SQuAD 2.0 Question Answering System

Stanford University

Chaonan Ye, Wanling Liu

{yec0214, liuwl} @ stanford.edu

Introduction

Reading and comprehending the human languages is a challenging task for machines, which requires understanding of natural languages and the ability to do reasoning over various clues. Question answering (QA) is one of popular problems in this field and has been actively researched in Natural Language Processing. QA has gained great significance and popularity since it has a wide range of applications, such as web search, e-learning and interactive voice response. In this project, we focus on designing a question answering system that has good performance on SQuAD 2.0.

Data

- **SQuAD 2.0 dataset**
- **150,000 questions in the format of <question, context, answer>**
- **50% of the questions that can be answered**
- **50% of the questions are not answerable using the given paragraph**
- **paragraphs are from Wikipedia**
- **questions and answers were crowdsourced using Amazon Mechanical Turk.**

Article: **Endangered Species Act**

Paragraph: *". . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."*

Question 1: **"Which laws faced significant opposition?"**

Plausible Answer: later laws

Question 2: **"What was the name of the 1937 treaty?"**

Plausible Answer: Bald Eagle Protection Act

Approach

- **BiDAF**
- **Character Embeddings**
- **Self - Attention Layers**
  - **From BiDAF ++**
  - **From Rnet**
- **BERTs**
- **Ensemble method**

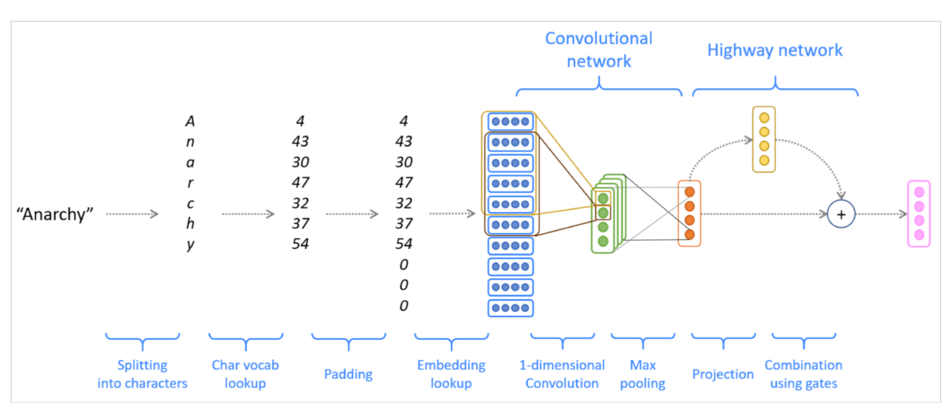

Figure 1: Character-based convolutional encoder[5]

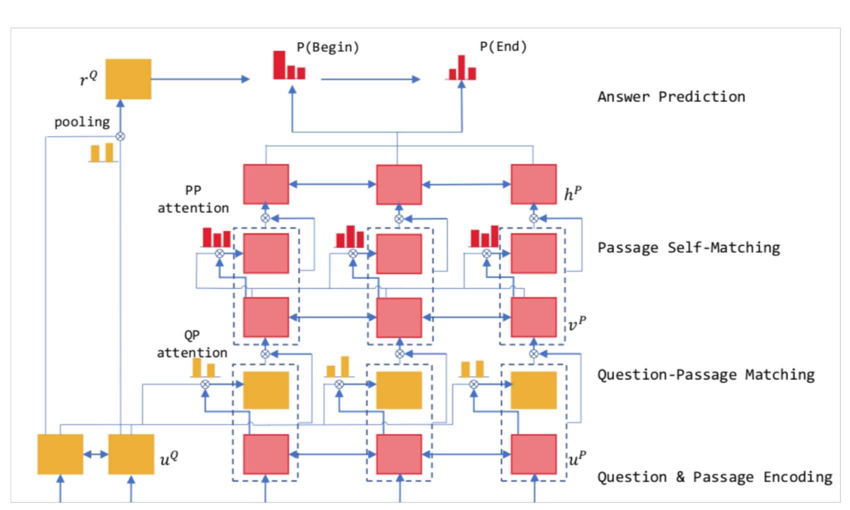

Figure 2: R-NET structure[3]

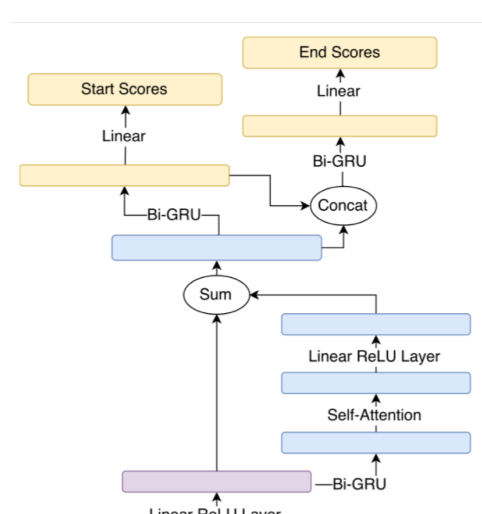

Figure 3: BiDAF++ structure[4]

Models

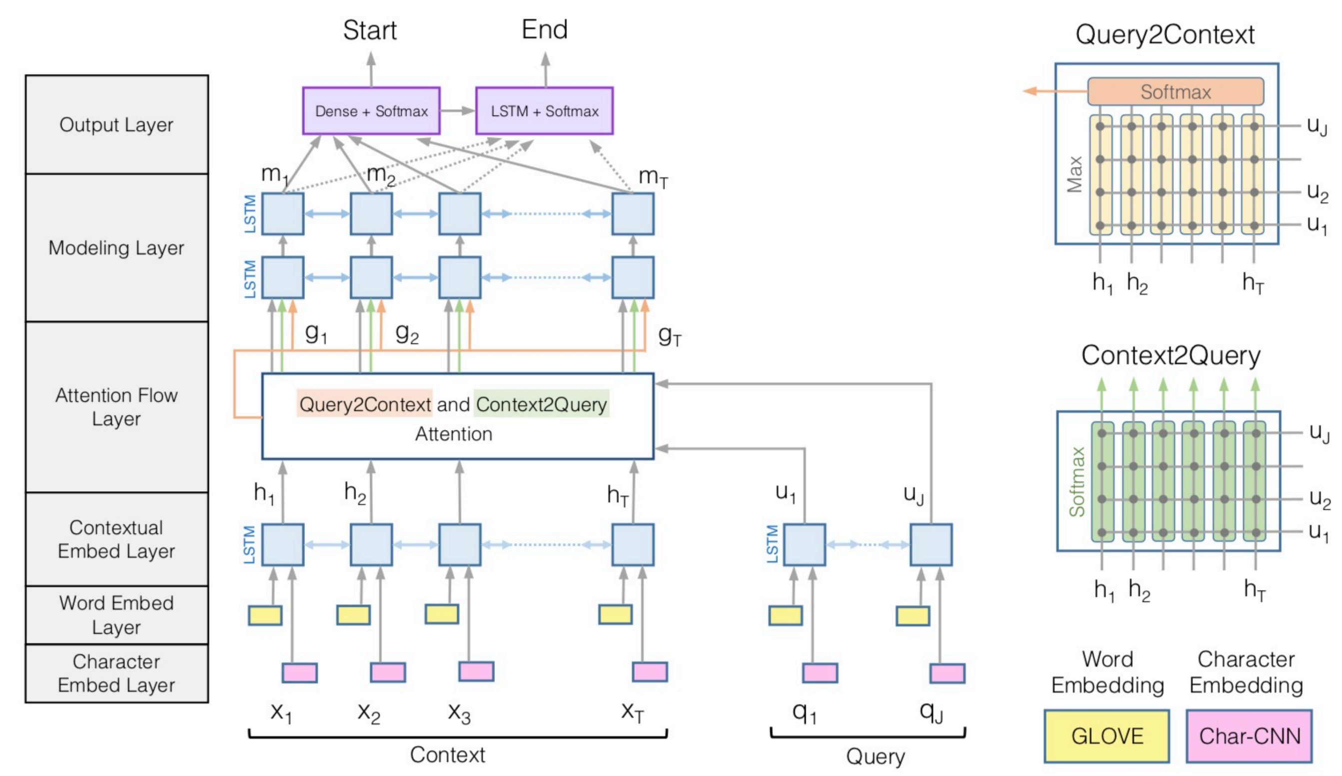

Figure 1: BiDirectional Attention Flow Model[1]

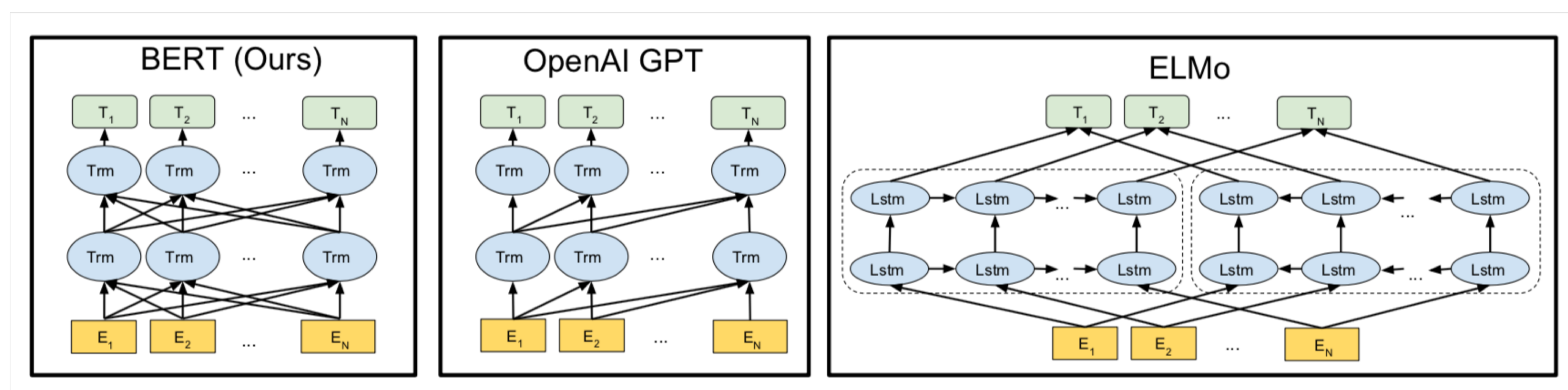

Figure 2: Differences in pre-training model architectures[2]

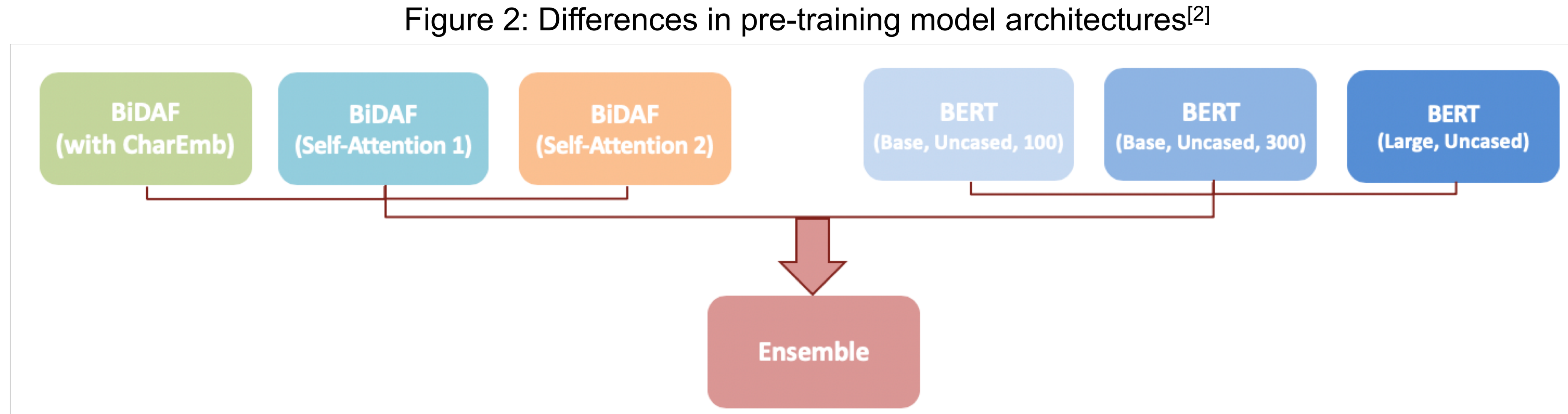

Figure 3: High-level structure of the final ensemble model

Experiments

- **The baseline BiDAF model with a character- level embedding layer performs a little better than the one without.**
- **The BERT-Base-Uncased model performs significantly better than BiDAF and BERT-Large-Uncased model performs better than BERT base model. When we increase the maximum sequence length, the result gets better.**
- **Self-attention layers do help model represent the question-related-context vector more suitably.**

Model	EM	F1
BiDAF with CharEmb	58.192	61.389
BERT-Base-Uncased (100)	71.356	73.859
BERT-Base-Uncased (300)	73.462	76.604
BERT-Large-Uncased	75.502	78.496

Table 1: The performance of baseline models

Model	EM	F1
SA(after Modeling layer)	55.08	59.98
SA(after Attention layer)	**57.60**	**61.33**
SA(Addition after Modeling layer)	56.36	59.92
Ref: BiDAF Starter Code	55	58

Table 2: Self-Attention (SA) from BiDAF++

- **Even only trained for 10 epochs, Self-Matching attention results are still comparable or even slightly higher than the BiDAF baseline model.**

Model	EM	F1
SMA v1	55.369	58.626
SMA v2	55.402	58.623
SMA v3	**56.125**	**59.548**

Table 3: Self-Matching attention (SMA) from R-NET

Results

We chose the ensemble model with the highest dev scores as our final model and tested it on the test set. This final ensemble model with 77.312 EM and 79.984 F1 on the test set is a reasonably good system.

Model	EM	F1
Ref: BiDAF Starter Code (dev)	55	58
Ensemble 1: One BERT + One BiDAF (dev)	75.683	78.552
Ensemble 2: Three BERTs + Three BiDAFs (dev)	**77.706**	**80.285**
Ensemble 2:(test)	**77.312**	**79.984**

Table 4: Ensemble results

Analysis

- **Question:** Who decides who gets to address the members of Parliament to share their thoughts on issues of faith?
- **Context:** The first item of business on Wednesdays is usually Time for Reflection, at which a speaker addresses members for up to four minutes, sharing a perspective on issues of faith. This contrasts with the formal style of "Prayers", which is the first item of business in meetings of the House of Commons. Speakers are drawn from across Scotland and are chosen to represent the balance of religious beliefs according to the Scottish census. Invitations to address Parliament in this manner are determined by the Presiding Officer on the advice of the parliamentary bureau. Faith groups can make direct representations to the Presiding Officer to nominate speakers.
- **Answer:** Presiding Officer
- **Prediction:** Presiding Officer

- **Question:** Who was Kaidu's grandfather?
- **Context:** Instability troubled the early years of Kublai Khan's reign. Ogedei's grandson Kaidu refused to submit to Kublai and threatened the western frontier of Kublai's domain. The hostile but weakened Song dynasty remained an obstacle in the south. Kublai secured the northeast border in 1259 by installing the hostage prince Wonjong as the ruler of Korea, making it a Mongol tributary state. Kublai was also threatened by domestic unrest. Li Tan, the son-in-law of a powerful official, instigated a revolt against Mongol rule in 1262. After successfully suppressing the revolt, Kublai curbed the influence of the Han Chinese advisers in his court. He feared that his dependence on Chinese officials left him vulnerable to future revolts and defections to the Song.
- **Answer:** Ogedei
- **Prediction:** N/A

If the question has keywords who are able to found in context and the answer is around the keywords, the model will correctly answer the question with high probability.

If the answer needs not only the context but also some logic, the model will return N/A at most of the time.

Conclusion

We introduce an ensemble of self-attention BiDAF models with character embedding and three fine-tuned BERT models. We use 2 types of self-attention layers in BiDAF in different locations. The experimental quantitative evaluations show that our model achieves the state-of-the-art results in SQuAD2.0. Therefore, our model is able to answer non-trivial questions by attending correct locations in the given context.

Reference

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Microsoft Research Asia Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks. Work-in-progress technical report posted on Microsoft.com, 2017.

[4] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723, 2017.

[5] http://web.stanford.edu/class/cs224n/assignments/a5.pdf