Extensive and Intensive: A BERT-based machine reading comprehension model with two reading strategies

Guoqi Zhang*, Chunlong Yao

Dalian Polytechnic University, Dalian 116034, China
**Corresponding author: 18108110005121@xy.dlpu.edu.cn*

Abstract: *Enabling machines to read, process, and understand natural language documents is a coveted goal of artificial intelligence. However, this task is extremely challenging, and most existing models lack the ability to perform complex reasoning. Considering that humans often read documents roughly first when understanding a problem, this paper proposes a new model that attempts to mimic the reasoning process of human readers. Our model performs a extensive read and a intensive read of the document separately, and then combines the information obtained from both reading methods to finally find a satisfactory answer. Finally, by experimenting within RACE dataset and comparing with the baseline model BERT, the feasibility and effectiveness of our proposed model can be illustrated.*

Keywords: *machine reading comprehension, neural network, attention mechanism.*

1. Introduction

Machine reading comprehension is a task that enables a computer to understand the semantics of a passage and answer related questions [1] [2] [3]. To accomplish the task, the model needs to deeply analyse the semantics of the text and the connections between the text and the question, and then provide an accurate answer based on the content in the text. In recent years, this task is receiving increasing attention from both industry and academic research communities.

In this paper, we focus on the multiple choice MRC task, which requires selecting the correct choice from a set of candidate answers based on a given passage and question. In this type of task, the correct answer to the question usually does not appear in the original text. Therefore, it requires a deeper understanding of the document than the traditional extractive reading comprehension.

The attention mechanism was first proposed in the field of computer vision. It mimics the attention mechanism of human vision, which focuses on only the most important parts of the complex information and ignores the other unimportant information. This mechanism was later also widely applied to various subtasks in the field of natural language processing, and obtained remarkable results[4].

For reading comprehension tasks, each word has a different importance in the sentence. Therefore, each word contributes differently to the final answer. The machine can use the interaction information between the passage and the question to infer which parts of the document are more important for answering the question. Some existing machine reading comprehension models [5] [6] [7] use an attention mechanism to compute the correlation matrix between each word in the question and each sentence in the document, and then mix the question information with the document information.

Since recent years, pre-trained language models such as BERT [8] have brought great advances to the field of machine reading comprehension. With a simple fine-tuning, SOTA results can be obtained on the English dataset SquAD [2] and surpass human performance. However, on those datasets that require further reasoning to answer, BERT is still far from human performance.We argue that the pre-trained language model lacks the ability to perform complex reasoning and thus cannot use information from paragraphs, questions, and choices for deep reasoning.

Considering the complexity of the questions, humans tend to answer the questions by first reading the passage, questions and choices roughly, followed by repeated readings of the text carefully to better understand the question. Based on this idea, we propose a new model to perform extensive and intensive reading of documents separately. The extensive reading module obtains the semantic information of the passage, questions and choices, while the intensive reading module uses the attention mechanism to perform multiple interactions between the passage, questions and choices to explore the relationship between them at a deeper level, thus simulating the repetitive reading behavior of humans when performing reading comprehension.

Specifically, our main contributions in this paper are as follows.

1) We propose a new model for multiple choice reading comprehension tasks, which attempts to imitate the reasoning process of human readers by performing a extensive and a intensive reading of the document separately, and then combining the information obtained from both reading strategies to finally find a satisfactory answer.

2) We conducted detailed experiments and comparisons of the proposed model with the baseline model. According to the experimental results, our model obtains significant improvements compared to BERT.

2. Model

We focus on the multiple choice MRC task, which can be described as follows: given a passage P, a question Q, and a set A of n answer candidates, where A={ , ,…, }. the task requires the model to select the correct answer from A .

Our proposed model consists of two modules: the extensive module and the intensive module, where the extensive module makes a rough judgment of the answer to the question, and the intensive module combines the information obtained from the extensive module with careful verification to arrive at the final answer.

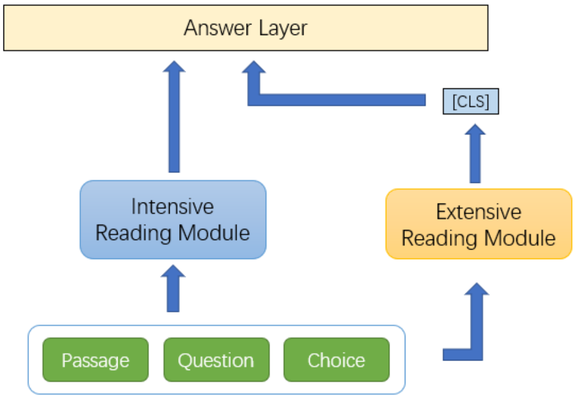

Figure 1: The framework of our model

2.1 Extensive reading module

The extensive reading module is the basis for the model to answer the question correctly, thus a powerful architecture is required.

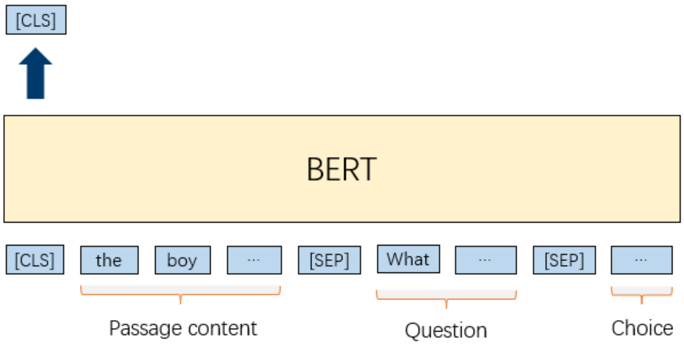

Figure 2: Extensive reading module

BERT [8] has become one of the most successful language representation models on various NLP tasks, including SQuAD. BERT is a pre-trained language model that solves the problem of multiple meanings of words compared to statically encoded word vectors, such as Word2vec [9] and Glove [10]. Due to its large number of parameters and network depth, BERT has a strong expressive power.

As shown in Figure 2, we use the pre-trained model BERT as the extensive reading module to encode the input. We splice the passage P, the question Q and the i-th choice A into a new sequence ([CLS] , P_1, ... , P_m , [SEP] , Q_1, ... , Q_n , [SEP] , A_1... , A_n), where m, n, k are the sequence lengths of P, Q, A. The spliced new sequence will be used as the input to the encoder BERT. Since the output vector [CLS] at the first position of BERT incorporates the semantic information of each word in the input text, we take the [CLS] vector as the representation vector after extensive reading of the question, passage and choices.

2.2 Intensive reading module

In our proposed model, the intensive reading module carefully verifies the information obtained from the extensive reading module and comes up with the final answer.It consists of two parts: Embedding layer and Matching layer.

2.2.1 Embedding layer

Machine reading comprehension requires answering question Q based on a given passage P and choices A. BERT handles this task by encoding P, Q, and A into a sequence as input. Although BERT has shown excellent performance, some recent studies [11] have argued that this approach may cause the semantics of one section to be influenced by the words in another section and it is difficult to accurately distinguish between passages, questions, and choices (< P, Q, A>) in the input sequence.

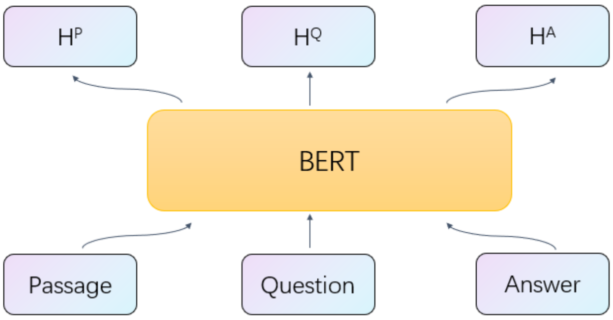

Figure 3: Embedding layer of intensive reading module

To solve the problem caused by attentional unfocusing, as shown in Figure 3, we input P, Q and A into BERT separately. The inputs are encoded as follows:

$$\begin{cases} H^P = Encode(P) \\ H^Q = Encode(Q) \\ H^A = Encode(A) \end{cases} \tag{1}$$

Where we use BERT as Encoder and Encode(·) returns the last layer of BERT output. , and are sequence representation of the passage, question and choice, respectively. In this way, the overall semantic attention of P, Q, and A will only be distributed on their own words and not distracted by other words. This approach makes it easier for the model to capture the semantic core words of the text.

2.2.2 Matching layer

To empower the model with stronger inference, we exploited the attention mechanism to fully exploit the information in the {P,Q,A} triplet. The attention weight matrix W_{PA} between passages H^P and choices H^A is calculated as follows:

$$\begin{cases} W_{PA} = \mathrm{Softmax}(\mathrm{H}^P W_1 (H^A)^T) \\ M^{PA} = WH^A \end{cases} \tag{2}$$

Where the representation of passage is $H^P \in \mathbb{R}^{|P| \times l}$, the representation of choice is $H^A \in \mathbb{R}^{|A| \times l}$, the trainable parameters $W_1 \in \mathbb{R}^{l \times l}$ and the attention weight matrix $W_{PA} \in \mathbb{R}^{|P| \times |A|}$ .So we get the new representation M^{PA} of the passage after combining the information of the choice.

In the same method, we can get the new representation after combining the information of the passage and the question:

$$\begin{cases} W_{PQ} = \mathrm{Softmax}(M^{PA} W_2 (H^Q)^T) \\ M^{PAQ} = W_{PQ} H^Q \end{cases} \tag{3}$$

Then, to get the final representation of each option, we use the max-pooling operation in the row direction.

$$\begin{cases} C^P = Pooling(M^{PAQ}) \\ C^{\mathrm{Q}} = Pooling(H^Q) \\ C^{\mathrm{A}} = Pooling(H^A) \end{cases} \tag{4}$$

So we can get the output C of the Matching layer:

$$C = concatenate(C^P; C^{\mathrm{Q}}; C^{\mathrm{A}}) \tag{5}$$

2.3 Answer Layer

Finally, the final answer is calculated in the Answer layer. The [CLS] vector obtained from the extensive reading module is stitched with the output of the intensive reading module as the input of the fully connected layer. s_i is the output of the fully connected layer corresponding to the i-th choice. Then the confidence score p_i of the i-th choice is calculated by mapping with the softmax function:

$$p_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)} \tag{6}$$

Finally, the choice with the highest confidence score is selected as the answer:

$$ans = \arg\max_i (p_i) \tag{7}$$

3. Experiments

3.1 Dataset and Setup

In this paper, the proposed model was evaluated on the RACE dataset [12].

RACE contains a total of 27,933 passages and 97,687 questions and is considered as one of the largest and most difficult datasets in multiple choice reading comprehension. Compared to CNN&Dailymail and SquAD, RACE is more focused on inference ability. Its correct answers are not necessarily directly reflected in the text, but can only be selected by analyzing the clues in the text and reasoning based on the context by understanding the text in depth from the semantic level.

In comparison, RACE has a longer average passage length and focuses more on the model's ability to make inferences based on understanding the original text. Therefore, we test our proposed model on RACE.

3.2 Implementation

In the experiments, the passage length of the BERT input was truncated to 380 words, and the question and choice lengths were truncated to 60 words. A dropout rate of 0.1 is applied to every BERT layer. In addition, for model training, the optimizer uses Adam, the learning rate is set to 8e-5, and the size of batch is 16.

3.3 Baseline Model

We divide the baseline models into two categories.

The first category is those that do not use pre-trained language models, such as Stanford Attentive Reader [13], Gated-Attention Reader [14], Co-Matching [15] and BiAttention [16], which mostly use static word vectors like Word2vec or Glove. Compared with pre-trained language models, these models do not combine contextual semantics well and therefore cannot solve the problem of multiple meanings of words.

The other category is pre-trained language models, such as BERT. The emergence of BERT has brought great progress to the field of machine reading comprehension. Particularly good results can be achieved by simply adding a simple network after BERT.

3.4 Main Results

We evaluated the proposed model in this paper on the RACE dataset and compared it with other models. We report the experimental results on RACE and its two subtasks: RACE-M and RACE-H. As shown in Table 1, the accuracy of our model is improved on the RACE dataset compared to BERT_BASE and BERT_LARGE. The experimental results show the superiority of our model in terms of accuracy.

Table 1: Experiment results on RACE dataset

Models	*Accuracy(dev)*		
	RACE-M	RACE-H	RACE
Human Ceiling Performance	95.4	94.2	94.5
Stanford Attentive Reader [13]	44.2	43.0	43.3
Gated-Attention Reader [14]	43.7	44.2	44.1
Co-Matching [15]	55.8	48.2	50.4
BiAttention [16]	57.7	47.4	50.4
BERT_BASE(our implementation)	70.9	62.1	64.8
Our Model(based on BERT_BASE)	**72.1**	**63.4**	**66.1**
BERT_LARGE(our implementation)	76.3	69.8	71.7
Our Model(based on BERT_ LARGE)	**77.5**	**70.6**	**72.6**

4. Conclusion

1) The accuracy of BERT is much higher than that of the model without pre-training. Moreover, our model improves in accuracy compared to BERT, indicating that our method can improve performance on stronger baseline models.

2) Although the BERT model performed well, its performance on the RACE dataset was still below the best human level. This indicates that there is still a gap between BERT and human reading comprehension, implying that there is much room for model improvement.

3) Our model achieves higher accuracy than BERT. The reason is that the intensive reading module of our model uses the attention mechanism to interact multiple times with articles, questions, and choices to explore the relationship between them at a deeper level. This illustrates the effectiveness of our proposed model and validates the importance of enhancing complex reasoning to improve machine reading performance.

4) The improvement of our model on RACE-M is more obvious compared to RACE-H. This is because RACE-H emphasizes more on the inference ability of the model, so our model has more advantages on RACE-H.

5. Summary

The presence of questions in machine reading comprehension tasks that require deep reasoning to answer emphasizes the importance of deep reasoning capabilities in MRC models. In recent years, the emergence of pre-trained language models such as BERT has brought great advances to the field of machine reading comprehension, and particularly good results can be achieved by simply adding a simple network. However, pre-trained language models lack the ability to perform complex inference using information from passages, questions, and selections.Inspired by human reading comprehension experience, we propose a machine reading comprehension model that integrates extensive and intensive reading. It uses the pre-trained language model BERT as an encoder, and exploits the attention mechanism for multiple interactions between passages, questions, and choices to explore their relationships at a deep level. We evaluated the proposed model on the MRC dataset RACE and compared it with the baseline models. The experimental results demonstrate the improved accuracy of our model over BERT, which shows the effectiveness of our proposed model and demonstrates that deep inference capability has an important impact on multiple choice MRC performance.

References

[1] Hermann K M, Kočiský T, Grefenstette E, et al. Teaching machines to read and comprehend [J]. arXiv preprint arXiv:1506.03340, 2015.
[2] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text [J]. arXiv preprint arXiv:1606.05250, 2016.
[3] Nguyen T, Rosenberg M, Song X, et al. MS MARCO: A human generated machine reading comprehension dataset [C]// CoCo@ NIPS. 2016.
[4] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.
[5] Yu A W, Dohan D, Luong M T, et al. Qanet: Combining local convolution with global self-attention for reading comprehension [J]. arXiv preprint arXiv:1804.09541, 2018.
[6] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension [J]. arXiv preprint arXiv:1607.04423, 2016.
[7] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension [J]. arXiv preprint arXiv:1611.01603, 2016.
[8] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.
[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J]. arXiv preprint arXiv:1310.4546, 2013.
[10] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
[11] Chen Z, Wu K. ForceReader: a BERT-based Interactive Machine Reading Comprehension Model with Attention Separation [C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 2676-2686.
[12] Lai G, Xie Q, Liu H, et al. Race: Large-scale reading comprehension dataset from examinations [J]. arXiv preprint arXiv:1704.04683, 2017.
[13] Chen D, Bolton J, Manning C D. A thorough examination of the cnn/daily mail reading comprehension task [J]. arXiv preprint arXiv:1606.02858, 2016.
[14] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension [J]. arXiv preprint arXiv:1606.01549, 2016.
[15] Wang S, Yu M, Chang S, et al. A co-matching model for multi-choice reading comprehension[J]. arXiv preprint arXiv:1806.04068, 2018.
[16] Tay Y, Tuan L A, Hui S C. Multi-range reasoning for machine comprehension [J]. arXiv preprint arXiv:1803.09074, 2018.