Simply Improving BiDAF Model Heuristically

Stanford CS224N Default Project: Building a QA system (IID SQuAD track)

Nathaniel Goenawan
Department of Computer Science
Stanford University
nathgoh@stanford.edu

Christopher Wong
Department of Computer Science
Stanford University
cwong7@stanford.edu

1 Key Information to include

We will be doing the default final project of building a QA system (IID SQuAD track). We do not have any external collaborators, and we are not sharing this project with any other class.

2 Research paper summary (max 2 pages)

Title	Making Neural QA as Simple as Possible but not Simpler
Year	2017
URL	https://arxiv.org/abs/1703.04816

Table 1: QA system tested on NewsQA dataset [1]

Background. Question answering (QA) is an important end-user task for natural language processing and information retrieval. The authors points out though that most QA systems are built in a *top-down* approach and thus propose complex architecture and raises the concern as to whether the aforementioned architecture justifies their empirical results. As such, the authors are proposing that their simple context/type matching heuristic can be used to provide a guideline for simple neural baseline architectures.

Summary of contributions. The authors introduced a simple context/type matching heuristic which can be used to provide a guideline for simple neural baseline architectures. Their RNN-based system shows itself to be an efficient neural baseline architecture for extractive question answering. Specifically, this system combines two important ideas that the authors propose are arguably integral and necessary for a successful, high-performing neural QA system; namely, the awareness of question words while processing the context and a composition function that goes above simple bag-of-words modeling, such as RNNs.

Limitations and discussion. The authors claim that their two proposed "ingredients" are integral for a "currently competitive QA system" after having trained on the SQuAD (consisting of questions about paragraphs of Wikipedia articles) and NewsQA dataset (consisting of questions about CNN news stories). While their statement may be true for English, it remains to be seen as to whether these "ingredients" are truly integral for any competitive QA system, even those for languages other than English, such as more linguistically "complicated" languages. Furthermore, the authors acknowledged that their model is not able to properly address linguistically motivated errors such as lack of co-reference resolution and context sensitive abbreviations, as well as struggling to capture basic syntactic sentence structure, especially with nested sentences with ignored punctuation and conjunctions, so their proposed model is not necessarily able to accurately deal with more linguistically complex sentences. In context, these limitations do not make the authors' result unconvincing but rather lead the way for future work in confirming the authors' proposed beliefs.

Why this paper? We chose "Making Neural QA as Simple as Possible but not Simpler" because it provided a basic understanding and requirements of what a question answering systems needs to have. The paper's direction is contrary to many currently proposed increasingly complex systems to solve QA, arguing for a more simple heuristic to guide neural baseline systems. The authors' proposed model was also tested on multiple datasets rather than just testing on SQuAD. The authors' proposal is something we feel has merit and potential applications for our proposed model.

Wider research context. This paper is especially relevant in the broader story of NLP research because it proposes a novel concept contrasting many aforementioned architectures, namely that an extractive QA system does not have to solve the complex reasoning types that Chen et. al (2016) [2] proposes to classify SQuAD instances in order to achieve competitive, state-of-the-art results. This is important to keep in mind, as the paper and its results show that more complex models may not necessarily always be the answer. This was indicated by the authors' comparison of their model to representative models that have a complex word-by-word interaction layer such as FASTQAExt (an alternate proposed model by the same authors) and the Dynamic Coattention Network proposed in 2017 by Xiong et al. [3], showing a halving in time needed and between two and four times less memory needed for the authors' model relative to the more complex models while maintaing similarly competitive results. However, the proposed model also maintains the seemingly ubiquitous problem of being unable to properly deal with linguistically complex sentences (specifically the ones previously mentioned in the limitations and discussion sections).

Title	QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension
Year	2018
URL	`https://arxiv.org/abs//1804.09541`

Table 2: Local Convolution + Global Self-Attention [4]

Background. The paper recognizes that over the past few years, the most successful models for machine reading comprehension and automated question answering generally employ recurrent models to process sequential inputs and some attention component to deal with long-term interactions. However, the authors recognize a weakness with the aforementioned models, namely that due to their recurrent nature, these models are often slow in both training and inference, with this problem being amplified in use with long texts; the slow training problem limits larger dataset usage for the models, while the slow inference problem prevents real-time application deployment.

Summary of contributions. In response to the previously mentioned problems, the authors of this paper propose to remove what they believe to be the source of the problems, the recurrent nature of these models, using exclusively convolutions and self-attentions to build the encoders that separate the question and context, learning the interactions between the two via standard attention. The resultant model is fully feedforward and its construction consists of separable layers suitable for parallel computation, and its results are both fast and accurate, surpassing the best published results on the SQuAD dataset while surpassing the speed for training and inference a competitive recurrent model by between 3x to 13x faster in training and 4x to 9x in inference.

Limitations and discussion. One of the limitations/flaws of the paper is that the authors made their argument using mainly their model evaluated on English Wikipedia text via SQuAD, as opposed to the previous paper, which also tested their model on another dataset (NewsQA). As a result, a more stronger argument for the authors could have been made had the authors also tested their concept and model on other datasets (especially of a different form than Wikipedia text). While the authors did test on a different dataset, TriviaQA, which is more challenging in that its examples have longer context and is much noisier, they only focuses on comparing with the single-paragraph reading baselines, avoiding actually testing on multi-paragraph reading methods and simply "believing" that their model can be plugged into other multi-paragraph reading methods to achieve similar or better performance that models with recurrent natures. Furthermore, the authors combine their model with a complementary data augmentation technique to enhance the training data, which paraphrases the examples by translating the original sentences from and back to English with an intermediate other language, diversifying phrasing and enhancing the number of training instances. The authors admit that the robustness of their model is likely because of its training with augmented data. In context, these limitations make the authors' result a bit more unconvincing, as the arguable cherry-picking of datasets with TriviaQA as well as the training data augmentation make it difficult to compare the authors' proposed models with other models that do not undergo the same experiences.

Why this paper? We chose "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension" because of its novel approach of removing the recurrent nature of the proposed model. The paper's direction is contrary to many currently proposed systems which use recurrent models to process sequential inputs. The arguable simplicy of the model was something of interest to us, as it would be incredibly interesting and valuable to try a simpler model lacking in a recurrent nature and testing how it performs against models not lacking that same nature. However, the authors' proposed model was not (at least, to us) really tested on multiple datasets, instead really just testing on SQuAD. Therefore, while we do believe that the authors' proposal is something we feel has merit and potential applications for our proposed model, we're looking at their proposed approach with what we believe to be a reasonable amount of caution.

Wider research context. This paper is especially relevant in the broader story of NLP research because it proposes a novel concept contrasting many aforementioned architectures, namely that a model does not need to have recurrent networks in order to achieve competitive, state-of-the-art results. This is especially interesting in context with what we know to be problems that tend to occur when applying deep learning methods to language; specifically, it is interesting that a model that discards the recurrent networks in favor of feed-forward architectures is able to perform competitively with state-of-the-art architectures despite not having recurrent networks for processing sequential inputs, something we have learned to be very integral to recent success with QA systems (and general NLP problems). Being able to forgo the recurrent nature of a model allows for parallel computation, as the sequential nature necessitated by RNNs prevent parallel computation (as they require tokens to be fed in order). Attempts have been made to replace recurrent networks with full convolution or full attention architectures (such as by Kim in 2014 [5] and Gehring et al. in 2017 [6]), and these models have been shown to be faster in not only QA tasks but also other tasks such as text classification, machine translation, and sentiment analysis.

3 Project description (1-2 pages)

Goal. The goal for our project is to determine the efficacy of implementing a context/matching heuristic on top of BiDAF. We are investigating whether or not the heuristic can help improve the model's performance on the SQuAD 2.0 dataset.

Our motivation for this goal comes from the resulting performances individually of context/matching heuristic and BiDAF. Individually, during the time of publishing, both implementations showed state-of-the-art performance on the SQuAD 1.0 dataset. Hence, we wanted to see if by combining the two implementations, we could see favorable if not potentially state-of-the-art performance on the SQuAD 2.0 dataset, which also contains unanswerable questions.

Task. Our task is to implement deep learning techniques for question answering on the SQuAD (Stanford Question Answering Dataset) 2.0. Our model will be given two inputs, a paragraph and a question about that paragraph. The output from our model should be a correct answer for the question.

Data. We will be using the SQuAD 2.0 dataset. The dataset is split into train, dev, and test sets. The train set contains 129,941 examples, all taken from the SQuAD 2.0 dataset. The dev set contains 6078 examples, roughly half of the official dev set, randomly selected. Finally, the test set contains 5915 examples, which come from the remaining dev set and are supplemented by hand-labeled examples. Preprocessing is provided by the starter code, first processing the train set and obtaining word and character vocabularies. With the obtained word and character vocabularies, we used it to process the dev and test sets to obtain the context and question features.

Methods. For our proposed model, we plan to first fully implement the BiDAF baseline model as described in [7]; specifically, we will implement the character-level embedding layer into the BiDAF baseline model, as the current baseline only implements the word-level embedding layer. Next, we will implement the context/matching heuristics as described in [1] to serve as a guideline for the fully implemented BiDAF model.

Baselines. For our baseline, we will use the provided baseline model in the starter code for IID SQuAD track. The baseline is a Bi-Directional Attention Flow for Machine Comprehension model based on the paper written by [7]. It should be noted that for the baseline, only the word-level embedding layer is implemented.

Evaluation. For a well-defined, numerical, automatic evaluation, we will be using Exact Match (EM) and F1 scoring. EM is a metric that measures the percentage of predictions that match any one of the ground truth answers exactly. F1 score is a measure of a model's accuracy on a dataset, the harmonic mean of precision and recall.

For some qualitative evaluations, we will do qualitative error inspection of predictions for the SQuAD dev set. From these inspections, we hope to highlight basic abilities that are missing in our model to reach human-level performance. For example, some error we could be potentially highlighting are lack of syntactic understanding or lack of fine-grained understanding of word meaning.

References

[1] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural qa as simple as possible but not simpler, 2017.

[2] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering, 2018.

[4] 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, 2018.

[5] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[6] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 06–11 Aug 2017.

[7] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension, 2018.