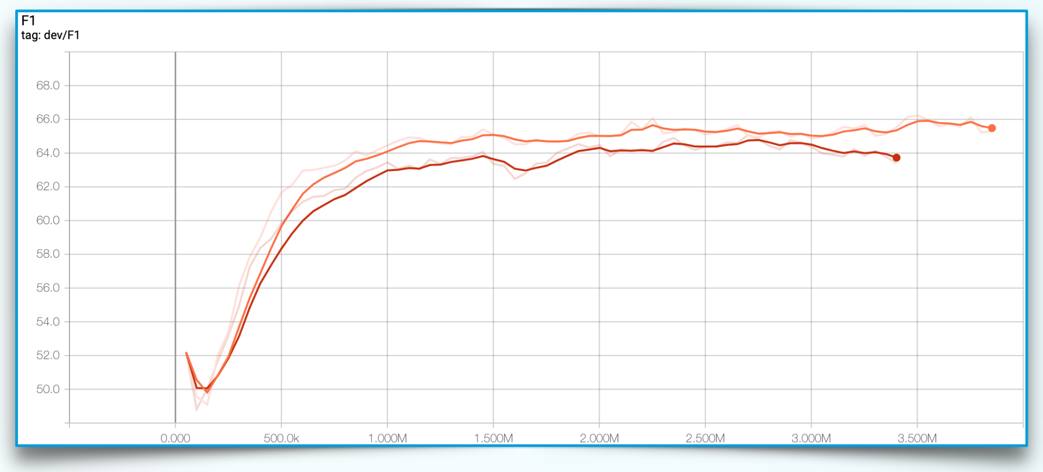

Figure 3: The experimental results of two models with one-layer and two-layer of embeddings.

◉ the one-layer model ◉ the two-layer model

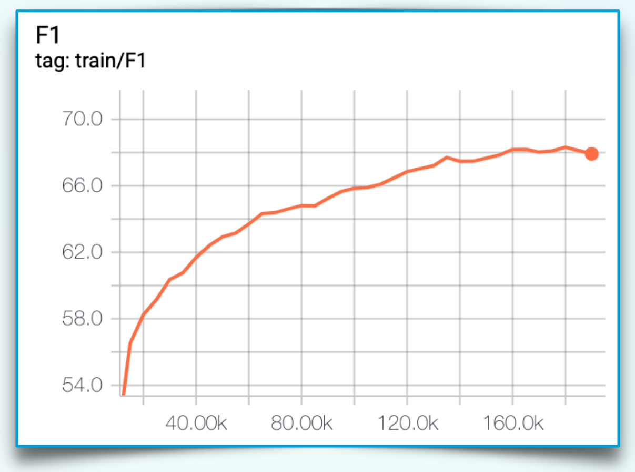

Figure 4: F1 score of QANet for comparison

Ablation Study

Model	F1	EM
Baseline	60.758	57.469
Baseline + Char Embedding	62.517	59.183
Baseline + Char Embedding + Learnable Weighted Average-attention	64.734	61.049
Baseline + Char Embedding+ Learnable Weighted Average-attention + Self-attention	66.241	62.679
Baseline + Char Embedding+ Learnable Weighted Average-attention + Self-attention+2-Layer RNN	63.098	60.025
QANets	68.013	64.251

Table 1: Model results at each implementation level

- The results of the BiDAF models at every improvement level is shown in Table 1. The last row of the table is the results of QANet.
- QANet serves as a comparison guideline, we did not submit QANet result to the learderboard, the result is from our own evaluation

Analysis

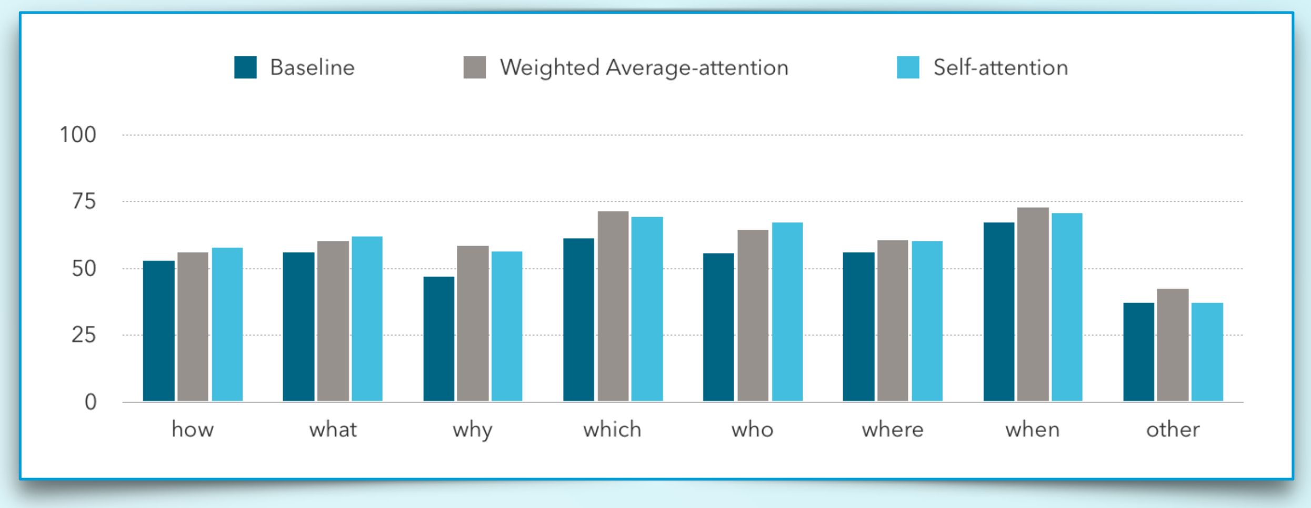

Figure: An break-down EM scores of different question types

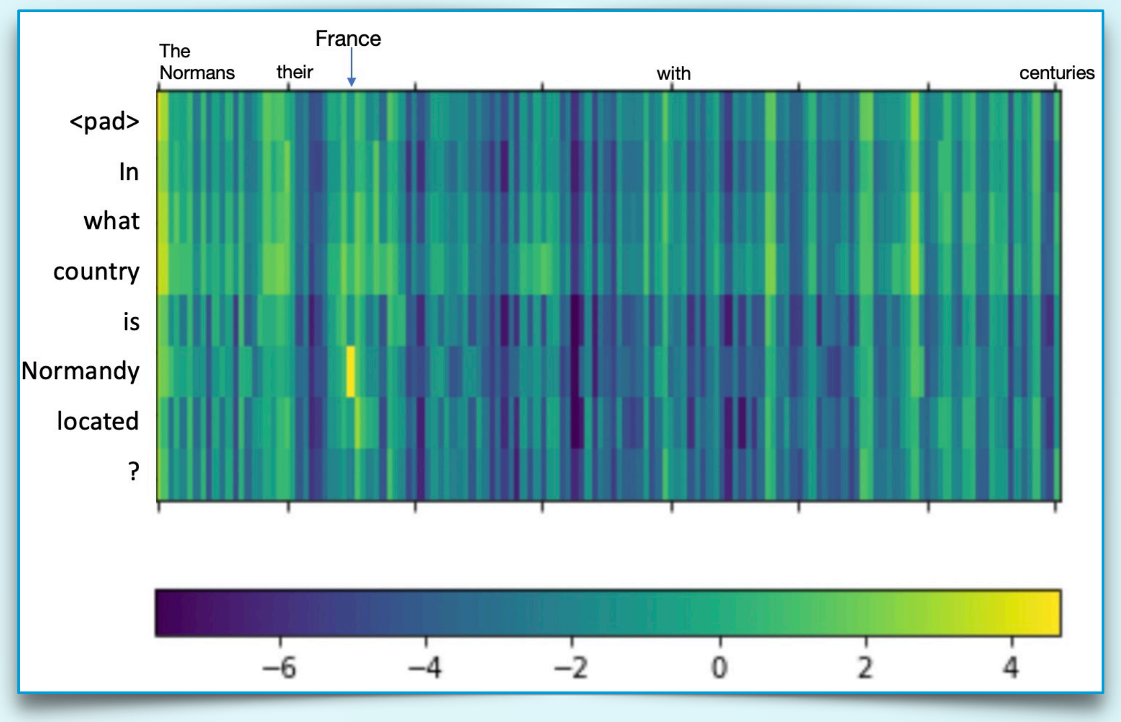

Figure: BiDAF attention visualization based on similarity matrix

Conclusion

- Adding character-embedding, weighted average-attention, and self-attention can boost the performance.
- The best F1 (66%) and EM (62%) scores were from our self-attention model.
- Adding extra layers might not help but make neural network more difficult to train.

Future Work

- Try different optimizer methods, such as AdaBound
- Add Elmo pre-trained embeddings
- Implement Transformer and Transformer-XL

Reference

[1] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," arXiv preprint arXiv:1611.01603, 2016
[2] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, andQuoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension.CoRR, abs/1804.09541, 2018.
[3] A. Vaswani & N. Shazeer & N. Parmar & J. Uszkoreit & L. Jones & A. N. Gomez & L. Kaiser & I. Polosukhin. 2017 Attention is all you need. In Neural Information Processing Systems (NIPS)