An Effective Multi-Stage Approach For Question Answering

Anonymous EMNLP-IJCNLP submission

Abstract

Machine reading comprehension is a key part of natural language understanding. Due to wide range of applications, machine reading comprehension has attracted considerable interest from both commercial and academic entities. Recently, deep neural network based models have been able to achieve near human performance on certain types of reading comprehension datasets (Rajpurkar et al., 2016; Devlin et al., 2018; Hermann et al., 2015; Seo et al., 2016). Newer and much harder datasets have been proposed that require more sophisticated reasoning capabilities. One such dataset was proposed by Khashabi et al.(2018) called MultiRC. In this work, we propose a multi-stage approach which significantly improves the previous state of the art results on MultiRC (Khashabi et al., 2018).

1 Introduction

Deep neural network (DNN) models have lead to significantly improved performance, close to or even better than humans, on existing span based reading comprehension datasets like the Stanford Question Answering Dataset (SQuAD) and cloze style datasets like CNN/Dailymail (Rajpurkar et al., 2016; Devlin et al., 2018; Hermann et al., 2015; Seo et al., 2016). For span based QA datasets, the answer being a span of the paragraph limits the difficulty of the question as the answer is present verbatim in the paragraph and does not test the model's ability to reason or infer across multiple sentences in a deeper way (Khashabi et al., 2018). Similarly, for cloze style datasets the answer is typically a single missing word from a sentence, usually an entity referred to elsewhere in the passage. This again requires limited reasoning ability (Chen et al., 2016).

Recently, multiple datasets have been proposed to test a question-answering model's ability to reason and perform inference across multiple sentences in a passage. On some of these datasets, the performance of the current state of the art QA models significantly lags behind the human performance. Examples of such datasets include MultiRC (Khashabi et al., 2018), OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018), etc. In particular, MultiRC (Khashabi et al., 2018) was designed specifically to be more challenging than the existing datasets.

In this paper, we propose a multi-stage approach to tackle the problem of performing inference using information from multiple sentences. The aforementioned sentences are usually spread across the passage. We evaluate our approach on two datasets, MultiRC and OpenBookQA. Our proposed approach significantly improves upon the current state of the art results on MultiRC dataset and matches the current state of art performance on OpenBookQA achieved by non-ensemble models.

2 Approach

The general idea behind our approach is to broadly emulate a multi-stage process usually followed by humans in multiple choice reading comprehension tasks. Typically, to answer questions related to a given passage one starts off by quickly reading the whole passage at the start to get a general idea of the passage topic and gist of important concepts mentioned in the passage. After the initial reading, one starts considering each question individually. For a particular question, the text from the question and all its associated options are used to identify the most relevant portions of the passage. Subsequently, one can usually eliminate one or more of the answer choices as being obviously irrelevant using the information gathered from the previous step. Finally, a suitable answer is chosen from the remaining candidates based on the information inferred using the previously identified relevant portions.

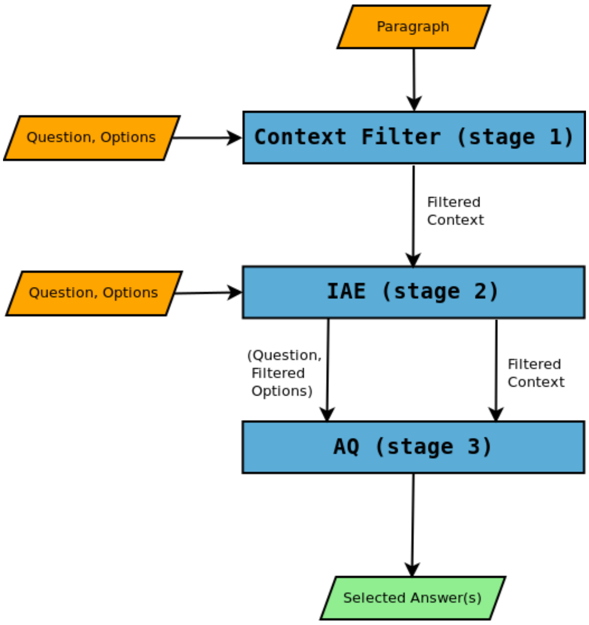

Figure 1: An illustration of the proposed approach

In the following subsections, we provide more details about each stage of our proposed approach:

2.1 Stage 1: Context Filtering (CF)

In the first stage, we train a model to classify, for each question, the portions of the given passage as either relevant or not. For this purpose, the question and all its candidate options are used. Portions classified as relevant for a particular question are extracted and used in the subsequent stages. We use sentences as the level of granularity for this extraction task. This helps us avoid truncating long passages arbitrarily which might lead to the elimination of relevant portions of the passage. Thus, in this stage, it is desirable that the model eliminates most of the irrelevant sentences and retains a reasonably high number of relevant sentences. We achieve this by choosing an operating point with a high precision and a reasonably good recall on the precision-recall curve for the model. This stage can be trained in a supervised manner by either using the information present in the dataset itself (if available) or using models trained on span based datasets, since models trained on span based datasets are effectively performing the same task as identifying relevant portions of the passage for each question.

2.2 Stage 2: Irrelevant Answer Elimination (IAE)

In the second stage, we train a dedicated model to classify every answer option using the corresponding question and the relevant portions of the passage extracted in the previous stage as either irrelevant or not. An option classified as irrelevant by the model is eliminated from further consideration. Here, the model's goal is to eliminate as many incorrect options as possible while avoiding elimination of possibly correct options. This stage reduces the number of candidate options for each question significantly. This leads to a significant reduction in the complexity of the task performed in the final stage.

2.3 Stage 3: Answering Questions (AQ)

After the previous two stages, we obtain a significant reduction in the complexity and size of the task. Thus, the level of reasoning abilities required of the final model to achieve high performance is significantly reduced. The filtered context and the remaining options are used to infer the correct answer(s).

3 Experiments

3.1 Experiment Settings

We use an open source PyTorch implementation of the pretrained BERT$_{\text{BASE}}$ made available by Hugging Face (2019) at each stage. We train each model for a maximum of 10 epochs. The models typically converge within 3 epochs.

Since we implement our approach by using BERT$_{\text{BASE}}$ to perform the tasks at each stage, we use the performance of BERT$_{\text{BASE}}$ on the full task as a baseline and provide that information for each experiment. We chose BERT$_{\text{BASE}}$ to perform the task at each stage due to its impressive performance across various NLP tasks (Devlin et al., 2018). We formulate the classification task as described in the original paper. We did not use BERT$_{\text{LARGE}}$ due to resource limitations.

We evaluate our proposed approach on two datasets: MultiRC (Khashabi et al., 2018) and OpenBookQA (Mihaylov et al., 2018). Furthermore, we also use RACE dataset (Lai et al., 2017) for pretraining our models while evaluating our approach on OpenBookQA. We provide more details in the following subsections.

Model	**EM**	**F1$_m$**	**F1$_a$**
BERT$_{\text{BASE}}$	17.73	68.46	66.11
Reading Strategies	22.60	71.50	69.20
Reading Strategies $^+$	21.80	73.10	70.50
CF + AQ *	25.29	69.77	67.98
CF + IAE + AQ *	**42.07**	**80.77**	**79.40**

Table 1: Model EM: exact match, $F1_m$: macro-average $F1$, $F1_a$: micro-average $F1$ scores (%) on MultiRC dev set, * indicates usage of BERT$_{\text{BASE}}$, $^+$ indicates ensemble model

3.2 MultiRC

MultiRC (Khashabi et al., 2018) is a multiple choice reading comprehension dataset. The questions included in this dataset were chosen to have answers which used information spread across multiple sentences. MultiRC does not have a fixed number of candidate answers for questions and there are usually multiple correct answer for a question (Khashabi et al., 2018). Furthermore, a majority of correct answers cannot be found verbatim in the context provided for answering the questions. These factors make it much more challenging than some of the existing multiple choice datasets, as evidenced by the large gap between the human and current state of the art deep learning model performance.

For the context filtering stage, we train a classification model on the MultiRC dataset which already includes information about the relevant sentences used for each question. We train the model using the sentences used information from the train set. Since the MultiRC dataset does not have a publicly available test dataset, we evaluate on the dev set. For the IAE stage, we consider each question-answer pair independently along with the filtered context obtained from the previous stage and classify the option as relevant or irrelevant. For the final stage, since the number of candidate options is not fixed and there can be a variable number of correct choices, we again consider each non eliminated answer independently for classification as correct or incorrect.

In table 1, we report the results of our experiments. To analyze the contribution of each component we also report the results obtained without the IAE stage. From the results achieved, it is evident that both stages contribute to the overall performance. Our approach significantly outperforms the existing state of the art (Sun et al., 2018).

3.3 OpenBookQA

OpenBookQA (Mihaylov et al., 2018) is a multiple choice reading comprehension dataset modeled to emulate the open book exam setting used to assess an individual's understanding of a particular topic. The questions in this dataset are based on elementary level science facts. In contrast to other reading comprehension datasets, OpenBookQA does not provide a separate context for each question. A common set of 1326 sentences containing elementary level science facts are provided which can be used to answer the questions. The questions are designed to test both common knowledge and the linguistic knowledge (Mihaylov et al., 2018) as opposed to just the linguistic knowledge tested by some other reading comprehension datasets like SQuAD (Rajpurkar et al., 2016).

For OpenBookQA dataset, apart from applying our approach in a way similar to how it was applied on the MultiRC dataset, we also apply an augmented version of our approach. The main difference between the non-augmented and augmented approach is that in the augmented approach, the models used in the last two stages are first pretrained on RACE (Lai et al., 2017) dataset and then finetuned on the target dataset. The reason for using the augmented approach is to acquire some common knowledge needed to answer questions (Mihaylov et al., 2018). RACE is one of the largest existing multiple choice dataset and existing work has (Pan et al., 2019) shown that pretraining on RACE leads to an improvement on a model's performance on OpenBookQA.

As mentioned earlier, OpenBookQA provides a common corpus of 1326 sentences for all the questions. Both the augmented and non-augmented approaches share the first stage. For OpenbookQA the first stage is split into two steps. First step is to pick out top 15 sentences from the common corpus for each question using TF-IDF and cosine similarity. We compute cosine similarity between the text obtained by concatenating question text and the corresponding option texts and all the sentences in the common corpus. For the second step, we use the same model that was used for the MultiRC dataset for context filtering, since OpenBookQA does not provide sentences used information in its training dataset. Furthermore, in contrast to MultiRC, we do not consider each of the sentence selected in the first step independently. We use the model to obtain relevance probabilities

Model	Accuracy
$\mathrm{BERT}_{\mathrm{BASE}}$	52.20
$\mathrm{BERT}_{\mathrm{LARGE}}$ (AllenAI, 2018)	60.40
$\mathrm{BERT}_{\mathrm{BASE}}$ **[CF + AQ]**	52.20
$\mathrm{BERT}_{\mathrm{BASE}}$ **[CF + IAE + AQ]**	**62.20**

Table 2: Non-augmented Model accuracies (%) on OpenBookQA test set

for each of the selected sentences for each question and then pick 6 sentences with highest relevance probability for each question.

We follow the same procedure for the second and third stage as described in the MultiRC section for OpenBookQA with two changes. The first change pertains to the fact that in OpenBookQA dataset each question has 4 candidate options with only 1 correct answer. This allows us to consider all 4 options at the same time while answering a question. Keeping this fact in mind, we train the models in the second and third stage to output probabilities for each option. In the second stage, the options with top 2 probabilities are eliminated (correspond to the 2 most irrelevant candidates) and in the third stage, the option with highest probability is picked as the final answer. The second change which applies only to the augmented approach is pretraining the models described earlier.

We report the performance of the non-augmented approach in table 2 and the augmented approach in table 3. Our non-augmented approach significantly outperforms the $\mathrm{BERT}_{\mathrm{BASE}}$ baseline and it also outperforms the $\mathrm{BERT}_{\mathrm{LARGE}}$ trained only on the OpenBookQA dataset (AllenAI, 2018). In table 2, we compare the performance of the augmented approach with other $\mathrm{BERT}_{\mathrm{LARGE}}$ based approaches, all of which include pretraining on other datasets in their training regimen. The augmented approach achieves the same result as the current non-ensemble state of the art approach reported by Pan et al (2019).

4 Discussion

Our approach logically divides the main task into three sub-tasks with much lower complexity. This allows each of the sub-tasks to be performed more effectively leading to the increase in overall performance.

The proposed multi-stage approach has some inherent advantages when compared to an end-to-end model. The output of each of the stages is highly interpretable and can be analyzed to improve the performance at each stage. The model used at each stage can be replaced with a model of more appropriate complexity depending on the dataset.

Model	Accuracy
$\mathrm{BERT}_{\mathrm{LARGE}}$(MS AI, 2019)	63.80
$\mathrm{BERT}_{\mathrm{LARGE}}$(Pan et al., 2019)	68.00 *
$\mathrm{BERT}_{\mathrm{LARGE}}$(Pan et al., 2019)	**69.60** +
$\mathrm{BERT}_{\mathrm{BASE}}$ **[CF + AQ]**	56.60
$\mathrm{BERT}_{\mathrm{BASE}}$ **[CF + IAE + AQ]**	68.00 *

Table 3: Augmented Model accuracies (%) on OpenBookQA test set, * indicates current non-ensemble state of the art, + indicates current state of the art ensemble model

The proposed multi-stage approach has lower resource requirements, particularly when making use of resource intensive models like BERT. Training such models as an end to end model or ensembling requires significant GPU memory.

For the MultiRC dataset, our model uses a comparable number of parameters to the previous state of the art (Sun et al., 2018) (which is based on OpenAI's GPT (Radford and Sutskever, 2018)), but significantly outperforms it on all metrics.

For the OpenbookQA dataset, our approach achieves comparable performance to the current state of the art which uses an ensemble. Since our model does not use ensembling and is based on $\mathrm{BERT}_{\mathrm{BASE}}$ rather than $\mathrm{BERT}_{\mathrm{LARGE}}$, it uses significantly fewer number of parameters.

5 Conclusion and Future Work

We proposed a multi-stage approach modeled on the process typically followed by humans to perform multiple choice reading comprehension task. Our approach significantly outperformed the previous state of the art on the MultiRC dataset and performed impressively on the OpenBookQA dataset indicating its effectiveness in handling challenging datasets requiring advanced reasoning capabilities.

In the future, we would like to develop an end to end architecture that can reasonably model the same multi-step process. We would also like to explore how the performance changes with the usage of more specialized model for different stages.

References

OpenBookQA Leaderboard Submission AllenAI. 2018. Bert-large (no kb). https://leaderboard.allenai.org/open_book_qa/submission/bga1grnl77gnifvufd00.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hugging Face. 2019. pytorch pretrained bert. https://github.com/huggingface/pytorch-pretrained-BERT.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. *CoRR*, abs/1809.02789.

OpenBookQA Leaderboard Submission MS AI. 2019. Bert multi-task (single model). https://leaderboard.allenai.org/open_book_qa/submission/biu0a2bloo2d3dptvh10.

Xiaoman Pan, Kai Sun, Dian Yu, Heng Ji, and Dong Yu. 2019. Improving question answering with external knowledge. *CoRR*, abs/1902.00993.

Alec Radford and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies.