Self-Attention for SQuAD Question Answering

Default Final Project (IID SQuAD)

Danny Tse
Department of Computer Science
Stanford University
dannytse@stanford.edu

1 Research Paper Summary

Title	R-Net: Machine Reading Comprehension with Self-Matching Networks
Group	Natural Language Computing Group, Microsoft Research Asia
Year	2017
URL	`https://www.microsoft.com/en-us/research/wp-content/uploads/2017/05/r-net.pdf`

Background. In this paper, the primary problem the authors worked to solve was developing a neural network with reading comprehension skills - in other words, given a passage or document, the model could answer questions related to the text that require logical reasoning. This work builds and improves on previous work, such as Bahdanau et al. (2014)'s *Neural Machine Translation by Jointly Learning to Align and Translate*, Rocktaschel et al. (2015)'s *Reasoning about Entailment with Neural Attention*, Wang and Jiang (2016)'s *Learning Natural Language Inference with LSTM*, and Wang and Jiang (2016)'s *Machine Comprehension Using Match-LSTM and Answer Pointer*. Utilizing novel methods such as a gated attention-based recurrent network and self-matching, the R-Net model described in this paper is able to achieve state-of-the-art (at time of publishing) results on both the Stanford Question Answering Dataset (SQuAD) and the Microsoft Machine Reading Comprehension (MS-MARCO) datasets.

Summary of contributions. The paper primarily focuses on two datasets: the Stanford Question Answering Dataset (SQuAD) and the Microsoft Machine Reading Comprehension (MS-MARCO) datasets. In the SQuAD dataset, answers to the questions are constrained to the given text passage, versus in the MS-MARCO dataset, answers to the questions are human-generated and are not constrained to the given document. As stated in the paper, the contributions can be showcased in three central points. First, the paper introduces a gated attention-based recurrent network. When compared to previous attention-based recurrent network models like Bahdanau et al. (2014), Rocktaschel et al. (2015), Wang and Jiang (2016), the proposed model contains an extra gate to weight parts of the given text based on their relevance to the question, and masks out more irrelevant components. Next, the paper introduces a self-matching mechanism, where the model aggregates information from the text to infer the answer. This self-matching mechanism is implemented by using gated attention-based recurrent networks on the text against itself (versus text against question from earlier). This way, the model can indeed aggregate relevant information from the text relevant to some current word from the text. Finally, combining both of these new methods, the model outperforms strong baselines, taking over first place on the SQuAD leaderboard (at time of publishing) and claiming the best published results on the MS-MARCO dataset (at time of publishing).

Limitations and discussion. Like all machine learning models, the one described in this paper is not perfect - it is prone to shortcomings and limitations. One limitation is the scarcity of data. The authors of the paper mentioned that though texts are abundant, it is difficult to come across texts along with question-answer pairs like in SQuAD. This does not reduce the credibility of the paper, as the model still outperformed many other models of its time, and took the top published result for MS-MARCO. However, the authors of the paper utilized a sequence-to-sequence question generation model from Zhou et al. (2017)'s *Neural Question Generation from Text: A Preliminary Study*, and failed to improve the R-Net's performance. This showcases how the sequence-to-sequence question generation model requires improvement; doing so would be an interesting experiment and would improve the R-Net's performance. Another limitation is syntax information, where the authors attempted and failed to utilize syntax information in the model. They used three methods: (1) POS tags, NER results, linearized PCFG tags and dependency labels, (2) tree-LSTM style module after encoding layer, and (3) dependency parsing. A final limitation mentioned in the paper is multi-hop inference, where the authors attempted adding multi-hop inference modules in the answer pointed layer, but did not get any improvements. Though these methods were claimed to have failed, the paper would have been stronger if the authors had included the accuracies using each method along with accuracies using combinations of these methods. These limitations showcase that the model is not perfect and improving it may be more difficult that some of the other models during its time, but the paper's credibility does not diminish in any way - the model's performance speaks for itself.

Why this paper? I chose this paper because it was a good introductory paper suggested by a TA in the handout. Also, the R-Net uses self-attention, something I have exposure to from lecture. As this is my first Machine Learning course, I hope to start with a smaller and more easily digestible paper. Furthermore, the model described in the paper was trained and evaluated on the SQuAD dataset, so there is some baseline/published result for how my developed model should perform. After reading the paper in some depth, I believe that I have a good foundational understanding of how R-Nets work, and can likely begin implementing the model.

Wider research context. Though this research paper solely focuses on Question-Answering and SQuAD/MS-MARCO, this paper also fits into the broader story of NLP research. For example, as discussed in the limitations section, the model's stagnation after training on a pseudo-dataset generated by the sequence-to-sequence question generation model in Zhou et al. (2017)'s *Neural Question Generation from Text: A Preliminary Study* reflects that the generated questions' quality needs to be improved. Furthermore, the paper itself brings some interesting self-attention techniques to the table, namely the gated attention-based recurrent network and self-matching mechanism described in the summary of contributions. The model evaluation results, combined with the success of variants of self-attention, makes these techniques appealing when devising future models. All in all, this paper both reveals shortcomings in the sequence-to-sequence question generation model, helps us build stronger question-answering models, possibly in different languages or different datasets (since this model was only trained and evaluated on SQuAD and MS-MARCO), and provides appealing techniques that can be applied in future language processing tasks.

2 Project description (1-2 pages)

Goal. The goal of this project is to develop a model that performs relatively well on SQuAD. To do this, we replicate the complex neural model described in the paper, including the contributions such as the gated attention-based recurrent network and self-matching mechanism. Furthermore, we wish to determine how these self-attention mechanisms compare to the given baseline RNN model for question-answering. Since the self-attention mechanism allows hidden states to consider previous hidden states, this model can record long-distance dependencies, and as a result have more complete answers to questions. The project is a direct application to the paper, even training on a similar dataset (SQuAD 2.0), so replicating and possibly making adjustments to improve the described will suffice. I chose the goal of replicating the model and observing if the model outperforms the baseline because I believe understanding self-attention is a crucial, effective technique, both for natural language processing and machine learning in general. Variants of self-attention have proven to be vastly successful in modern natural language processing models (ex. usage of multi-headed self-attention in a Transformer Encoder-Decoder block introduced in Vaswani et al. (2017)'s *Attention Is All You Need*), thus in theory, this model should outperform our Bidirectional Attention Flow baseline.

Task. The task is similar to that of the chosen paper, as it was also trained and evaluated on SQuAD. In other words, given a passage **P** and a question **Q** from SQuAD, we want our model to return an answer **A** within the span of the passage, or indicate that no such answer exists. For example, as listed in the paper, an example question-answer pair from SQuAD is as follows:

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901, which he (Morgan) had caused**. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

Answer: Panic of 1901.

Data. We will be using the Stanford Question Answering Dataset 2.0 (SQuAD 2.0), which has the following splits:

1. train (129,941 examples)
2. dev (6078 examples)
3. split (5915 examples)

In this dataset, passages are selected from the English Wikipedia (usually 100 150 words), questions are crowd-sourced, and each answer is either contained within the passage, or does not exist at all (not answerable). Also, each answerable SQuAD question has three answers, each from a different crowd worker.

Methods. The methods used will be the ones stated in the paper. We will replicate and utilize R-Net, an end-to-end neural network designed for reading comprehension and answering questions. Our model will consist of four parts, (1) a recurrent network encoder to build representations for questions and passages separately, (2) the gated matching layer as described in the contributions section, (3) the aggregating self-matching layer also described in the contributions section, and (4) a pointer-network based answer boundary prediction layer.

Baselines. In terms of baseline, we will use the one provided in the repository, based on Bidirectional Attention Flow. This baseline model uses learned character-level embeddings in addition to word embeddings . Furthermore, if we decide to make adjustments to the model described in the paper, we could also use the model's results as a baseline, as it was trained and evaluated on the SQuAD dataset.

Evaluation. To evaluate model performance, we use two well-defined, numerical, automatic evaluation metrics: Exact Match (EM) and F1 score. The Exact Match (EM) captures the percentage of the prediction that completely matches with one of the ground truth answers. For example, if the ground truth was "Panic of 1901" and the model responded with "1901", then the ground truth score would be 0. On the other hand, the F1 score is less strict - it is the harmonic mean of precision and recall. An example of an F1 score is with "1901" (model answer) and "Panic of 1901" (ground truth). Since the model answer is completely contained in the ground truth, it would have 100% precision, though only 33.3% recall, as it only included one out of the three words in the ground truth answer. Thus we have an F1 score of $2 \times 100 \times 33.3/(100 + 33.3) \approx 50.0\%$.

We will be comparing these evaluations against the Bidirectional Attention Flow baseline's performance, and also looking to stay close to or slightly exceed the paper's scores, as this is a re-implementation with possible adjustments.

References