Advancing with Adversaries: Comparing LSTMs Across Adversarial Inputs

Angela Chen, Darian Martos, Jerold Yu

Problem

Recently, more effort has been made to increase robustness against adversarial examples in reading comprehension systems. Robust systems are suggested to have "real language understanding abilities" [1] and are more transferrable to real-world question answering tasks (e.g. social media posts). Modern approaches attempt to model more complex relationships between the question and context [2] or encourage the identification of an adversarial example. We look to explore these approaches in more detail.

Data/Task

We will train and evaluate our model on the SQuAD 2.0 dataset. SQuAD 2.0 contain examples that have unanswerable questions.

Example:

Paragraph: King David I of Scotland, whose elder brother Alexander I had married Sybilla of Normandy, was instrumental in introducing Normans and Norman culture to Scotland, part of the process some scholars called the "Davidian Revolution."
Question: What did Sybilla of Normandy introduce to Scotland?
Answer: N/A
Model Predicts: Normans and Norman culture

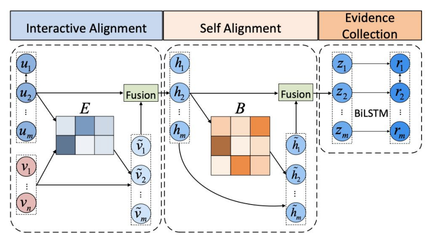
Using non-PCE models, we want to improve on the baseline QA system. The baseline QA model predicts non-existent answers on passages that don't contain an answer to a given question. By improving upon no-answer conditions, our goal is to improve our baseline QA scores.

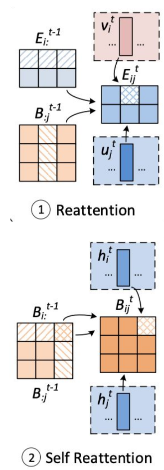
Approach

Character embeddings: added using CNN's to better capture the internal structure of words and predict OOV words better

Reattention: to capture more complex interactions between the question and context, we modify BiDAF attention to a reattention mechanism, a multi-round alignment architecture:

$$\tilde{E}^t_{ij} = \mathrm{softmax}(E^{t-1}_{i:}) \cdot \mathrm{softmax}(B^{t-1}_{:j})$$
$$E^t_{ij} = f(v^t_i, u^t_j) + \gamma \tilde{E}^t_{ij}$$
$$\tilde{B}^t_{ij} = \mathrm{softmax}(B^{t-1}_{i:}) \cdot \mathrm{softmax}(B^{t-1}_{:j})$$
$$B^t_{ij} = \mathbb{1}_{(i \neq j)} \left(f(h^t_i, h^t_j) + \gamma \tilde{B}^t_{ij} \right)$$





Results

Model	EM	F1
BiDAF	56.298	59.920
BiDAF + Char Embed	58.394	62.413
BiDAF + Char Embed + Reattn	59.121	62.979

Analysis

In order to analyze the effect of reattention, we compare the full model's performance with the character embeddings-only model. In general, the full model is able to model complex interactions between question and context better:

Question: How did peace start?
Context: The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North. It began with a dispute over control of the confluence of the Allegheny and Monongahela rivers, called the Forks of the Ohio, and the site of the French Fort Duquesne and present-day Pittsburgh, Pennsylvania. The dispute erupted into violence in the Battle of Jumonville Glen in May 1754, during which Virginia militiamen under the command of 22-year-old George Washington ambushed a French patrol.
Answer: N/A
Char Embed Prediction: with a dispute over control of the confluence of the Allegheny and Monongahela rivers
Reattn Prediction: N/A

However, the full model did not do any better in terms of adversarial input (i.e. unanswerable questions)

Future Work

Adversarial evaluation. Adversarial data available for SQuAD 1.1; would be useful to evaluate models on updated data sets.

Improving on unanswerable questions. Predicting no-answer will improve performance on adversarial data; can implement a no-answer reader or a modified objective loss function.

References

[1] R. Jia and P. Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Empirical Methods in Natural Language Processing (EMNLP) 2017*.

[2] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, and M. Zhou. 2018. Reinforced Mnemonic Reader for Machine Reading Comprehension. In *27th International Joint Conference on Artificial Intelligence (IJCAI)*.