Neural Question Answering on SQuAD 2.0

Yiren Zhou[1]
[1]Department of Computer Science, Stanford University
Research Mentors: Elaine Sui

Introduction

This work aims to investigate innovative designs of model architectures that can help boost performance on the SQuAD 2.0 dataset, without using pre-trained language models. Since around half of the questions are unanswerable, it is important for the model to tactfully abstain from answering. The main contribution is a self-implemented QANet architecture with extensions on the embedding layer and the output layer. By using unified encoding for the context and question before feeding into context-query attention and employing threshold-based answer verification during testing, the model achieves stronger out-of-sample performance than the original QANet baseline. With a novel debiased ensemble method, the model achieves an EM score of 67.68 and F1 score of 70.53 on the test leaderboard for the IID SQuAD track.

Dataset

The SQuAD 2.0 dataset contains IID (context, question, answer) triples. The training data consists of 129941 examples. The dev and test set both consist around 6k examples. The raw data is pre-processed to tokens and then represented by the pre-trained GloVe word vectors combining with randomly initialized trainable character embeddings.

Model Architecture

Baseline: QANet

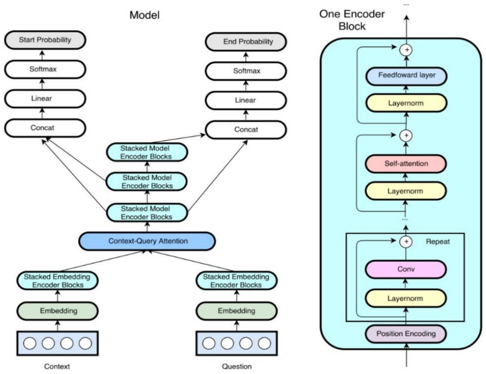


Figure 1: The QANet architecture. Source: [3]

Extension

Unified encoding (Unified QANet) Inspired by [8], we implemented a unified embedding layer, which encodes the concatenated question and context representation. It performs the same operations as the QANet input embedding layer, with an additional trainable segment embedding to indicate whether the word belongs to context or passage segment. For a single word representation x_j, the final output is the sum of the original output of highway network and the segment embedding:

$$x_j = \text{highway}(W[x_j^w; x_j^c]) + x_{s_j}, \ \ x_{s_j} \in \mathbb{R}^h, s_j \in \{0, 1\}$$

Threshold-based answerable verification (TAV) From the original QANet paper, $c_0 = \text{OOV}$ (Out of Vocabulary) is inserted at the beginning of each paragraph as a start placeholder, and span $S = \{c_0\}$ will indicate a no-answer output. Among all the valid starting and ending position pairs, $(0, 0)$, which indicates no answer, is only one of these $l(l+1)/2$ possible cases (l is the sequence length of the context). This makes QANet frequently give plausible answers to unanswerable queries.

Ensemble

Assuming there are k models $\mathcal{M}_1, ..., \mathcal{M}_k$ from a I.I.D model distribution, then the majority voting answer score is approximately proportional to predicted density. Therefore, we define the new length-debiased score by:

$$\text{score}_A = \sum_{i=1}^{k} \mathbb{I}\{\mathcal{M}_i(C, Q) = A\}$$

$$\text{score}_A^{debiased} = \text{score}_A \times \frac{P_l(l = l_A)}{P_l^{\mathcal{M}}(l = l_A)}$$

Results

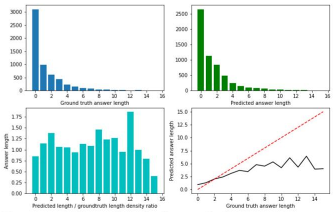
Model	EM	F1	AvNA
BiDAF (Baseline)	58.01	61.26	68.27
BiDAF + Character Embeddings	60.36	63.41	69.37
QANet	66.27	69.45	75.32
UnifiedQANet	66.32	69.72	**76.02**
UnifiedQANet (1 more embedding encoder)	66.09	69.49	74.79
UnifiedQANet (1 more model encoder)	65.94	69.38	75.47
UnifiedQANet + TAV1	65.48	68.65	74.09
UnifiedQANet + TAV2	**67.13**	**69.97**	75.52
Ensemble	69.50	72.35	N/A
Ensemble + Length debiasing	**70.32**	**73.10**	N/A

The Unified QANet performed slightly better than the QANet, but the difference is not significant enough.

The Unified QANet with TAV2 achieved the highest performance. The result is expected since a perfect score for a specific example can be achieved by successfully predicting unanswerability

Finally, we see that the two ensemble of the seven models both achieve significant improvement in EM and F1 score on the development set.

Analysis



The upper two plots in figure 2 show the histogram of ground truth answer length and predicted answer length. The lower left is the density ratio of the predicted length and ground truth length.