Url: https://youtu.be/mYx5mmRwp8g

For poster see next page

Reading Comprehension for SQuAD 2.0 using U-Net

Qingyun Wan
qywan@stanford.edu

Motivation

Machine Reading Comprehension and Question Answering are challenging tasks in natural language processing as they require a system to understand a given context and parse various kinds of questions to select the correct text span from the context as their answers. SQuAD is designed for evaluating such reading comprehension systems by extracting contexts and building questions from Wikipedia.

However, such dataset suffers from allowing random guesses out of the context when the questions are virtually unanswerable. SQuAD 2.0 is therefore constructed to overcome this weakness by including unanswerable questions so the systems that predict plausible answers for unanswerable questions will be punished. The goal is to implement a non-PCE model that achieve competitive performance on this dataset compared with the baseline model BiDAF..

Data

Around 150k questions in SQuAD 2.0 with almost half unanswerable split into:

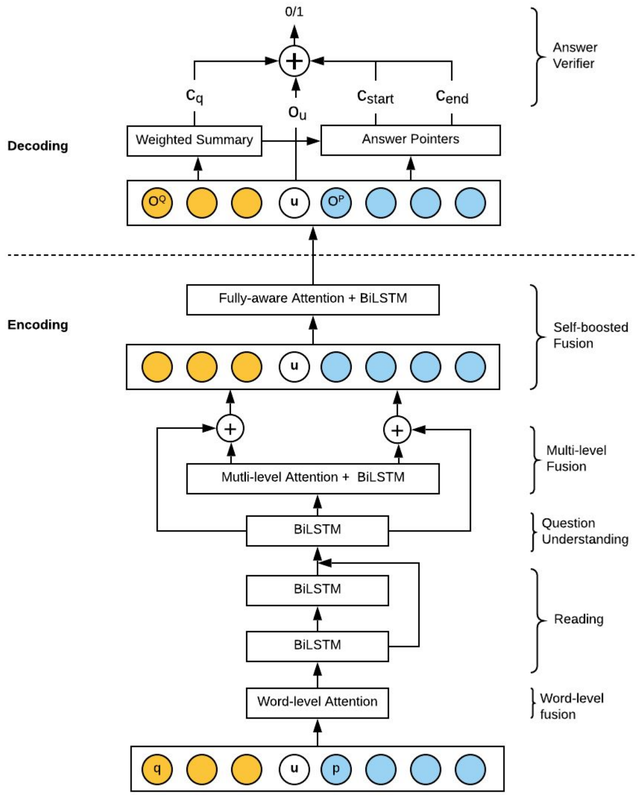
- train (129941): SQuAD 2.0 training set.
- dev (6078): randomly selected half of the dev set
- test (5915): remaining questions from the dev set, plus hand-labeled examples.

Task

- Train an **end-to-end multi-task** model that is able to verify if question is unanswerable and predict the correct answer span from the context for answer questions. The model is based on U-Net[1]
- Ablation study of he best model on both input features and model components to understand the model and ensembling to reveal potential improvement.

Methodology

U-Net = Universal Node + FusionNet[2] + Answer Pointer + Answer Verifier



Results

Model	EM %	F1 %
baseline	56.38	59.75
U-Net with Glove only	57.42	61.76
U-Net with Glove + plausible answer	59.60	63.52
U-Net with Glove + plausible answer + additional features	64.51	68.79
U-Net with Glove + plausible answer + additional features + character embedding	65.41	69.48

Performance of single models with different features on the dev set

Analysis

- Ablation

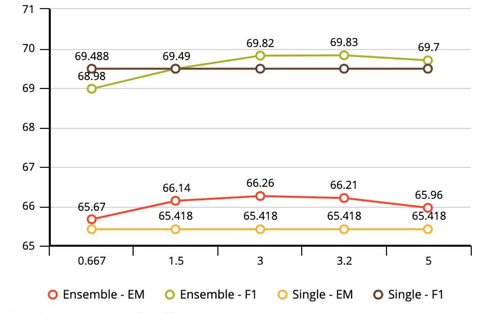
	EM change in %	F1 change in %
No high-level hidden states	-2.22	-2.2
Use multi-head self-attention	-0.58	-0.66
More weight on answer pointer's loss	-1.95	-0.76
More weight on answer verifier's loss	-0.28	-0.14

- Error Analysis

Context: ...A key distinction between analysis of algorithms and computational complexity theory is that the former is devoted to analyzing the amount of resources needed by a particular algorithm to solve a problem, whereas the latter asks a more general question about all possible algorithms that could be used to solve the same problem...
Answer:"computational complexity theory"
Baseline model prediction:"computational complexity theory"
U-Net model prediction:"theoretical computer science are analysis of algorithms and computability theory. A key distinction between analysis of algorithms and computational complexity theory"

- Ensemble

Ensemble with the baseline BiDAF model:



Performance of different ensemble weights on the dev set

Conclusion

The U-Net model achieves small performance gain compared to the baseline without additional features which further largely improves this model. The U-Net model has some fundamental problems in narrowing down the correct answer span compared to the baseline so if we can improve the model structure (basically FusionNet) of U-Net learning from the BiDAF, it might perform much better.

[1] Sun F, Li L, Qiu X, et al. U-Net: Machine Reading Comprehension with Unanswerable Questions[J]. arXiv preprint arXiv:1810.06638, 2018.
[2] Huang, H.; Zhu, C.; Shen, Y.; and Chen, W. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. CoRR abs/1711.07341.