**INTRODUCTION**

In [1], Adams et al. proposed QA4Net, a question-answering model based on convolutions and self-attention instead of recurrent neural networks. At the time of publication of the paper (Aug 23, 2018), the most successful question-answering models were generally based on recurrent neural networks (RNNs) with some attention mechanisms. One weakness of these models is that their recurrent nature precludes parallel computation, making training and inference slow. These were efforts to speed up the RNNs by avoiding the desired attention [2] by replacing the context-query attention model. This model has been tested on the SQuAD v1.1 dataset, and its F1 scores were around 77.

QA4Net solves this problem by moving away from RNNs instead of using convolutions and self-attention as the main building blocks. This approach gives the model a significant speed advantage (especially between 4 or 12 times faster in training and 4 to 8 times in inference) over its RNN counterparts. Taking advantage of this speed boost, the authors trained the model with augmented data and achieved an F1 score of 84.4 on the SQuAD dataset, which was significantly better than the best published result at the time.

QA4Net is PyTorch was implemented in this project and tested on SQuAD 2.0. While the transition to SQuAD 2.0 is straightforward, it is challenging to reproduce the performance, especially the speed, reported in the original paper.

**DISCUSSION**

The work presented in this paper is a direct descendent of QA4Net. This model is intended as a direct replacement for QA4Net in the QA4Net+ model.

*This paper was co-authored by a student in the NLP course at Stanford University.*

**REFERENCES**


