Exploring and Exploiting Multi-Granularity Representations for Machine Reading Comprehension

Nuo Chen
ADSPLAB, School of ECE
Peking University, Shenzhen, China

Chenyu You
Department of Electrical Engineering,
Yale University, CT, USA

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. The analysis shows 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 phenomenon will mislead the model to make wrong judgement and degrade the performance. 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. To this end, we propose a novel approach called Adaptive Bidirectional Attention-Capsule Network (ABA-Net), which adaptively exploits the source representations of different levels to the predictor. Furthermore, due to the better representations are at the core for boosting MRC performance, the capsule network and self-attention module are carefully designed as the building blocks of our encoders, which provides the capability to explore the local and global representations, respectively. Experimental results on three benchmark datasets, i.e., SQuAD 1.0, SQuAD 2.0 and COQA, demonstrate the effectiveness of our approach. In particular, we set the new state-of-the-art performance on the SQuAD 1.0 dataset.

1 Introduction

Machine reading comprehension (MRC) (Chen et al., 2021c; You et al., 2021b; You et al., 2020b) is a long-standing task that aims to teach 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. In recent years, the deep neural networks, including the recurrent neural networks (RNNs) and convolutional neural networks (CNNs) (Zhu et al., 2018; Zhang, 2019) have been introduced to extract representations of source sequence efficiently and make great progress in MRC. However, for RNNs-based models, the undeniable deficiency is that due to the sequential dependency, its learning ability of long sentences is insufficient and it is unsuitable for parallel training (Vaswani et al., 2017a); For CNNs-based models, in which the pooling operation incorrectly discards spatial information and does not consider the hierarchical relationship between extracted features (Liu et al., 2019b).

Recently, the attention-enhanced multi-layer encoder, e.g., Transformer (Vaswani et al., 2017a), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019c) and XLNet (Yang et al., 2019b), which is based solely on attention mechanisms (Bahdanau et al., 2015) and eliminates recurrence entirely, has been proposed and has established the state-of-the-art in multiple challenging MRC datasets (Rajpurkar et al., 2016; Joshi et al., 2017; Reddy et al., 2019). However, under the multi-layer deep learning setting, the representations of source sequence will become more coarse-grained from fine-grained with more...
Figure 1: An visualization example of a Question-Answer pair for a passage.

encoder layers. Following in (Huang et al., 2018), Figure 1 illustrates that as the multi-layer encoder processes the source sequences, each input word will gradually gather more and more related information as more layers are used, resulting in more abstract representations, i.e., from fine-grained to coarse-grained representations that are more likely to be similar with other positions (Homogeneous Phenomenon). For those representations output by different encoder layers, the common practice for the current answer predictor is to draw information (coarse-grained representations) only from the final encoder layer. However, it should be intuitive that coarse-grained representations are good at expressing the overall meaning of the source sequence, but are less precise in the finer details. If we also exploit fine-grained representations for the answer predictor, it will help the predictor find source information more precisely and give answers that are more accurate. As we observe in Figure 1, due to the baseline model only focus on the coarse-grained representations, it gives an incorrect answer. In contrast, our method exploits fine-grained representations for the answer predictor, it will help the predictor find source information more precisely and give answers that are more accurate. As we observe in Figure 1, due to the baseline model only focus on the coarse-grained representations, it gives an incorrect answer. In contrast, our method exploits fine-grained representations for the answer predictor, it will help the predictor find source information more precisely and give answers that are more accurate. As we observe in Figure 1, due to the baseline model only focus on the coarse-grained representations, it gives an incorrect answer.

In this paper, we propose a novel framework called Adaptive Bidirectional Attention-Capsule Network (ABA-Net) to dynamically provide multi-granularity source representations for better predicting the correct answer. As shown in Figure 2, the proposed approach builds connections with each encoder layer, so that the ABA-Net not only can exploit the coarse-grained representations, of the source sequence, which is instrumental in language modeling, but also can exploit the fine-grained representations of the source sequence, which is helpful to predict more precise answers. Hence, the answer predictor is encouraged to use source representations of different granularity, exploiting the expressive power of the model. Furthermore, due to better textual feature representations are at the core for boosting MRC performance, we propose to leverage the capsule network and self-attention module as the building blocks of encoders that separately encode the passage and question to obtain a more effective understanding of each word, i.e., the local and global representations. Last, before finally decoding to the probability that each position is the beginning or end of the answer span, we use our stacked encoder blocks to encode the generated representation again in the model encoding layer. Our experimental results show that each component has substantial gains in predicting answers correctly.

In general, our main contributions in this paper are as follows:

- We propose a novel approach called ABA-Net, which can encourage the answer predictor to take full advantage of the expressive power of the multi-layer encoder through exploring and exploiting the multi-granularity representations of source sequences.

- We propose to introduce the capsule network and self-attention module, which can better encode complicated text features and can capture the global representations, as our basic encoder block for obtaining effective representations. Our encoder blocks also can be stacked to form a hierarchy for multi-step interactions, which can gradually capture mutual relation and better transfer knowledge.

- Specifically, the proposed approach achieves the state-of-the-art performance on the SQuAD 1.0 dataset, showing the concrete benefit of exploring and exploiting multi-granularity representations.
Figure 2: An overview of the ABA-Net architecture (left). ABA-Net employs Adaptive Bidirectional Attention to reasonably exploit multi-granularity representations (red lines), and specific implementation details can be seen in the Section 3.2.2. We use the same Encoder Block (right) throughout the model, only varying the number of Conv-Pri-Dig and layernorm layers for each block. Notice that, we use depthwise separable convolutions (Chollet, 2017) instead of traditional ones. Following (Yu et al., 2018), we use a residual connection between each layer of the Encoder Block and share weights of the passage and question encoder in the embedding layer, as well as the three modeling encodes.

2 Related Work

Machine Reading Comprehension. In recent years, MRC (Chen et al., 2021a; Chen et al., 2021b; You et al., 2020a; Chen et al., 2021c) has been an active field in NLP. Much of this popularity can be attributed to the release of many annotated and publicly available datasets, such as CNN/Daily News (Hermann et al., 2015), WikiReading (Hewlett et al., 2016), SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), DuReaderrobust (Tang et al., 2020), Molweni (Li et al., 2020), etc. Based on these challenging datasets, a great number of end-to-end approaches have been proposed, including BiDAF (Seo et al., 2016), DCN (Xiong et al., 2016), R-Net (Wang et al., 2017). In MRC tasks, attention mechanism (Dong et al., 2020a; Gao et al., 2020; Zheng et al., 2020; Zhu et al., 2020; You et al., 2021a; Chen et al., 2022) have become an essential part to capture dependencies without considering their distance in the input/output sequences. Recently, some works show that well pre-trained models are powerful and convenient for downstream tasks, such as R-Trans (Liu et al., 2019a), DCMN+ (Zhang et al., 2019a), ALBERT (Lan et al., 2020) and GF-Net (Lee and Kim, 2020), which facilitate us to take pre-trained models as our backbone encoder.

Capsule Network. Capsule Networks were initially proposed by Hinton (2011), which demonstrate the robustness in learning meaningful representations than vanilla CNNs. In (Sabour et al., 2017), it referred to a capsule as a set of neurons whose activity vectors represent instantiation parameters of a
particular type of entity, such as an object or part thereof. Then Hinton (2018) introduced a novel iterative routing procedure between each capsule layer, which is based on the EM algorithm and shows a higher level of accuracy on the smallNORB (LeCun et al., 2004) dataset. Over the past few years, remarkable progress and research have been made with the capsule network in natural language processing (NLP). Dong (2020b) employs Capsule Network based on BiLSTM for the sentiment analysis. Zhao (2019), Wu (2020) and Yang (2019a) investigate Capsule Network for text classification.

Our work is different from our predecessors in the following aspects: (1) We propose a new attention mechanism to help MRC models leverage multi-granularity representations of each word reasonably; (2) We innovatively combine the capsule network with self-attention in a residual encoder block to capture the local and global representations, respectively; (3) To the best of our knowledge, our model is the first to apply the capsule network in MRC.

3 Model

3.1 Problem Formulation

In this paper, the reading comprehension task is defined as follows: Given a passage/paragraph with \( n \) words \( P = \{p_1, p_2, p_3, ..., p_n\} \) 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.

3.2 Model Overview

Generally, our model contains four major parts: an embedding layer, an interaction-attention layer, a model encoding 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 this layer, we use Glove (Pennington et al., 2014) as fundamental word embedding, and then they are passed through a BiLSTM (Hochreiter and Schmidhuber, 1997) and our proposed Encoder block separately. Moreover, we utilize ALBERT (Lan et al., 2020) contextualized word embedding. Interaction-Attention layer plays a key role in ABA-Net, which employs Adaptive Bidirectional Attention to distinguish and utilize multi-granularity presentations of each word reasonably. Model Encoding Layer encodes the generated representations again by our encoder blocks. Output Layer leverages the output of the previous layer to compute the final answer span by a bilinear projection.

3.2.1 Embedding Layer

Word-Char Embedding. We obtain the fixed word embedding\(^1\) from pre-trained Glove (Pennington et al., 2014) word vectors. Following (Yu et al., 2018), we obtain the character-level embedding of each word using CNN. Following (Seo et al., 2016), a two-layer highway network in (Srivastava et al., 2015) is used on the top. As results, we obtain the final embeddings for the tokens for \( P \) as a matrix \( E_p \in \mathbb{R}^{d \times n} \), and tokens in \( Q \) as a matrix \( E_q \in \mathbb{R}^{d \times m} \).

Contextualized Word Embedding. Since the context encoding generated by the pre-trained model plays an important role in machine reading comprehension, we employ ALBERT (Lan et al., 2020) as contextualized embedding\(^2\) like SDNET(Zhu et al., 2018). ALBERT generates \( L \) layers of hidden states for all BPE (Sennrich et al., 2015) tokens in a sentence/passage and we employ a weighted sum of these hidden states to obtain contextualized embedding. In detail, given a word \( w \), which is tokenized to \( S \) BPE tokens \( w = a_1, a_2, ..., a_s \), and ALBERT generates \( L \) hidden states for each BPE token, \( h^l_w \), \( 1 \leq l \leq L \), \( 1 \leq s \leq S \). The contextual embedding \( \text{ALBERT}_w \) for word \( w \) is then a per-layer weighted sum of average ALBERT embedding, with weights \( \theta_1, ..., \theta_L \):

\(^1\)We map the symbolic/surface feature of \( P \) and \( Q \) into neural space via word embeddings, 16-dim part-of-speech (POS) tagging embeddings, 8-dim named-entity embeddings and 4-dim hard-rule features.

\(^2\)Instead of adding a scoring layer to pre-trained models as proposed in many question answering models, we use the transformer output from ALBERT-large as contextualized embedding in our encoding layer.
\[
\text{ALBERT}_w = \sum_{l=1}^{L} \theta_l \sum_{s=1}^{S} h_l^s
\]

**BiLSTM.** In this component, we use two separate bidirectional RNNs (BiLSTMs (Hochreiter and Schmidhuber, 1997)) to learn the contextualized understanding for \(P\) and \(Q\).

\[
\begin{align*}
  h_{P,k}^1, \ldots, h_{P,k}^m &= \text{BiLSTM}(h_{P,k}^{1-1}, \ldots, h_{P,k}^{m-1}) \\
  h_{Q,k}^1, \ldots, h_{Q,k}^m &= \text{BiLSTM}(h_{Q,k}^{1-1}, \ldots, h_{Q,k}^{m-1})
\end{align*}
\]

where \(1 \leq k \leq K\) and \(K\) is the number of BiLSTM layers. Thus, we get the representation: \(\hat{C}_p \in \mathbb{R}^{d \times 2n}\) for passage and \(\hat{C}_q \in \mathbb{R}^{d \times 2m}\) for question.

**Encoder Block.** In this component, we leverage and extend the strategy in (Yu et al., 2018) to construct our own encoder block architecture. As showed in Figure 2, we propose the following basic building block: \([\text{convolution-PrimaryCaps-DigitCaps-layer} \times n + \text{self-attention-layer} + \text{feed-forward-layer}]\) to encode the contextual information of both passages/paragraphs and questions respectively. In detail, the kernel size is 7, the number of filters is \(d = 128\) and the number of Conv-Pri-Dig layers within a block is 5. Following (Yu et al., 2018), we adopt the multi-head attention mechanism defined in (Vaswani et al., 2017b), which calculates the weighted sum of all positions, for each position in the input, called a query or keys, based on the similarity between the query and key as measured by the dot product. Each of these basic operations (Conv-Pri-Dig/self-attention/ffn) is placed inside a residual block in (Yu et al., 2018). For an input \(x\) and a given operation \(f\), the output is \(f(\text{layernorm}(x)) + x\), meaning there is a full identity path from the input to output of each block. As results, final representation is defined as: \(C_p\) for passage and \(C_q\) for questions.

### 3.2.2 Interaction-Attention Layer

**Adaptive Bidirectional Attention.** After multiple layers extract different levels of presentations of each word, we conduct Adaptive Bidirectional attention from question to passage and passage to question based on all layers of generated representations to build connection of all layers of representation, which is showed in Figure 3. We define a novel concept of ‘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:

\[
\begin{align*}
HOS_p &= [C_w; C_c; E_p; ALBERT_p; C_p; \hat{C}_p] \\
HOS_q &= [Q_w; Q_c; E_q; ALBERT_Q; C_q; \hat{C}_q]
\end{align*}
\]

where \(C_w/Q_w\) and \(C_c/Q_c\) are word/char embedding matrixs for passage and question respectively.

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

\[
\begin{align*}
\lambda^p \times HOS^T_p &= \hat{HOS}^T_p \\
\lambda^q \times HOS^T_q &= \hat{HOS}^T_q
\end{align*}
\]
where the matrix $\lambda$ is a trainable matrix. 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.

However, since the data aggregation requires extensive computation resource, we specially train a set of selection parameters to reduce the dimension of vectors: $\alpha_1, \alpha_2, ..., \alpha_k$ to choose the three most relevant semantic vectors of $\hat{HOS}_p$ and $\hat{HOS}_q$. Thus, the input of bidirectional multilevel attention is defined as: $HOS^P$ for passages, $HOS^Q$ for questions.

**Bidirectional Attention.** In this component, our model dynamically computes attention of the embeddings from previous layers each time, are allowed to flow through to the downstream modeling layer. Like most high performing models, such as (Xiong et al., 2016), (Seo et al., 2016), (Yu et al., 2018), we construct passage-question and question-passage attention respectively. The attention function in (Seo et al., 2016) 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 rendering a similarity matrix $H \in \mathbb{R}^{n \times m}$. $H$ is computed as:

$$H = \text{dropout}(f_{\text{attention}}[HOS^P, HOS^Q])$$

(6)

After that, we use the strategy of (Yu et al., 2018) to normalize each row of $H$ by applying the softmax function, getting a matrix $\hat{H}$. Then the passage-to-question attention is computed as $M = \hat{H} \cdot HOS^Q^T$. Following DCN (Xiong et al., 2017), we compute the column normalized matrix $\hat{H}$ of $H$ by softmax function, and the question-to-passage attention is $S = \hat{H} \cdot \hat{H}^T \cdot HOS^P^T$. At last, our model use a simple concatenation as following to get the final representation, which shows good performance in our experiments: $O = [HOS^P; M; HOS^P \odot M; HOS^P \odot S]$.

### 3.2.3 Model Encoding Layer

The input of this layer is $O$, which encodes the question-aware representations of context words. This layer use the same building blocks as the Interaction Layer except that Conv-Pri-Dig layer number is 2 within a block and the total number of blocks are 4. Following (Yu et al., 2018), we share weights between each of the 3 repetitions of the model encoder.

### 3.2.4 Output Layer

This layer aims to calculate two scores for each word in the passage, corresponding to the probability that the answer begins and ends at that word. Like most existing models, the last output layer is task-specific. We adopt the following strategy (Yu et al., 2018) to predict the probability of each position in the passage/paragraph being the start or end of an answer span, if the question is answerable. More concretely, the distributed probabilities of the starting and ending index for the answer span are defined as:

$$P_{\text{begin}} = \text{softmax}(W_1[B_1; B_2])$$

(7)

$$P_{\text{end}} = \text{softmax}(W_2[B_2; B_3])$$

(8)

where $W_1$ and $W_2$ are two trainable variables and $B_1$, $B_2$, $B_3$ are respectively the outputs of the three model encoders, from bottom to top. If the question is unanswerable, the final output is an empty string. In addition, the proposed model can be customized to other comprehension tasks, e.g. selecting from the candidate answers, by changing the output layers accordingly.

**Objective** Following (Yu et al., 2018), the objective function is defined as the negative sum of the log probabilities of the predicted distributions indexed by true start and end indices, averaged over all the training examples:

$$L(\theta) = -\frac{1}{N} \sum_i^N [\log(p_{y_i^1}) + \log(p_{y_i^2})]$$

(9)

where $y_i^1$ and $y_i^2$ are respectively the ground truth starting and ending position of example $i$, and contains all the trainable variables.
Table 1: The performances of single models on different datasets.

| Single Model               | SQuAD 1.0 EM/F1 | SQuAD 2.0 EM/F1 | COQA EM/F1 | F1(average) |
|----------------------------|----------------|----------------|------------|-------------|
| SAN (Liu et al., 2018)     | 76.8/84.2      | 68.6/71.4      | -          |             |
| BiDAF++ (Shen et al., 2018)| 77.6/84.9      | 65.6/68.7      | 69.5       |             |
| QANet (Yu et al., 2018)    | 80.9/87.8      | 65.4/67.2      | -          |             |
| Bert_{large} (Devlin et al., 2018) | 85.1/91.8    | 79.9/83.1      | 74.4       |             |
| SDNet (Zhu et al., 2018)   | -              | 76.779.8       | 76.6       |             |
| SGNNet (Zhang et al., 2020a)| -              | 85.1/87.9      | -          |             |
| SemBERT_{large} (Zhang et al., 2019b) | -          | 86.1/88.8      | -          |             |
| RoBERTa_{large} (Liu et al., 2019d) | -            | 86.8/89.8      | 84.9       |             |
| ALBERT_{large} (Lan et al., 2019) | 89.1/94.6    | 86.8/89.6      | 85.4       |             |
| XLNet_{large} (Yang et al., 2019b) | 89.9/95.0  | 87.9/90.7      | 84.6       |             |
| Retr-Reader on ELECTRA (Zhang et al., 2020b) | -           | 89.5/92.0      | -          |             |
| Dev :ABA-Net               | 90.1/95.4      | 88.8/91.3      | 88.4       |             |
| Test :ABA-Net              | 90.4/95.8      | 89.6/91.7      | 88.5       |             |

Table 2: An ablation study of our model. The reported results are obtained on the test set and we use the original ABA-Net as the base model. Diff to B-M is the abbreviation of Difference to Base Model. sep\- conv is the abbreviation of depthwise separable convolutions.

|                  | SQuAD 1.0 Diff to B-M | SQuAD 2.0 Diff to B-M | COQA Diff to B-M | F1(average) |
|------------------|------------------------|-----------------------|------------------|-------------|
| Base ABA-Net     | 90.2/95.4              | 89.6/91.7             | 88.4             |             |
| - self-attention in encoders | 88.8/94.2   | -1.4/-1.2             | 88.0/90.4       | -1.6/-1.3   | 87.0        | -1.4     |
| - Capsule Network in encoders | 88.7/93.7  | -1.5/-1.7             | 88.5/90.4       | -1.1/-1.3   | 87.3        | -1.1     |
| - Adaptive Bidirectional Attention | 88.6/93.3  | -1.6/-2.1             | 87.8/89.5       | -1.8/-2.2   | 85.9        | -2.5     |
| replace sep\- conv with normal conv | 89.7/95.0  | -0.5/-0.5             | 89.2/91.2       | -0.4/-0.5   | 87.9        | -0.3     |

4 Experiments

In this section, we conduct experiments to study the performance of our model. We will primarily benchmark our model on the SQuAD 2.0 dataset (Rondeau and Hazen, 2018) and SQuAD 1.0 dataset (Rajpurkar et al., 2016). We also conduct similar studies on COQA dataset (Reddy et al., 2019), a conventional Q&A dataset, to show that the effectiveness and efficiency of our model are general. Detailed introduction of these datasets and experimental settings can be found in Appendix A.

4.1 Results

The F1 and Exact Match (EM) are two metrics of accuracy for evaluating MRC models performance. F1 measures the part of the overlapping mark between the predicted answer and groundtruth, and if the prediction is exactly the same as the ground truth, the exact match score is 1, otherwise it is 0. We show the results in comparison with other single models in Table 1. From the table, the accuracy (EM/F1) performance of our model is on par with the state-of-the-art models. In detail, ABA-Net achieves the state-of-the-art results on SQuAD 1.0 and competitive performance on SQuAD 2.0 and COQA compared with other baseline models. More concretely, ABA-Net improves EM/F1 score to 90.4/95.8, compared with previous state-of-the-art models on SQuAD 1.0; ABA-Net also achieves an EM score of 89.6 and an F1 score of 91.7 on SQuAD 2.0. Furthermore, our model achieves an F1 (average) score of 88.5 on COQA, which is competitive with the best documented result.
4.2 Ablation Study And Analysis

Ablation Study. We conduct ablation studies on components of the proposed model. The validation scores on the test set are shown in Table 2. As can be seen from the table, the Adaptive Bidirectional Attention plays a crucial part in our model: both F1 and EM drop dramatically by almost 2 percent if it is removed. Self-attention in the encoders is also a necessary component that contributes $1.4/1.2,1.6/1.3,1.4$ to the ultimate performance on three datasets. More significantly, the combination of Capsule Network and self-attention is significantly better than the combination of RNNs or CNNs and self-attention.

Analysis. We interpret these phenomena as follows: Adaptive Bidirectional Attention is able to distinguish and utilize multi-granularity representations of each word reasonably. The capsule network captures the local structure of the context and questions while the self-attention is able to model the global context representations, interactions and further focus on the most important part of the interaction. Hence they are complimentary to but cannot replace each other. Concretely, capsule networks have the ability for learning hierarchical relationships between consecutive layers by using routing processes without parameters and additionally improve the generalization capability. The use of separable convolutions instead of traditional convolutions also has a prominent contribution to performance, which can be seen by replacing the depthwise separable convolutions with conventional convolutions, resulting in slightly worse accuracy.

5 Discussions & Visualization

Effect of Adaptive Bidirectional Attention. We additionally perform experiments to prove the effectiveness of our proposed Adaptive Bidirectional Attention mechanism. 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 3) on SQuAD 2.0. As can be seen from Table 3, after adding Adaptive Bidirectional Attention, the performance of these models could be improved to varying degrees. This also proves the versatility of this attention mechanism.

Effect of Capsule Network. We additionally perform experiments to understand this mechanism, which is shown in Table 4. We first adopt normal convolutions to replace the Capsule Network and the effect of the model decreases heavily on three benchmark datasets. We attribute it to the fact that Capsule Network is better than CNNs in extracting rich text representation. Then we employ a shared two-layers BiLSTM to replace the Capsule Network, which gives $0.9/1.1$ and $0.6/1.0$ EM/F1 decrease, $0.3$ F1 decrease in our experiments. Notice that, due to the limitations of the existing capsule network, its performance on the complex multi-round dialogue reading comprehension dataset(COQA) does not perform better than other two datasets.

Effect of the total number of stacked blocks. Our model uses stacked encoder blocks to form a multi-step interactions. As showed in Figure 4, we observe that the performance of our model drops heavily when encountered with different number of blocks. Under the premise that the number of blocks is 4, ABA-Net achieves the best results on three datasets. We interpret this phenomena as follows: When the number of encoder blocks is less than 4, the local and global text features captured by the model are

Table 3: An comparison study of Adaptive Bidirectional Attention on SQuAD 2.0.

| Base Model          | EM/F1   |
|---------------------|---------|
| BiDAF++ (Shen et al., 2018) | 65.6/68.7 |
| QANet (Yu et al., 2018)    | 65.4/67.2 |
| SAN (Liu et al., 2018)     | 68.6/71.4 |
| SDNet (Zhu et al., 2018)   | 76.7/79.8 |
| SGNet (Zhang et al., 2020a) | 85.1/87.9 |

+ Adaptive Bidirectional Attention
| BiDAF++ (Shen et al., 2018) | 67.2/70.0 |
| QANet (Yu et al., 2018)     | 66.9/69.1 |
| SAN (Liu et al., 2018)      | 70.1/73.0 |
| SDNet (Zhu et al., 2018)    | 78.5/81.1 |
| SGNet (Zhang et al., 2020a) | 86.7/89.2 |

Figure 4: F1 accuracy of various stacked encoder blocks on three benchmark datasets.
Figure 5: Visualization of Syntax-guided attention (left) and Adaptive Bidirectional attention (right). 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 wrong judgment. In contrast, this phenomenon does not happen in our Adaptive Bidirectional attention.

| The Effect of Capsule Network | SQuAD 1.0 | SQuAD 2.0 | COQA |
|-------------------------------|-----------|-----------|------|
| Base ABA-Net                  | 90.2/95.6 | 89.6/91.7 | 88.5 |
| replace Capsule with normal conv | 88.3/93.6 | 88.4/90.3 | 87.5 |
| replace Capsule with sep conv  | 88.8/94.1 | 88.8/90.8 | 88.0 |
| replace Capsule with BiLSTM   | 89.3/94.5 | 89.0/90.7 | 88.1 |

Table 4: An comparison study of the Capsule Network.

not enough, and it is easy to overfit. On the contrary, if the number exceeds, the model extracts many repeated contextual representations, which will hurt the performance.

Visualization. To have an insight that how Adaptive Bidirectional Attention works, we draw attention distributions of the Syntax-guided attention of SGNet (Zhang et al., 2020a) and our proposed Adaptive Bidirectional attention, as shown in Figure 5. The visualization verifies that benefiting from Adaptive Bidirectional attention, our model is effective at distinguishing and utilizing different layers of presentation of each word in the context and the query, guiding the downstream layer to collect more relevant semantics to make predictions.

6 Conclusion

In this paper, we propose a novel end-to-end machine reading comprehension (MRC) model Adaptive Bidirectional Attention-Capsule Net-work (ABA-Net) for exploring and exploiting multi-granularity representations of source sequences. In particular, our ABA-Net can adaptively exploit the source representations of different levels to the predictor. More concretely, our core innovation is to propose a novel attention mechanism called Adaptive Bidirectional Attention, which guides 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. Furthermore, to explore better representations, we incorporate the capsule network and self-attention module as the building blocks of encoders. The experimental results show that our model achieves competitive performance compared to the state-of-the-art on three public benchmark datasets. The systematic analysis demonstrates the effectiveness of each component in our model and shows that our model can learn effective representations for MRC and is capable of understanding rich context and answering complex questions. Future work can
include an extension of employing Adaptive Bidirectional Attention to other question answering tasks.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Nuo Chen, Fenglin Liu, Chenyu You, Peilin Zhou, and Yuexian Zou. 2021a. Adaptive bi-directional attention: Exploring multi-granularity representations for machine reading comprehension. In *ICASSP*, pages 7833–7837. IEEE.

Nuo Chen, Linjun Shou, Min Gong, Jian Pei, and Daxin Jiang. 2021b. From good to best: Two-stage training for cross-lingual machine reading comprehension.

Nuo Chen, Chenyu You, and Yuexian Zou. 2021c. Self-supervised dialogue learning for spoken conversational question answering. *CoRR*, abs/2106.02182.

Nuo Chen, Linjun Shou, Ming Gong, Jian Pei, and Daxin Jiang. 2022. Bridging the gap between language models and cross-lingual sequence labeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1909–1923, Seattle, United States, July. Association for Computational Linguistics.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yanru Dong, Peiyu Liu, Zhenfang Zhu, Qicai Wang, and Qiuyue Zhang. 2020a. A fusion model-based label embedding and self-interaction attention for text classification. *IEEE Access*, 8:30548–30559.

Yongfeng Dong, Yu Fu, Liqin Wang, Yunliang Chen, Yao Dong, and Jianxin Li. 2020b. A sentiment analysis method of capsule network based on bi lstm. *IEEE Access*, 8:37014–37020.

Lianli Gao, Xuanhan Wang, Jingkuan Song, and Yang Liu. 2020. Fused GRU with semantic-temporal attention for video captioning. *Neurocomputing*, 395:222–228.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018*, Conference Track Proceedings. OpenReview.net.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. 

Hongxuan Tang, Jing Liu, Hongyu Li, Yu Hong, Hua Wu, and Haifeng Wang. 2020. Dureaderrobust: A chinese dataset towards evaluating the robustness of machine reading comprehension models. CoRR, abs/2004.11142.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. CoRR, abs/1706.03762.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 189–198.

Yujia Wu, Jing Li, Jia Wu, and Jun Chang. 2020. Siamese capsule networks with global and local features for text classification. Neurocomputing, 390:88–98.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. CoRR, abs/1611.01604.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Min Yang, Wei Zhao, Lei Chen, Qiang Qu, Zhou Zhao, and Ying Shen. 2019a. Investigating the transferring capability of capsule networks for text classification. Neural Networks, 118:247–261.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020a. Towards data distillation for end-to-end spoken conversational question answering. CoRR, abs/2010.08923.

Chenyu You, Nuo Chen, and Yuexian Zou. 2020b. Contextualized attention-based knowledge transfer for spoken conversational question answering. arXiv preprint arXiv:2010.11066.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021a. Mrd-net: Multi-modal residual knowledge distillation for spoken question answering. In Zhi-Hua Zhou, editor, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 3985–3991. International Joint Conferences on Artificial Intelligence Organization, 8. Main Track.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021b. Self-supervised contrastive cross-modality representation learning for spoken question answering. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 28–39.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. CoRR, abs/1804.09541.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019a. DCMN+: dual co-matching network for multi-choice reading comprehension. CoRR, abs/1908.11511.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware bert for language understanding. arXiv preprint arXiv:1909.02209.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020a. Sg-net: Syntax-guided machine reading comprehension. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 9636–9643. AAAI Press.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020b. Retrospective reader for machine reading comprehension. arXiv preprint arXiv:2001.09694.
Xuanyu Zhang. 2019. Mc2: Multi-perspective convolutional cube for conversational machine reading comprehension. In *ACL*.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019. Towards scalable and reliable capsule networks for challenging NLP applications. *CoRR*, abs/1906.02829.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. *CoRR*, abs/2005.05806.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *CoRR*, abs/1812.03593.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Dual multi-head co-attention for multi-choice reading comprehension. *CoRR*, abs/2001.09415.