Question Answering with Self-Attention and Residuals

Luyao Hou
Stanford University

Introduction

Machine question answering has led to lots of applications like chatbots and improved search engines. Most current question answering models use recurrent neural networks (RNNs) to encode sequential information inherent within texts. However, due to their sequential nature, RNN models can be slow to train and difficult to scale up.

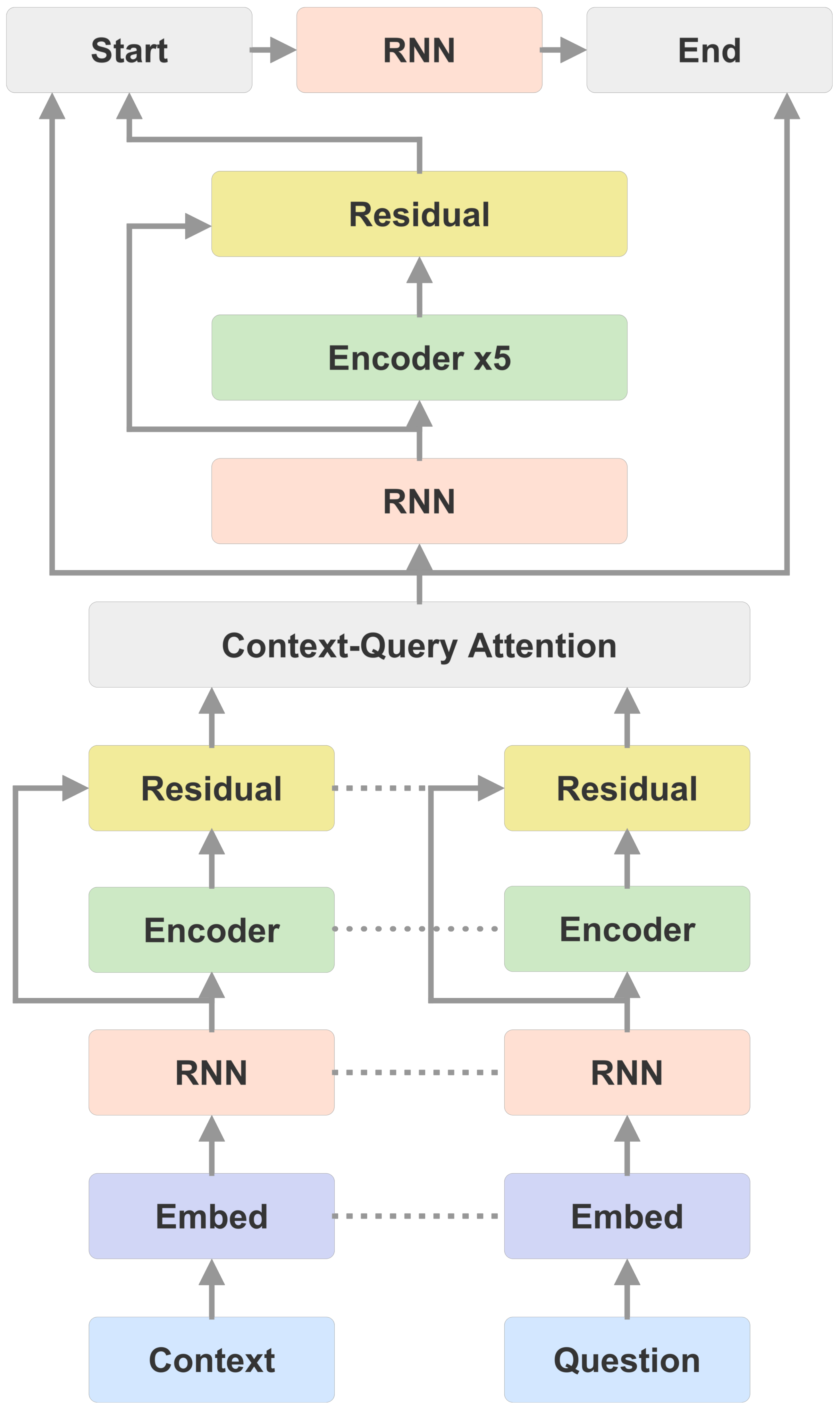
In this project, we explore the possibility of combining RNNs with convolution and self-attention to speed up training as well as improving model performance. With the SQuAD2.0 dataset, the baseline BiDAF model achieved a F1 score of 59.92 while our model achieved a F1 of 60.90 with less training steps.

Dataset

We used SQuAD2.0 dataset to both train and evaluate our model. SQuAD2.0 contains context/question pairs where there is either no answer to a question or the answer must be a span of the contexts.

In particular we split the dataset into 129,941 examples for training, 6078 examples for dev and 5921 examples for test.

Approach

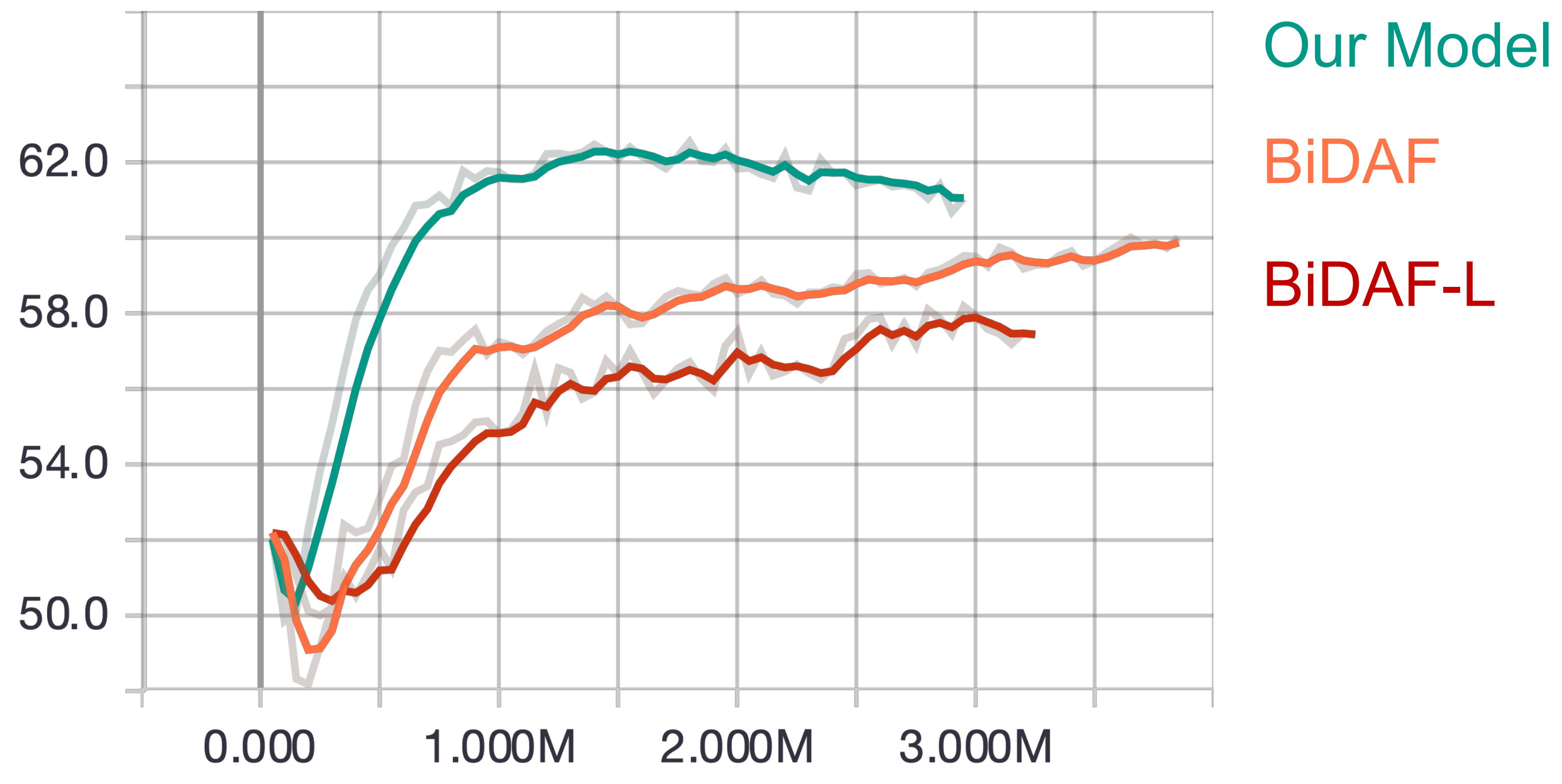


Each Residual layer has learnable parameters that determine how much each input is summed into its output. Each encoder is composed of stacked convolution, self-attention[1] and linear layers as proposed in QANet[2].

Results

	EM	F1	Steps
BiDAF	56.30	59.92	3.9M
BiDAF-L	54.55	57.94	3.0M
Our Model	**57.36**	**60.90**	**1.5M**

F1 score vs training steps



Analysis and Conclusions

From the F1 score plot we observe that our model converges to a higher F1 more quickly than the baseline BiDAF model as the residual connections provide shortcuts. However, the early overfitting of our model suggests that further tuning may produce higher F1 scores. Future work could focus on replacing more RNNs by self attentions to build more parallelism into the model.

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998-6008. 2017.
[2] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.