ADAPTIVE BI-DIRECTIONAL ATTENTION: EXPLORING MULTI-GRANULARITY REPRESENTATIONS FOR MACHINE READING COMPREHENSION

Nuo Chen\textsuperscript{1}, Fenglin Liu\textsuperscript{1}, Chenyu You\textsuperscript{2}, Peilin Zhou\textsuperscript{1}, Yuexian Zou\textsuperscript{1,3,*}

\textsuperscript{1}ADSPLAB, School of ECE, Peking University, Shenzhen, China
\textsuperscript{2}Department of Electrical Engineering, Yale University, CT, USA
\textsuperscript{3}Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Recently, the attention-enhanced multi-layer encoder, such as Transformer, has been extensively studied in Machine Reading Comprehension (MRC). To predict the answer, it is common practice to employ a predictor to draw information only from the final encoder layer which generates the coarse-grained representations of the source sequences, i.e., passage and question. Previous studies have shown that the representation of source sequence becomes more coarse-grained from fine-grained as the encoding layer increases. It is generally believed that with the growing number of layers in deep neural networks, the encoding process will gather relevant information for each location increasingly, resulting in more coarse-grained representations, which adds the likelihood of similarity to other locations (referring to homogeneity). Such a phenomenon will mislead the model to make wrong judgments so as to degrade the performance. To this end, we propose an approach called Adaptive Bidirectional Attention, which adaptively exploits the source representations of different levels to the predictor. Experimental results on the benchmark dataset, SQuAD 2.0 demonstrate the effectiveness of our approach, and the results are better than the previous state-of-the-art model by 2.5\% EM and 2.3\% F1 scores.

Index Terms— Natural Language Processing, Question Answering, Machine Reading Comprehension, Multi-Granularity Representation

1. INTRODUCTION

Machine reading comprehension (MRC) \cite{1, 2, 3} is a long-standing task that aims to teach the machine how to read and comprehend a given source sequence, i.e., passage/paragraph, then answer its corresponding given questions automatically.

It has large amounts of real application scenarios such as question answering and dialog systems.

Recently, the attention-enhanced multi-layer encoder, e.g., Transformer \cite{4}, ALBERT \cite{5}, RoBERTa \cite{6}, and XLNet \cite{7}, which is based solely on attention mechanisms \cite{8, 9} and eliminates recurrence entirely, has been proposed and has established the state-of-the-art in multiple challenging MRC datasets \cite{10, 11, 12}. However, under the multi-layer deep learning setting, the representations of source sequence will become more coarse-grained from fine-grained with the growing number of encoder layers. Following in \cite{13}, Figure 1 illustrates that as the multi-layer encoder processes the source sequences, each input word will gradually gather related information increasingly, resulting in more coarse-grained representations, which adds the likelihood of similarity to other locations (referring to homogeneity). Such a phenomenon will mislead the model to make wrong judgments so as to degrade the performance. To this end, we propose a novel approach called Adaptive Bidirectional Attention, which adaptively exploits the source representations of different levels to the predictor. Experimental results on the benchmark dataset, SQuAD 2.0 demonstrate the effectiveness of our approach, and the results are better than the previous state-of-the-art model by 2.5\% EM and 2.3\% F1 scores.

In this paper, we argue that it would be better if the predictor could exploit representations of different granularity from the encoder, providing different views of the source sequences, such that the expressive power of the model could be fully utilized. As a consequence, we propose a novel method called Adaptive Bidirectional Attention to dynamically provide multi-granularity source representations for...
better predicting the correct answer. As shown in Figure 3, the proposed approach builds connections with each encoder layer, so that the MRC model not only can exploit the coarse-grained representation, of the source sequence, which is instrumental in language modeling, but also has the ability to exploit the fine-grained representations of the source sequence, which help predict more precise answers. Hence, the answer predictor is encouraged to use source representations of different granularity, exploiting the expressive power of the model. Our experimental results show that Adaptive Bidirectional Attention has substantial gains in predicting answers correctly.

2. MODEL

2.1. Problem Formulation

In this paper, the reading comprehension task is defined as follows: Given a passage/paragraph with \( l \) words \( P = \{p_1, p_2, p_3, ..., p_l\} \) and a question with \( m \) words \( Q = \{q_1, q_2, q_3, ..., q_m\} \) as inputs, the goal is to find an answer span \( A \) in \( P \). If the question is answerable, the answer span \( A \) exists in \( P \) as a sequential text string; Otherwise, \( A \) is set to an empty string indicating that the question is unanswerable. Formally, the answer is formulated as \( A = \{p_{\text{begin}}, ..., p_{\text{end}}\} \). In the case of unanswerable questions, \( A \) denotes the last token of the passage.

2.2. Model Overview

Generally, typical MRC models contain four major parts: an embedding layer, several encoding layers, an attention layer and an output layer, as shown in Figure 2. Embedding Layer is responsible for encoding each token in the passage and question into a fixed-length vector. In the Encoding Layers, the models move forward to extract contextual cues from the surrounding words and sentences to form higher level granularity representations with some efficient methods such as transformers, BiLSTMs [14] and CNNs, to name a few. Next, Attention Layer plays a key role in MRC development, which has the ability of extracting relationship between words at various positions of context and question, and even focuses on the most important part of the interaction. Finally, the Output Layer decodes to the probability that each position is the beginning or end of the answer span so as to predict an answer.

2.3. Adaptive Bidirectional Attention

As illustrated in Section 1, we propose Adaptive Bidirectional Attention to encourage the answer predictor (output layer) to take full advantage of the expressive power of the multi-layer encoders through exploring and exploiting the multi-granularity representations of source sequences.

To simplify notation, we first define the output of Embedding layer are \( E_p \) for passage and \( E_q \) for question, separately. Then the output of Encoding layers are formulated as \( C^1_p, C^2_p, ..., C^m_p \), \( C^1_q, C^2_q, ..., C^m_q \), sequentially. Next, the output of the Attention layer is \( A \). After multiple layers extracting different levels of presentations of each word, we conduct Adaptive Bidirectional Attention from question to passage and passage to question based on generated representations to take advantage of all layers of representation reasonably, which is
Fig. 3. An overview of Adaptive Bidirectional Attention. ⊙ is the element-wise multiplication.

showed in Figure 3. We define a novel concept – 'history of semantic', which denotes multi-granularity representations extracted by the model before the current layer. Thus history-of-semantic vectors can be defined as:

\[
H_{OS}^p = [E_p; C^1_p; C^2_p; \ldots; C^n_p; A] \quad (1)
\]

\[
H_{OS}^q = [E_q; C^1_q; C^2_q; \ldots; C^n_q; A] \quad (2)
\]

Adaptive. In order to obtain multi-level representations connection between the encoder layers. We have designed the following function:

\[
\lambda^p \odot H_{OS}^T \quad \lambda^q \odot H_{OS}^T = H_{OS}^T \quad (3)
\]

where the matrix \( \lambda \) is trainable. Notice that at the beginning of training, in order to retain the original semantics of each layer, we will initialize the first column of this matrix to all 1, the remaining columns are all 0.

Bidirectional Attention. In this component, we are aiming to dynamically compute attention of the embedding vectors from previous layers each time, as well as the multi-granularity representations generated from the different layers are allowed to flow through to the downstream output layer. Like most high performing models, such as [15, 16], we construct passage-to-question and question-to-passage attention, respectively. The attention function in [16] is used to compute the similarity score between passages and questions. First, we calculate the similarities between each pair words in passages and questions, and render a similarity matrix \( H \in \mathbb{R}^{n \times m} \). \( H \) is computed as:

\[
H = \text{dropout}(f_{\text{attention}}[H_{OS}^p, H_{OS}^q]) \quad (4)
\]

After that, we use the strategy of [2] to normalize each row of \( H \) by applying the softmax function, and then get a matrix \( \hat{H} \). Then the passage-to-question attention is computed as \( M = \hat{H} \cdot (H_{OS}^q)^T \). Following DCN [17], we compute the column normalized matrix \( \overline{H} \) of \( H \) by softmax function, and the question-to-passage attention is \( S = \overline{H} \cdot \overline{H}^T \cdot (H_{OS}^p)^T \).

At last, our method use a simple concatenation as following to get the final representation, which shows good performances in our experiments:

\[
I = [\hat{H}_{OS}^p; M; \hat{H}_{OS}^q \odot M; \overline{H}_{OS}^p \odot S] \quad \text{[21]}
\]

3. EXPERIMENTS

In this section, we conduct a series of experiments to study the performance of our method. We will primarily benchmark our method on the SQuAD 2.0 dataset [21].

3.1. Dataset And Experimental Settings

3.1.1. Dataset

We evaluate our method on SQuAD 2.0 dataset [21], a new MRC dataset which is a combination of SQuAD 1.0 and additional unanswerable question-answer pairs. Specifically, the number of questions that can be answered is about 100K;

<table>
<thead>
<tr>
<th>Base Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++ [18]</td>
<td>65.6</td>
<td>68.7</td>
</tr>
<tr>
<td>+ Adaptive Bidirectional Attention</td>
<td>67.2</td>
<td>70.8</td>
</tr>
<tr>
<td>QANet [2]</td>
<td>65.4</td>
<td>67.2</td>
</tr>
<tr>
<td>+ Adaptive Bidirectional Attention</td>
<td>66.9</td>
<td>69.1</td>
</tr>
<tr>
<td>SAN [3]</td>
<td>68.6</td>
<td>71.4</td>
</tr>
<tr>
<td>+ Adaptive Bidirectional Attention</td>
<td>70.1</td>
<td>73.7</td>
</tr>
<tr>
<td>SDNet [19]</td>
<td>76.7</td>
<td>79.8</td>
</tr>
<tr>
<td>+ Adaptive Bidirectional Attention</td>
<td>78.5</td>
<td>81.7</td>
</tr>
<tr>
<td>SGNet [20]</td>
<td>85.1</td>
<td>87.9</td>
</tr>
<tr>
<td>+ Adaptive Bidirectional Attention</td>
<td>87.6</td>
<td>90.2</td>
</tr>
</tbody>
</table>
Fig. 4. Visualization of Syntax-guided Attention (left) and Adaptive Bidirectional Attention (right). The example is the same as the showed in Figure 3. For Syntax-guided attention, *who* focuses on *Michael* and 23, and there is a great similarity between 23 and 24, which would mislead the model to make the wrong judgment. In contrast, this phenomenon does not happen in our Adaptive Bidirectional attention.

while the unanswerable questions are around 53K. For the SQuAD 2.0 challenge, MRC models must answer questions not only where possible, but also when the paragraph does not support any answers.

3.1.2. Baseline Models

To prove the generality and effectiveness of our proposed method, we choose several previous state-of-the-art QA models. In a nutshell, the architectures of QA models are briefly summarized below. BiDAF++ [18] is a very representative work for machine comprehension, which proposes a Bi-Directional Attention Flow to obtain query-aware contextual representations. QANet [2], an atypical but very effective QA model, which completely contains no recurrent networks and consists of stacked encoder blocks with convolution networks and self-attention. SAN [3] is an alternative multi-step reasoning neural network for MRC, simple yet robust. SDNet[19] and SGNet [20] both can be viewed as BERT-based models, which take BERT as their backbone encoders.

3.2. Results

In this study, we use F1 and EM as evaluation metrics. F1 measures the part of the overlapping mark between the predicted answer and the ground-truth answer, and if the prediction is exactly the same as the ground truth, the exact match (EM) score is 1, otherwise it is 0. Specifically, we add Adaptive Bidirectional Attention on some end-to-end MRC models to compare with their initial version (i.e. base models in Table 1) on SQuAD 2.0. As can be seen from Table 1, after adding Adaptive Bidirectional Attention, the performance of these models could be improved to varying degrees. As for SGNet[20], our method even improves EM and F1 scores by 2.5% and 2.3%, separately. This also proves the versatility of this attention mechanism.

4. VISUALIZATION

In this component, to have an insight that how Adaptive Bidirectional Attention works, we draw attention distributions of the Syntax-guided Attention of SGNet [20] and our proposed Adaptive Bidirectional attention, as shown in Figure 4. The visualization verifies that benefiting from Adaptive Bidirectional Attention, our method is effective at distinguishing and utilizing different layers of presentation of each word in the context and the question, guiding the downstream layer to collect more relevant semantics to make predictions so as to help MRC models in predicting the answer more accurately.

5. CONCLUSION

In this paper, we propose a novel attention mechanism Adaptive Bidirectional Attention to explore and exploit multi-granularity representations of source sequences for machine reading comprehension. In particular, our method can adaptively exploit the source representations of different levels to the predictor. More concretely, Adaptive Bidirectional Attention is used to guide attention learning and reasonably leverages source representations of different levels for question answering by dynamically capturing the connection between all layers of representations of each word. The experimental results show the effectiveness of our approach on the public benchmark dataset, SQuAD 2.0. Future work can include an extension of employing Adaptive Bidirectional Attention to other natural language processing tasks.
6. REFERENCES


