Question Answering Training on SQUAD(QATS)

Carson Yu Tian Zhao, Louise Qianying Huang, and Mi Jeremy Yu

Department of Statistics, Stanford University,
{czhao333, qyhuang, miyu1996}@stanford.edu

Problem

The tasks of machine comprehension (MC) and question answering (QA) have gained a significant amount of scholarly attention over the past few years. The goal is to teach machines to read, process and comprehend text and answer questions given a passage or a document.

Our system combines ideas from some of the best performing QA systems. On one hand, we improve upon the provided BiDAF baseline [1] by incorporating character level embedding and adding an extra layer of R-Net inspired multiplicative self-attention after the bidirectional attention layer. On the other hand, we re-implement the QANet architecture [2].

Data/Task

In this paper, we specifically focus on the Stanford Question Answering Dataset (SQuAD) 2.0 SQuAD 2.0 combines the 100,000 questions in SQuAD 1.1 with over 50,000 new, unanswerable questions. Such characteristics pose additional challenges to neural QA systems. Now, the system not only needs to answer questions accurately, but also needs to comprehend information sufficiency to determine whether a question is actually answerable. The main task for our model is to do reading comprehension and determine if the question has an answer or not based on the context. If it does, then the model tries to get the answer from the sub-phrases of the paragraph.

Result

	Dev Set F1 / EM	Test Set F1 / EM
Single Model		
BiDAF CS 224N Baseline	58.02 / 54.85	59.920 / 56.298
BiDAF + Char-CNN	63.14 / 59.65	
BiDAF + Char-CNN + self-attention	65.18 / 62.17	
QANet (1 attention head, D_{model} = 96)	**68.57 / 65.25**	
QANet (4 attention heads, D_{model} = 96)	68.10 / 64.75	
QANet (4 attention heads, D_{model} = 128	67.64 / 63.87	
Ensemble Model		
3 QANet ensemble	69.83 / 66.68	66.72 / 63.16
3 QANet + 2 best-performing BiDAF (5 ensemble)	**70.12 / 67.13**	**68.10 / 64.75**

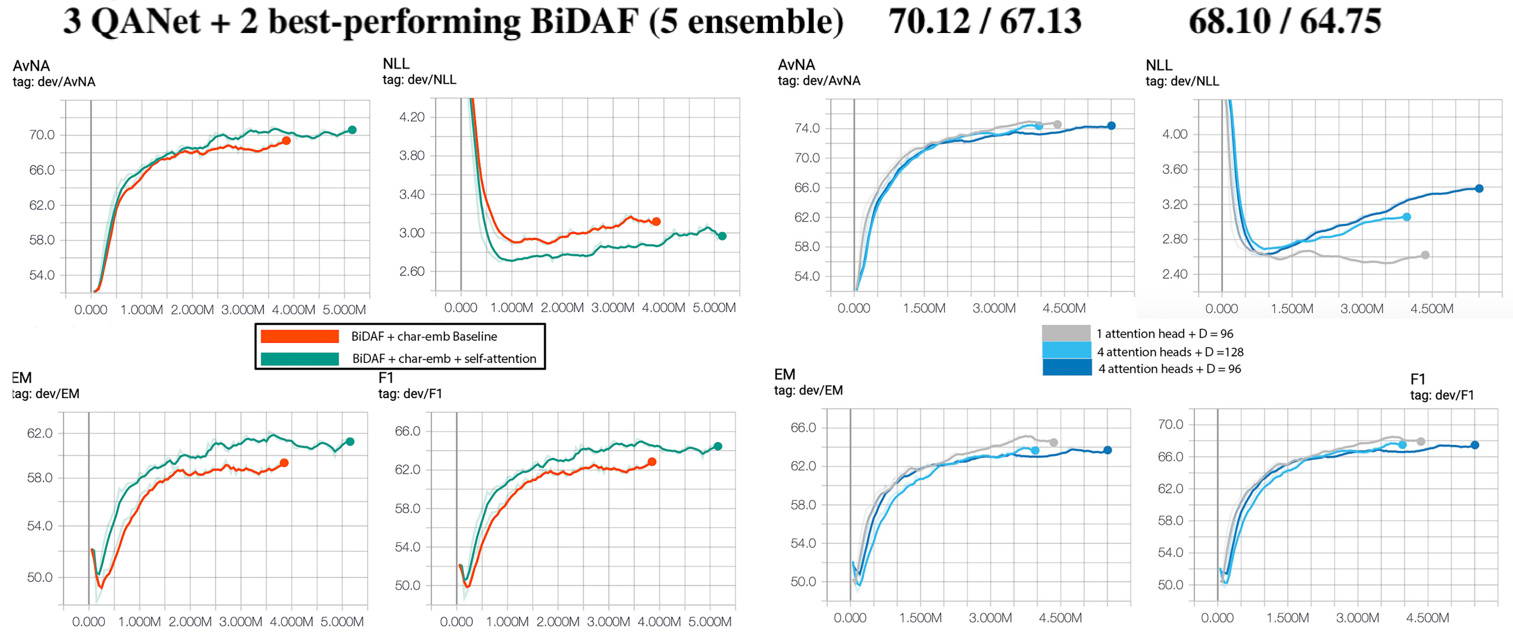

Figure 2, the Tensorboard result for BiDAF(left) and QANet(right)

Approach

The baseline of our work is the BiDAF model [1], which uses a bi-directional attention flow layer to combine both **Context-to-Query and Query-to-Context attention**. We added a **self-attention layer** on top of the BiDAF baseline. QANet instead uses **stacked self-attention** and demonstrates how a model architecture with only attention mechanisms can outperform those with recurrent and convolutional layers. The following diagram is a high-level summary of our model structure.

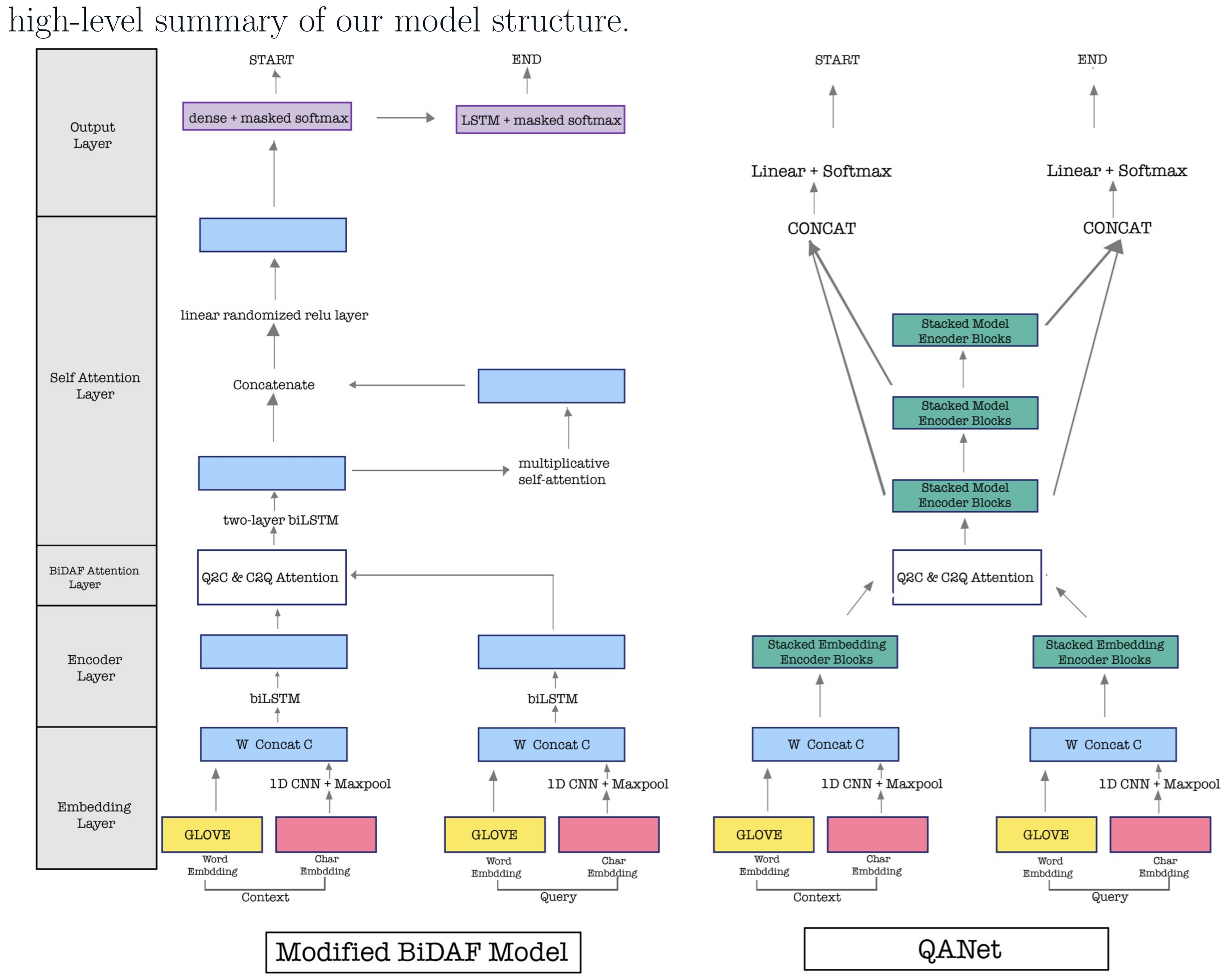

Figure 1:Architecture for BiDAF an QAnet

Examples

Context: Fourth, national courts have a duty to interpret domestic law "as far as possible in the light of the wording and purpose of the directive"…
Question: Which courts have a duty to interpret domestic law as far as possible?
Answer: national courts
QANet Answer: national courts
Modified BiDAF Answer: Fourth, national courts

Context: …Because oil was priced in dollars, oil producers' real income decreased…from then on, they would price oil in terms of a fixed amount of gold.
Question: Why did oil start getting priced in terms of gold?
Correct answer: oil was priced in dollars, oil producers' real income decreased
Prediction: No Answer

Context: …the totality of the exhaust steam cannot evacuate the cylinder, choking it and giving excessive compression (\"kick back\").[citation needed]…
Question: What is another term for excessive compression?
Correct answer: kick back
Prediction: kick back").[citation

Context: …The two forces finally met in the bloody Battle of Lake George between Fort Edward and Fort William Henry….
Question: Who won the battle of Lake Niagara?
Answer: No Answer
QANet Answer: Fort Edward and Fort William Henry
Modified BiDAF Answer: No Answer

Analysis

Compared to BiDAF, QANet often predicts answers with a more precise span boundary. In example 1, BiDAF prediction incorrectly includes **irrelevant words** to answer the question. QANet demonstrates better reading comprehension ability than BiDAF in general

Overcomprehension of QANet can be a caveat and limits its performance. Compared to BiDAF in example 2, QANet is more likely to generate predictions for unanswerable questions.

Example 3 demonstrates one of the most common mistakes of our model that it **fails to decide on the correct boundary of the answer**, and the answer is partially correct. This might also be due to the fact that there is an inherent ambiguity in deciding the boundary for an answer.

We found that our model **struggles to determine whether a question is answerable based on the context**, as shown in example 4. Such mistakes are much more common in "how" and "why" questions, which require deeper logical reasoning than other types of questions.

For Modified BiDAF, adding character embeddings and a self-attention layer boosts the performance drastically. From Tensorboard, it is apparent that the performance after adding the self-attention layer outperforms the model without self-attention from the very beginning. The Tensorboard plots for QANet show that increasing the number of attention heads and the size of model leads to overfitting on the Dev set after 1.5 million iterations.

Conclusions

We implemented an end-to-end neural Question Answering system for reading comprehension task on SQuAD 2.0. **The final ensemble model achieves 70.12 F1 and 67.13 EM on the dev set, and 68.10 F1 and 64.75 EM on the hidden test set.** From the results and the error analysis, we found that our model can effectively comprehend the provided context and synthesize information, but it struggles to determine whether a question is answerable. It also struggles to understand the contexts and questions that require more reasoning. Further work includes making the QANet model faster, exploring alternative attention methods like Transformers-XL, and ensembling using average logits instead.

References

[1] Seo et. al. "Bidirectional attention flow for machine comprehension." In: *arXiv preprint arXiv:1611.01603* (2016).

[2] Yu et. al. "QAnet: Combining local convolution with global self-attention for reading comprehension". In: *arXiv preprint* (2018).