Question Answering on SQuAD 2.0 Dataset

Stanford University

Zhihan Jiang, Wensi Yin

CS 224N

Motivation

Machine question answering, which answers a question based on a given context, has gained great interests in recent years in the NLP area. By using various deep learning approaches such as recurrent neural networks and attention mechanisms, researchers have closed the gap between machine performance and human performance. We researched on the state-of-the-art techniques and built an end-to-end neural network model to tackle this problem.

Context: The immune system is a system of many biological structures and processes within an organism that protects against disease. To function properly, an immune system must detect a wide variety of agents, known as pathogens, from viruses to parasitic worms, and distinguish them from the organism's own healthy tissue.
Question: The immune system protects organisms against what?
Answer: disease

Dataset

SQuAD 2.0 is a machine learning comprehension dataset on Wikipedia articles with more than 150k questions. Different from SQuAD 1.1, there are roughly half of the questions that are unanswerable using the provided paragraph, which makes the prediction task more challenging.

For our training and evaluating purpose, we use the 224n-customized version of the SQuAD 2.0 dataset, and we split the dataset roughly by 90%, 5% and 5% for train, validation and test set. We measure the performance via two metrics: **Exact Match** (EM) score and **F1** score.

Result

The table below shows the EM score and F1 score of the models we built with each component added upon each other.

Model	F1	EM
Baseline BiDAF (w/o char emb)	60.61	57.52
Add char-emb	62.66	59.35
Add POS/NER features	62.78	59.10
Add QANet Encoder block	63.90	60.28
Add exact match (EM) feature	65.44	62.31
Add self-matching attention	**66.88**	**63.75**
Ensemble	**68.08**	**65.27**

Our test split results for **single model** is: **F1 63.66, EM 60.42, ensemble model: F1 64.70, EM 61.83.**

Approach

Embedding Layer: Our embedding vector consists of 4 parts: GloVe pre-trained word embedding (300-d), character-level embedding (200-d), part-of-speech (POS) (16-d) and name-entity-recognition (NER) (8-d) embedding, exact-match (EM) (3-d) feature.

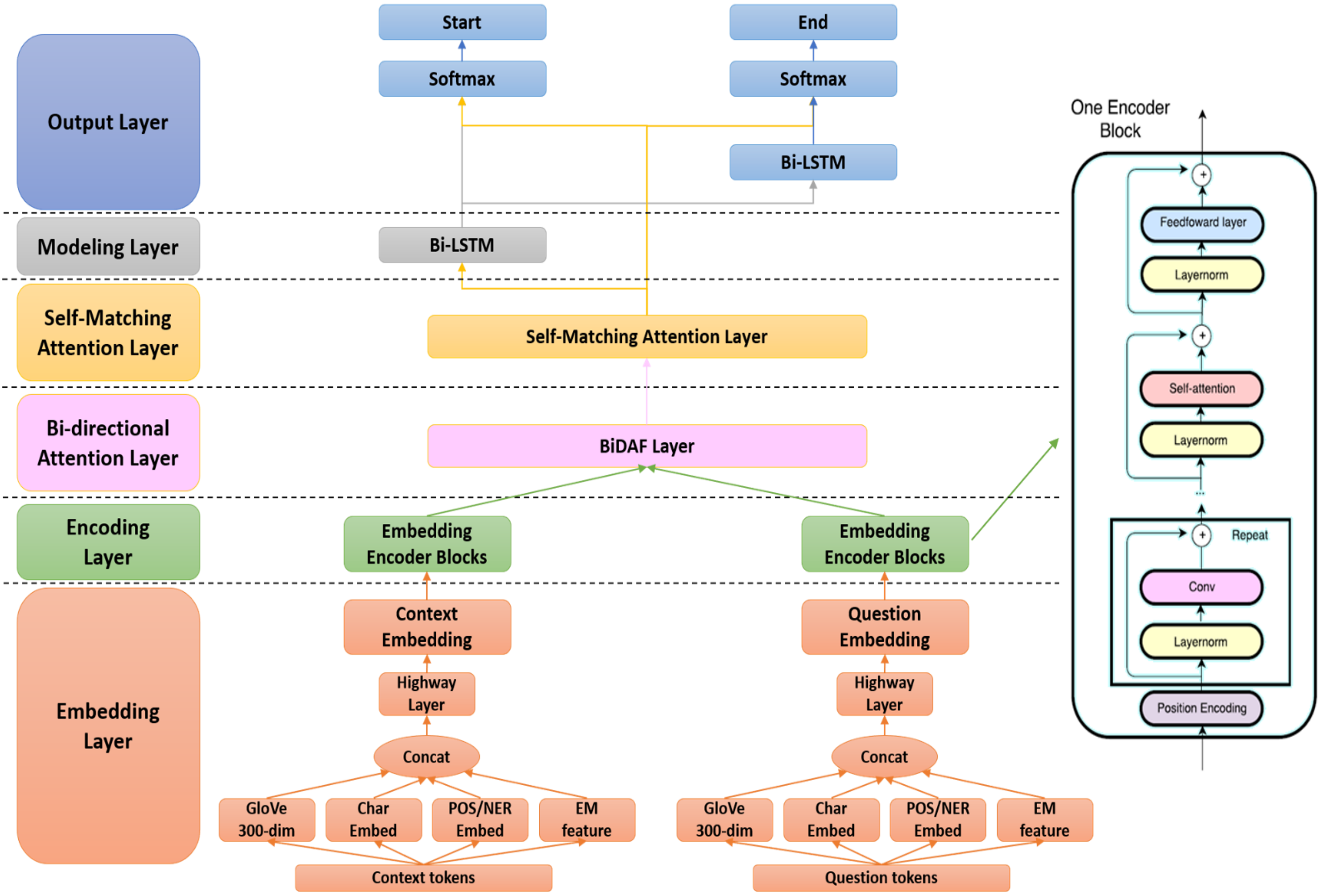
Encoding Layer: We follow the embedding encoder block design from QANet [2], which consists of one position encoding layer, four depthwise-separable convolution layers, one self-attention (multi-head attention) layer, one feed-forward layer.

Bi-directional Attention Layer: Follow the implementation in [1]. Combine the context representation and the question representation by concatenating the context-to-question attention and the question-to-context attention.

Self-matching Attention Layer: We use self-matching attention (inspired by RNet [3]) to dynamically collect evidence from the whole passage and encode them into the passage representation h_t^P. $h_t^P = BiLSTM(h_{t-1}^P, [v_t^P, c_t])$, where $c_t = att(v^P, v_t^P)$ is the scaled dot-product self attention of the whole passage (v^P) against itself.

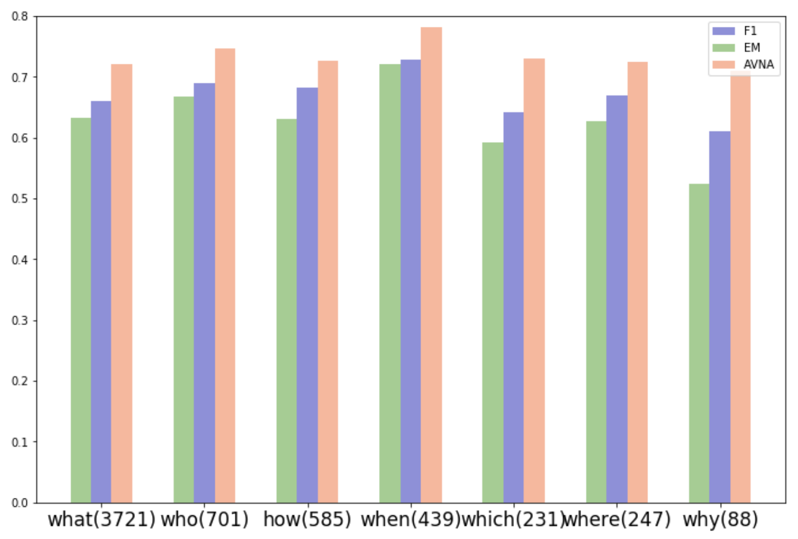
Modeling Layer: The modeling layer is a one-layer bidirectional LSTM which refines the sequence of vectors after the attention layer.

Output Layer: The output layer combines the attention layer and modeling layer output and produces a vector of probabilities corresponding to each position in the context: $p_{start}, p_{end} \in \mathbb{R}^N$. The span of the answer is chosen by finding the greatest $p_{start}(i) \cdot p_{end}(j)$, where i is the starting position and j is the ending position.

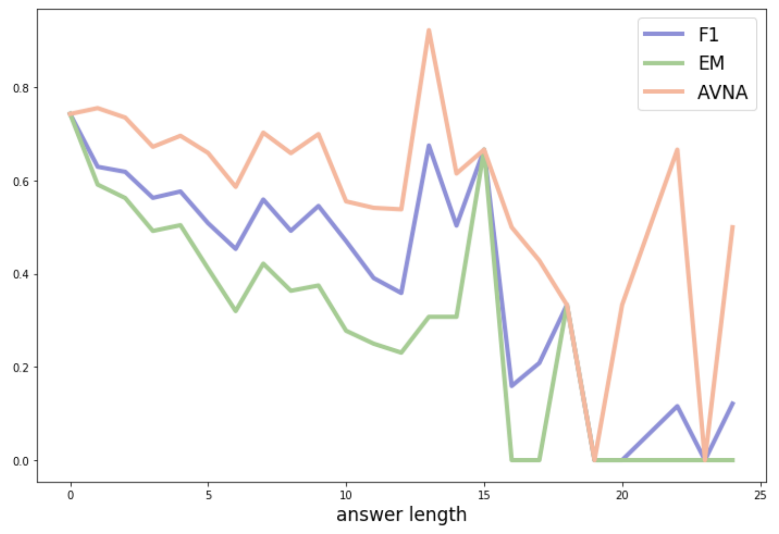


Analysis

Error Analysis: We find the model is better at answering "who", "when", "where" questions, but weaker at predicting the answers to "what", "how", especially "why". This shows that the model does a great job at finding obvious connections between words in the passage, but lacks the ability to understand the deeper logic and reasoning behind it.



Also, the model is better at generating short answers rather than long answers. This is because the model is weak at modeling long-term dependency.



[1] M. J. Seo et al., "Bidirectional attention flow for machine comprehension,"CoRR, vol. abs/1611.01603, 2016
[2] A. W. Yu et al, "Qanet: Combining local convolution with global self-attention for reading comprehension,"CoRR, vol. abs/1804.09541, 2018
[3] W. Wang et al, "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), vol. 1, pp. 189–198, 2017.