DUMA: Reading Comprehension With Transposition Thinking

Pengfei Zhu, Zhuosheng Zhang, Hai Zhao, and Xiaoguang Li

***Abstract*—Multi-choice Machine Reading Comprehension (MRC) requires models to decide the correct answer from a set of answer options when given a passage and a question. Thus, in addition to a powerful Pre-trained Language Model (PrLM) as an encoder, multi-choice MRC especially relies on a matching network design that is supposed to effectively capture the relationships among the triplet of passage, question, and answers. While the newer and more powerful PrLMs have shown their strengths even without the support from a matching network, we propose a new DUal Multi-head Co-Attention (DUMA) model. It is inspired by the human transposition thinking process solving the multi-choice MRC problem by considering each other's focus from the standpoint of passage and question. The proposed DUMA has been shown to be effective and is capable of generally promoting PrLMs. Our proposed method is evaluated on two benchmark multi-choice MRC tasks, DREAM, and RACE. Our results show that in terms of powerful PrLMs, DUMA can further boost the models to obtain higher performance.**

***Index Terms*—Attention network, machine reading comprehension, pre-trained language model.**

TABLE I
An Example of DREAM Dataset

Passage	Woman: *So, you have three days off, what are you going to do?* Man: *Well, I probably will rent some movies with my friend bob.*
Question	*What will the man probably do?*
Answer options	1) *Ask for a three-day leave.* 2) *Go out with his friend.* **3) √ *Watch films at home.***

TABLE II
Improvements of Several Prior Models for Representative PrLMs (Sorted by Releasing Time) on RACE Dataset

		+OCN	+WAE	+DCMN
BERT$_{\text{large}}$	71.0	71.7(+0.7)*	73.1(+2.1)	75.8(+4.8)*
XLNet$_{\text{large}}$	80.1	80.9(+0.8)	81.8(+1.7)	82.8(+2.7)*
ALBERT$_{\text{xxlarge}}$	86.6	87.2(+0.6)	87.3(+0.7)	85.7(-0.9)
ELECTRA$_{\text{large}}$	86.1	86.3(+0.2)	86.9(+0.8)	84.9(-1.2)

I. Introduction

Machine Reading Comprehension has been an active topic and challenging problem. Various datasets and models have been proposed in recent years [3], [16], [38]. For an MRC task, given a passage and a question, the task can be categorized as *generative* or *selective* according to its answer style [2]. *Generative* tasks require the model to generate answers according to the passage and the question but not limited to spans of the passage. On the other hand, *selective* tasks give the model several candidate answers from which to select the best one. Multi-choice MRC is a typical task of *selective* type and is the focus of this paper. Table I shows one example of the DREAM dataset [28], whose task is to select the best answer among three candidates given a particular tuple of passage and question.

The kernel method for a model to solve the MRC problem is a two-level hierarchical process, 1) representation encoding, which is done by an encoder such as PrLM, and 2) capturing the relationship among the triplet of passage, question and answer, which has to be carefully handled by various matching networks such as OCN [25] and DCMN [38]. With the development of PrLMs, matching network design tends to become more complicated to achieve more effective improvements.

Table II shows that the newer variant of the PrLM such as ALBERT [17] has shown its power even without the support from a proper matching network. At the same time, the previous models[1] [14], [25], [38] either provide very limited improvements or even cause a drop in performance of the PrLMs [6], [8], [17], [37]. This motivates us to develop a more effective mechanism to support powerful PrLMs. Instead of designing more complicated matching network patterns, we choose a going-back-to-the-basics way to draw inspiration from human experience on solving MRC problems, which intuitively is to first **1)** quickly read through the overall content of the passage, question and answer options to build up a global impression, followed by a **transposition thinking** process: **2)** based on dedicated

Manuscript received May 29, 2021; revised September 19, 2021, November 8, 2021, and December 17, 2021; accepted December 21, 2021. Date of publication December 28, 2021; date of current version January 8, 2022. This work was supported in part by the Key Projects of National Natural Science Foundation of China under Grants U1836222 and 61733011, and in part by the Chinese National Key Laboratory of Science and Technology on Information System Security. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jing Huang. *(Corresponding author: Hai Zhao.)*

Pengfei Zhu, Zhuosheng Zhang, and Hai Zhao are with the Department of Computer Science and Engineering, Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhupf97@sjtu.edu.cn; zhangzs@sjtu.edu.cn; zhaohai@cs.sjtu.edu.cn).

Xiaoguang Li is with the Huawei Noah's Ark Laboratory, Shenzhen 518129, China (e-mail: lixiaoguang11@huawei.com).

Digital Object Identifier 10.1109/TASLP.2021.3138683

[1] We have re-implement OCN and WAE and have obtained codes of DCMN through personal communication with its authors. In addition, the results denoted with * are from original papers.

information from question and answer options, re-considerate details of the passage and collect supporting evidences for the question and answer options, **3)** based on dedicated information from the passage, re-consider the question and answer options to decide the correct option and exclude wrong options. When humans are re-reading the passage, they tend to extract key information according to their impression of question and answer options, and it is the same when re-reading question and answer options. It can be regarded as a bi-directional process in terms of transposition thinking pattern, and we adopt an attention-inside network design to simulate this procedure, whose details are shown in the following Section *Model*.

Since the time the attention mechanism was originally proposed [1] for Neural Machine Translation, it has been widely used in MRC tasks to model the relationships between the passage and the question and it effectively enhances nearly all kinds of tasks [26], [38], [39]. The attention mechanism computes the relationships of each word representation in one sequence to a target word representation in another sequence and aggregates them to form a final representation, which is commonly named passage-to-question attention or question-to-passage attention.

Transformer [33] uses the self-attention mechanism to represent dependencies and relationships between different positions in one single sequence. This is an effective method to obtain representations of sentences for global encoding. Since [8], [21] used it to improve the structure of PrLMs [19], many kinds of PrLMs have been proposed to constantly refresh records of all kinds of tasks [17], [18]. For PrLMs, the more layers and larger hidden size they use, the better the performance they achieve. Benefiting from large-scale self-supervised training data and multiple stacked layers, PrLMs are able to encode sentences into very deep and precise representations. Moreover, [17] reveals that decoupling the word embedding and the contextualized representation and applying parameter sharing through a very deep model can reduce the model size significantly while sustaining the performance. At the same time, continually extending the hidden size leads to even higher performance. Parameter sharing may make it easier to train a deep model, and a larger hidden size may benefit learning to encode more informative representations. However, training an LM is a time consuming and labor intensive task requiring engineering effort to explore parameter settings. The larger the model is, the more resources it consumes and the harder it is to implement. Moreover, despite the great success they achieve on different tasks, we find that for MRC tasks, using self-attention of the Transformer to model sequences is far from sufficient. Regardless of how deep the structure is, it suffers from the nature of self-attention, which is to concatenate all the tokens of the question, passage, and answer options into one sequence. As a consequence, each token is treated equally when using a linear projection to transform into the same representation subspace in order to apply the attention operation, which then results only in a global relationship being inferred. However, for MRC tasks, the passage and the question are remarkably different in contents and literal structures, and the relationship between them necessarily needs to be carefully considered. However, previous models [1], [26], [38] obtain only limited improvements when applied on top of PrLMs even though they use very complicated structures.

Rather than seeking a complicated matching network pattern, we, inspired by the human thinking experience in solving MRC problems, put forward a new network design named **DU**al **M**ulti-head Co-**A**ttention (DUMA) to sufficiently capture relationships among the passage, question and answer options for multi-choice MRC; as a result, it may effectively improve the performance when cooperating with newer and more powerful PrLMs. Our model is based on the Multi-head Attention, which is the kernel module of the Transformer. Similar to BiDAF [26] and DCMN [38], we use the bi-directional method to obtain sufficient modeling of relationships. The contributions of this work can be summarized as:

1) For multi-choice MRC tasks, we investigate the effects of previous models over Pre-trained Language Models.
2) We propose a new **DU**al **M**ulti-head Co-**A**ttention (DUMA) model, which simulates the procedure of solving MRC tasks by humans and shows its effectiveness and superiority to previous models through extensive experiments.
3) We show that the proposed DUMA can effectively improve PrLMs to obtain higher performance on two benchmark multi-choice MRC tasks, DREAM and RACE.

II. Related Works

A. Pre-Trained Language Model

Emerging large-scale deep PrLMs have led the field of Natural Language Processing (NLP) to a new era. Trained on a huge amount of self-supervised data crawled from the internet, they can learn dynamic and informative representations of language. Embedding from Language Models (ELMo) [20] uses a deep biLSTM architecture to train a Language Model on the 1B Word Benchmark [4]. Generative Pre-trained Transformer (OpenAI GPT) [22] first introduces Transformer [33] into training large Pre-trained Language Models using a uni-directional method for modeling and is trained on the BooksCorpus dataset [41]. Afterward, the most influential PrLM, namely, BERT [8] was proposed, which introduces a bi-directional modeling method and a Masked Language Model (MLM) pre-training task. Since then, many kinds of PrLMs have been proposed: Generalized Autoregressive Pre-training (XLNet) [37], which proposes Permutation Language Modeling; Robustly Optimized BERT Pre-training approach (Roberta) [18], which applies many techniques for better training; A Lite BERT (ALBERT) [17], which proposes to share parameters among the modeling layers; and ELECTRA [6], which uses a replaced token detection method for language modeling and significantly improves training efficiency.

PrLMs substantially improve the performance on various MRC datasets, which are considered the most challenging tasks in NLP. MRC requires the model to have a deep understanding of language and to be able to do some reasoning, which PrLMs can achieve by learning from large-scale self-supervised data. PrLMs continuously refresh the leaderboards of several MRC datasets: SQuAD [23], SQuAD 2.0 [24], RACE [16], DREAM [28], etc. However, despite the great success of PrLM, its ability to solve MRC problems can still be further improved. After using PrLM as an encoder, a well-designed matching network is able to further mine the deep related information to the benefit of reasoning.

B. *Multi-Choice Machine Reading Comprehension*

Multi-choice Machine Reading Comprehension (MCRC) is a representative MRC task that requires the model to select one answer from multiple candidates and usually needs reasoning over the question and the passage. Many works aim at designing networks for better matching question, passage and answer choices, which is also the main focus of this work. Other methods aim at adopting better training strategies or more training data, such as MMM [13], which proposes a two-stage course-to-fine training method, and [12], which adopts all kinds of QA datasets to enhance the model for the RACE dataset.

C. *Attention Mechanism and Matching Networks*

[1] first proposes the attention mechanism for Neural Machine Translation. The joint learning of alignment and translation effectively improves the performance. Since then, the attention model has been introduced to all kinds of NLP tasks, and various architectures have been proposed [5], [9], [32], [36]. [26] uses a multi-stage architecture to hierarchically model representation of the passage and uses a bi-directional attention flow. These works, which were proposed before the emergence of PrLMs, are able to model the representations well on top of traditional encoders such as Long Short-Term Memory (LSTM) [11]. In fact, the experimental results show that they can still improve the representations of PrLMs, but the improvements are suboptimal.

Based on PrLMs, [25] proposes a method to model relationships and interactions among answer options to the benefit of distinguishing them. [14] integrates a model that learns to select the wrong answer. [38] proposes a sentence selection method to select more important sentences from the passage to improve the matching representations and considers interactions among answers for multi-choice MRC tasks. Even though the matching network design becomes more complicated, it cannot fully exploit powerful PrLMs and even cause a drop in performance when applied on newer PrLMs[2].

In a word, when applied on top of PrLMs, previous models are not effective enough to improve the performance by a large margin. Thus, inspired by the experience of humans solving MRC problems, we design a new model that can effectively utilize well-modeled representations of PrLMs to achieve better performance.

III. Task Definition

Multi-choice MRC tasks have to handle a triplet of passage P, question Q, and answer A. When given the passage and question, the model is required to produce a correct answer. The passage consists of multiple sentences, and its content can be dialogue, story, news, and so on, depending on the domain of the dataset. The questions and corresponding answers are single sentences, which are usually much shorter than the passage. The target of multi-choice MRC is to select the correct answer from the candidate answer set $A = \{A_1, \ldots, A_t\}$ for a given passage and question pair $< P, Q >$, where t is the number of candidate answers. Formally, the model needs to learn a probability distribution function $F(A_1, A_2, \ldots, A_t | P, Q)$.

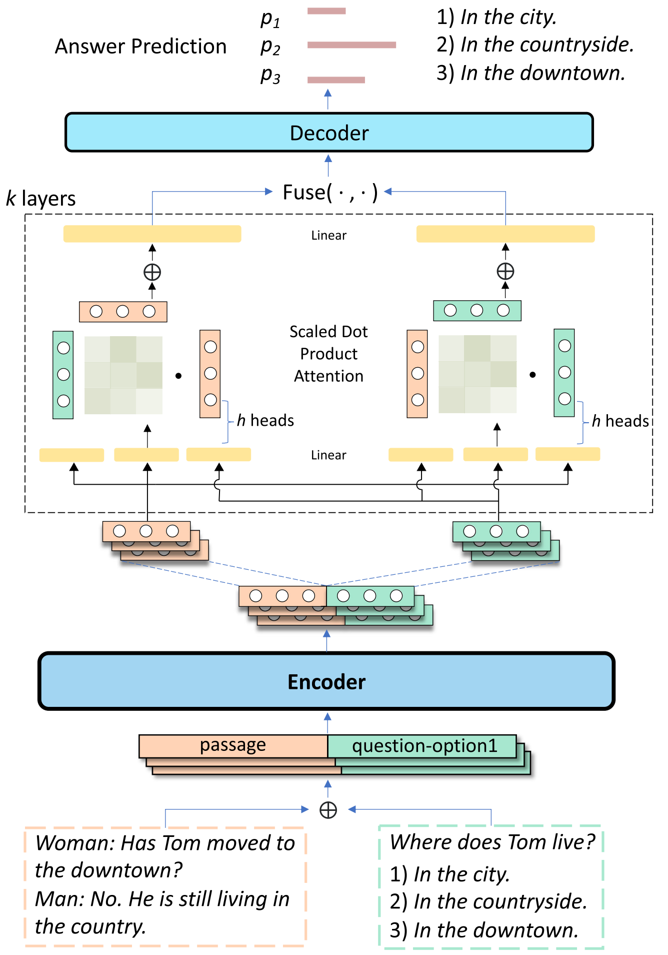

Fig. 1. The overall architecture. Our proposed DUMA is between the Encoder and Decoder.

IV. Model

Fig. 1 illustrates the overall architecture of our model. The encoder takes text inputs to form a global sequence representation, which is similar to human reading through the whole content for the first time to obtain an overall impression. The decoder performs the answer prediction, which is similar to human aggregating all the information to select the correct answer option. Our proposed Dual Multi-head Co-Attention (DUMA) layer is between the encoder and the decoder, which simulates a human transposition thinking process to capture relationships of key information from the passage, question, and answer options.

[2] As shown in Table II.

A. Encoder

To encode input tokens into representations, we take PrLMs as the encoder. To obtain a global contextualized representation, for each candidate answer, we concatenate it with its corresponding passage and question to form one sequence and then feed it into the encoder. Let $P = [p_1, p_2, \ldots, p_m]$, $Q = [q_1, q_2, \ldots, q_n]$, and $A = [a_1, a_2, \ldots, a_k]$ denote the sequences of passage, question and a candidate answer, respectively, where p_i, q_i, and a_i are tokens. The adopted encoder with encoding function $Enc(\cdot)$ takes the concatenation of P, Q and A as the input, namely, $E = Enc(P \oplus Q \oplus A)$. The encoding output E has the form $[e_1, e_2, \ldots, e_{m+n+k}]$, where e_i is a vector of fixed dimension d_{model} that represents the respective token.

B. Dual Multi-Head Co-Attention

We use our proposed Dual Multi-head Co-Attention module to calculate attention representations of the passage and question-answer. Fig. 1 shows the details of our proposed DUMA, which may be stacked as k layers. The following formula assumes $k = 1$ for simplicity. Our model is based on the Multi-head Attention module [33]. The proposed DUMA reuses the architecture of Multi-head Attention, while for the inputs, $\mathcal{K}$ and $\mathcal{V}$ are the same but $\mathcal{Q}$ is another sequence representation (note that $\mathcal{Q}$ here denotes *Query* from the original paper, different from previous Q in this paper. $\mathcal{K}$ and $\mathcal{V}$ are *Key* and *Value*, respectively). We first separate the output representation from the Encoder to obtain $E^P = [e_1^p, e_2^p, \ldots, e_{l_p}^p]$ and $E^{QA} = [e_1^{qa}, e_2^{qa}, \ldots, e_{l_{qa}}^{qa}]$, where e_i^p, e_j^{qa} denote the i-th and j-th token representations of passage and question-answer, respectively, and l_p, l_{qa} are the lengths. Then, we calculate the attention representations in a bi-directional way, that is, take 1) E^P as *Query*, E^{QA} as *Key* and *Value*, and 2) E^{QA} as *Query*, E^P as *Key* and *Value*.

$$\mathrm{Attention}(E^P, E^{QA}, E^{QA}) = \mathrm{softmax}\left(\frac{E^P (E^{QA})^T}{\sqrt{d_k}}\right) E^{QA}$$

$$\mathrm{head_i} = \mathrm{Attention}(E^P W_i^Q, E^{QA} W_i^K, E^{QA} W_i^V)$$

$$\mathrm{MHA}(E^P, E^{QA}, E^{QA}) = \mathrm{Concat}(\mathrm{head_1}, \ldots, \mathrm{head_h}) W^O$$

$$\mathrm{MHA}_1 = \mathrm{MHA}(E^P, E^{QA}, E^{QA})$$

$$\mathrm{MHA}_2 = \mathrm{MHA}(E^{QA}, E^P, E^P)$$

$$\mathrm{DUMA}(E^P, E^{QA}) = \mathrm{Fuse}(\mathrm{MHA}_1, \mathrm{MHA}_2) \quad (1)$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$ are parameter matrices, d_q, d_k, d_v denote the dimension of *Query* vectors, *Key* vectors and *Value* vectors, h denotes the number of heads, $MHA(\cdot)$ denotes Multi-head Attention and $DUMA(\cdot)$ denotes our Dual Multi-head Co-Attention. The $Fuse(\cdot,\cdot)$ function first uses mean pooling to pool the sequence outputs of $MHA(\cdot)$ and then aggregates the two pooled outputs through a fusing method. In Subsection *Investigation of the Fusing Method*, we investigate three fusing methods, namely, element-wise multiplication, element-wise summation and concatenation.

TABLE III
STATISTICAL DATA OF DREAM AND RACE DATASET. # DENOTES THE NUMBER. "EXTRACTIVE" MEANS THE ANSWERS ARE SPANS OF THE PASSAGE, AND "ABSTRACTIVE" MEANS THE ANSWERS ARE NOT SPANS

	DREAM	RACE
# of source documents	6,444	27,933
# of total questions	10,197	97,687
Train/Dev/Test split	3:1:1	18:1:1
Extractive (%)	16.3	13.0
Abstractive (%)	83.7	87.0
Average answer length	5.3	5.3
# of answers per question	3	4
Avg./Max. # of turns per dialogue	4.7 / 48	-
Avg. passage length	85.9	321.9
Vocabulary size	13,037	136,629

As shown in Fig. 1, the right part of DUMA calculates question-answer-aware passage representation, which simulates human re-reading details in the passage while keeping in mind the question and answer, and the left part calculates passage-aware question-answer representation, which simulates re-considering the question-answer with a deeper understanding of the passage. The $Fuse(\cdot,\cdot)$ function denotes fusing all the key information before deciding which is the best answer option.

C. Decoder

Our model decoder takes the outputs of DUMA and computes the probability distribution over answer options. Let A_i denote the i-th answer option, $O_i \in \mathbb{R}^l$ denote the output of the i-th $< P, Q, A_i >$ triplet, and A_r denote the correct answer option. The loss function is computed as:

$$O_i = \mathrm{DUMA}(E^P, E^{QA_i})$$

$$L(A_r|P, Q) = -\log \frac{\exp(W^T O_r)}{\sum_{i=1}^{s} \exp(W^T O_i)} \quad (2)$$

where $W \in \mathbb{R}^l$ is a learnable parameter and s denotes the number of candidate answer options.

V. EXPERIMENTS

Our proposed method is evaluated on two multi-choice MRC benchmarks, DREAM and RACE. Table III shows their data statistics, which indicates that RACE is a large-scale dataset covering a broad range of domains, and DREAM is a small dataset presenting the passage in a form of the dialogue.

a) DREAM: DREAM [28] is a dialogue-based dataset for Multi-choice Reading Comprehension. It is collected from English exams. Each dialogue has multiple corresponding questions, and each question has three candidate answers. The most important feature of the dataset is that most of the questions are non-extractive and need reasoning based on more than one sentence. Thus the dataset, even though small, is still challenging.

b) RACE: RACE [16] is a large dataset collected from middle and high school English exams. Most of the questions also need reasoning, and domains of passages are diverse, ranging from news and story to ads.

TABLE IV
RESULTS ON DREAM DATASET. RESULTS WITH MULTI-TASK LEARNING ARE REPORTED BY [34].

model	dev	test	source
BERT$_{\text{large}}$ [8]	66.0	66.8	leaderboard
BERT$_{\text{large}}$+WAE [14]	-	69.0	
XLNet$_{\text{large}}$ [37]	-	72.0	
RoBERTa$_{\text{large}}$+MMM [13]	88.0	88.9	
ALBERT$_{\text{xxlarge}}$ [17]	89.2	88.5	
ALBERT$_{\text{xxlarge}}$+DUMA	**89.9**	**90.5**	our model
+multi-task learning [34]	-	**91.8**	

TABLE V
RESULTS ON RACE DATASET

model	test (M/H)	source
HAF [40]	46.0(45.0/46.4)	publication
MRU [31]	50.4(57.7/47.4)	
HCM [35]	50.4(55.8/48.2)	
MMN [30]	54.7(61.1/52.2)	
GPT [21]	59.0(62.9/57.4)	
RSM [29]	63.8(69.2/61.5)	
OCN [25]	71.7(76.7/69.6)	
XLNet [37]	81.8(85.5/80.2)	
XLNet$_{\text{large}}$ + DCMN+ [38]	82.8(86.5/81.3)	
ALBERT(ensemble) [17]	89.4(91.2/88.6)	leaderboard
Megatron-BERT (single) [27]	89.5(91.8/88.6)	
ALBERT-SingleChoice + transfer learning [12]	**90.7(92.8/89.8)**	
Megatron-BERT (ensemble) [27]	90.9(93.1/90.0)	
ALBERT-SingleChoice + transfer learning (ensemble)	**91.4(93.6/90.5)**	
ALBERT$_{\text{xxlarge}}$	86.6(89.0/85.5)	
ALBERT$_{\text{xxlarge}}$+DUMA	**88.0(90.9/86.7)**	our model
ALBERT$_{\text{xxlarge}}$+DUMA (ensemble)	**89.8(92.6/88.7)**	

TABLE VI
COMPARISON WITH ALBERT BASELINE ON RACE DATASET

model	dev	test (M/H)
ALBERT$_{\text{xxlarge}}$	87.4	86.6(89.0/85.5)
ALBERT$_{\text{xxlarge}}$ +DUMA	**88.1**(+0.7)	**88.0(90.9/86.7)**(+1.4)

A. Evaluation

For multi-choice MRC tasks, the evaluation criterion is accuracy with $acc = N^{+}/N$, where N^{+} denotes the number of examples the model selects the correct answer, and N denotes the total number of the evaluation examples.

B. Experimental Settings

Our model takes ALBERT$_{\text{xxlarge}}$ as the Encoder and uses $k = 2$ layers of DUMA. We make the left and right parts of DUMA, and all the layers share parameters. Using the PrLM, our model training is done through a fine-tuning method for both tasks.

Our codes are written based on Transformers,[3] and the ALBERT model version we use is *v2*. The results of the ALBERT [17], ELECTRA [6] and BERT [8] models as baselines are our re-running for the purpose of fair comparison with applying additional modules on them.

For the DREAM dataset, the learning rate is 1e-5, the batch size is 8 and the warmup steps are 100. We train the model for 2 epochs in 4 hours. For the RACE dataset, the learning rate is 1e-5, the batch size is 8 and the warmup steps are 1000. We train the model for 3 epochs in 2 days. For each dataset, we use FP16 training from Apex[4] to accelerate the training process. We train the models on eight nVidia P40 GPUs. In Section *Analysis Studies*, for other re-running or re-implementation, including PrLM baselines and PrLM plus other models for comparison, we use the same learning rate, warmup steps, and batch size as mentioned above.

We choose the result on the dev set that has stopped increasing for three checkpoints (382 steps for DREAM and 3000 steps for RACE). To obtain stable results, we run experiments 5 times with different random seeds and select the median as the ultimate performance.

C. Results

Tables IV, V and VI show the experimental results.[5] For the RACE dataset, "(M/H)" denotes performance on only the middle school or high school part of the dataset.

Megatron-BERT [27] is a variant of BERT [8], which has 8.3 billion parameters and is nearly 40 times larger than the largest size of ALBERT; therefore, it is usually very hard to apply in practice with the present commonly available computation power, and its results are not strictly comparable to our ALBERT+DUMA. ALBERT-SingleChoice + transfer learning [12] adopts AutoML [10] to search for better parameters and uses additional training data.

On both the RACE[6] and DREAM[7] benchmarks, our DUMA outperforms previous attention models or matching network designs by a large margin. It can be further improved by the multi-task learning method MMM [13], [34].

VI. ANALYSIS STUDIES

We perform ablation experiments on the DREAM dataset to investigate key features of our proposed DUMA, such as attention modeling ability, structural simplicity, bi-directional setting, and low coupling.

A. Comparison With Vanilla Self-Attention and Transformer Block

We investigate whether the improvements are simply caused by the increase in parameters. Thus, we conduct the experiments of ALBERT plus vanilla Multi-head Self-attention [33], whose inputs $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{V}$ are all concatenations of passage, question, and answer. The results shown in Table VII indicate the effectiveness of our bi-directional co-attention model design.

[3] https://github.com/huggingface/transformers

[4] https://github.com/NVIDIA/apex

[5] Our codes are available at https://github.com/pfZhu/duma_code.

[6] http://www.qizhexie.com/data/RACE_leaderboard

[7] https://dataset.org/dream/

TABLE VII
COMPARISON AMONG VANILLA MULTI-HEAD SELF-ATTENTION, DUMA AND TB-DUMA ON DREAM DATASET

model	dev	test
ALBERT$_{\text{base}}$	64.51	64.43
+Vanilla SA	66.27	66.34
+DUMA	67.06	**67.56**
+TB-DUMA	**67.79**	67.17

TABLE VIII
COMPARISON AMONG DIFFERENT MODELS ON DREAM DATASET

model	ALBERT$_{\text{base/xxlarge}}$	ELECTRA$_{\text{large}}$
baseline	64.4/88.5	88.2
+Soft Attention [1]	65.4(+1.0)/88.9(+0.4)	88.8(+0.6)
+BiDAF [26]	65.6(+1.2)/89.3(+0.8)	89.1(+0.9)
+OCN [25]	65.8(+1.4)/89.2(+0.7)	89.0(+0.8)
+WAE [14]	66.5(+2.1)/89.9(+1.4)	89.5(+1.3)
+DCMN[8] [38]	63.3(-1.1)/87.8(-0.7)	87.7(-0.5)
+DUMA	**67.6(+3.2)/90.5(+2.0)**	**89.8(+1.6)**

Moreover, we observe that the original Transformer Block (TB) [33] consists not only of the *Multi-head Attention* module but also of the *Layer Normalization (LN)* and the *Feed-Forward Network (FFN)*. In consideration of the extensive application and great success of TB for global encoding [8], [17], [18], we investigate whether the Transformer Block models the co-attention relationships better than Multi-head Attention using TB-based DUMA (TB-DUMA). The experimental results shown in Table VII indicate that TB-DUMA has no obvious difference from our DUMA in modeling relationships. However, our proposed DUMA has a more succinct structure and equally effective performance.

B. Comparison With Related Models

We compare our attention model with several representative works, which have been discussed in Section *Related Works*. Soft Attention [1] and BiDAF [26] were originally based on traditional encoders such as LSTM [11], and DCMN [38], OCN [25], and WAE [14] are based on BERT [8]. For a fair comparison with Soft Attention, we simply use it to replace the attention score computing in our model.

Table VIII compares the effectiveness of various model designs, and our proposed DUMA outperforms all other models. The performance of Soft Attention is much lower than that of DUMA, which indicates that the DUMA's similarity in structure with ALBERT (both use Multi-head Attention) makes it utilize the information from encoded representation better. Although BiDAF has been a successful attention model for a long time, it is suboptimal for PrLMs. WAE uses an ensemble model design with nearly twice the number of parameters as our model. DCMN adopts a much more complicated model structure design for better matching, but the results with ALBERT and ELECTRA are not satisfactory. This indicates that it may be specially optimized for a specific PrLM, while our DUMA achieves the absolutely highest accuracy with an intuitive structure design. In fact, our DUMA has a good generalization ability because it also works well with many kinds of PrLMs.

TABLE IX
COMPARISON AMONG DIFFERENT IMPLEMENTATION OF THE FUSING METHOD ON DREAM DATASET. THE LAST THREE ROWS ARE OUR DUMA APPLYING THREE KINDS OF IMPLEMENTATIONS

model	dev	test
ALBERT$_{\text{base}}$	64.51	64.43
element-wise multiplication	65.29	64.58
element-wise summation	66.32	65.51
concatenation	**67.06**	**67.56**

TABLE X
COMPARISON OF NUMBER OF PARAMETERS AMONG DIFFERENT MODELS. THE MODELS ARE SAME AS LISTED IN TABLE VIII.

model	parameter number
ALBERT$_{\text{base}}$	11.7M
+Soft Attention [1]	13.5M (+1.8M) (+15.4%)
+BiDAF [26]	12.0M (+0.3M) (+2.6%)
+OCN [25]	14.8M (+3.1M) (+26.5%)
+WAE [14]	23.4M (+11.7M) (+100%)
+DCMN [38]	19.4M (+7.7M) (+65.8%)
+DUMA	13.5M (+1.8M) (+15.4%)

We note that BiDAF [26] also employs a bi-directional setting that may seem somehow related to ours. However, they are not comparable and are quite different in many aspects. BiDAF is a classical method that is designed for MRC tasks of span extraction type. Our DUMA is different in the following aspects: 1) Motivation. BiDAF focuses on the classical span extraction MRC problem [23] and is motivated by improving the uni-directional attention mechanism with bi-directional attention. In contrast, our work aims at solving more challenging multi-choice MRC whose question and passage have less overlap and often needs some reasoning to solve. Further, DUMA is motivated by continuously improving PrLMs and studying the superiority of co-attention compared with PrLMs' self-attention. 2) Methodology. Compared with BiDAF's hierarchical architecture, DUMA models the attention in one step using "Multi-head Scaled Dot-product Attention," which not only is computationally faster and space-efficient but also makes the model learn different information by multiple attention heads. 3) Application. DUMA adopts PrLMs as the encoder and enhances interactions among question, passage, and answer options based on representations encoded by powerful enough PrLMs, while BiDAF models the interaction based on traditional weaker encoders such as LSTM.

C. Investigation of the Fusing Method

We investigate different implementations of the fusing function from (1), namely, element-wise multiplication, element-wise summation and concatenation. The results are shown in Table IX. We see that concatenation is optimal because it retains the matching information and lets the network learn to fuse them dynamically.

[8] The results of ALBERT+DCMN are our re-running of the official codes which we obtained through personal communication with its authors.

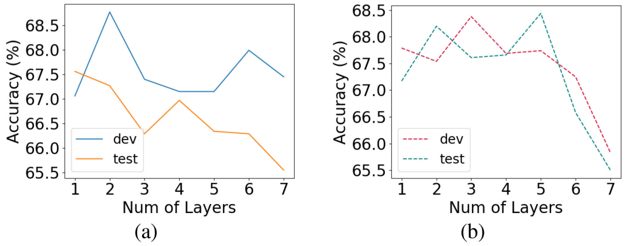

Fig. 2. (a) Different numbers of DUMA layers on the DREAM dataset. (b) Different numbers of TB-DUMA layers on the DREAM dataset.

D. *Number of Parameters*

We compare the number of parameters among different models in Table X. BiDAF requires the least model enlargement; however, it is far less effective than our model. In addition, our model enlargement is far less than that of DCMN. In a word, DUMA can obtain the best performance while requiring a small model enlargement.

E. *Number of DUMA Layers*

We stack 2 layers of DUMA to make passage and question-answer interact more than once to obtain deeper representations. In addition, we make different layers share parameters, which is the same as ALBERT.

Fig. 2(a) shows the results. We can see that as the number of layers increases, the performance fluctuates, and too many layers even leads to a slight drop. It is much like when humans solve MRC tasks where excessive thinking and hesitation may make them misunderstand the meaning of some information. For the network with the current number of parameters, it shows that interacting twice is enough to capture the key information, and that stacking too many layers may disorder the well-modeled representations and make the model harder to train.

Note that PrLMs [8], [17], [18] stacks Transformer Blocks (described in Subsection *Comparison With Vanilla Self-Attention and Transformer Block*) instead of Multi-head Attention modules. This raises the doubt whether the lack of the LN and FFN makes DUMA unsuitable for stacking a deeper network. Hence, we further conduct an experiment with TB-DUMA (the same as described in Subsection *Comparison With Vanilla Self-Attention and Transformer Block*). The experimental results in Fig. 2(b) show the same performance trend as the original DUMA, which again verifies the effectiveness of our DUMA design.

F. *Effect of Bi-Direction*

As established by [38], bi-directional matching is a very important feature for sufficiently modeling the relationship between passage and question. To investigate this effect, we perform experiments on two settings, namely, P-to-Q only and Q-to-P only. In other words, we respectively remove the right part and left part of DUMA. Table XI shows the results. We see that for the bi-directional model, the overall improvement is 2.84%, while for the uni-directional model, the improvement is only 2.06% at most. The setting of bi-direction effectively improves the performance, which reveals its efficiency for modeling the relationship and this conclusion is in line with the conclusion of [38]. Additionally, it is also in agreement with our intuitive understanding that all the passage, question, and answer options should be deliberated.

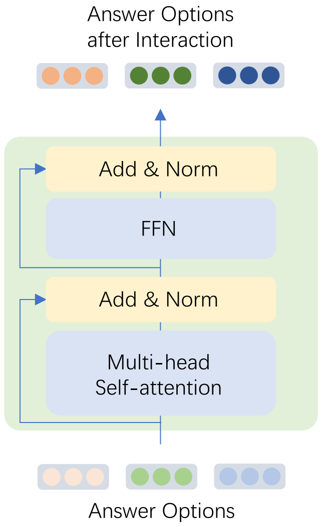

Fig.3. Answer option comparison module.

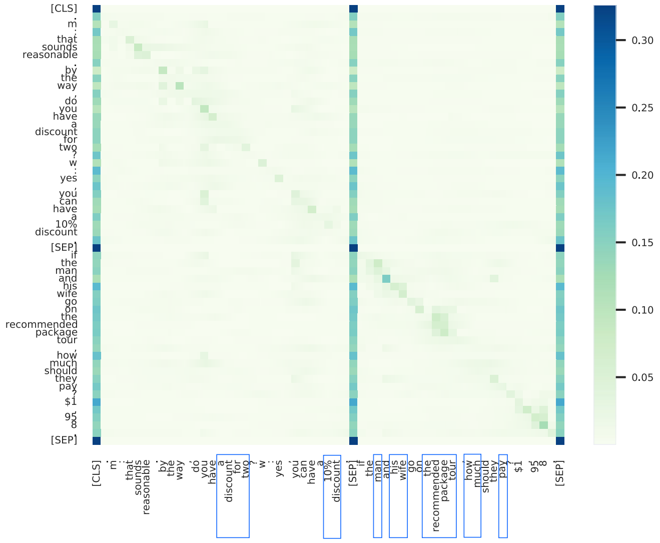

Fig. 4. Attention heatmap of the ALBERT$_{\text{base}}$ baseline.

TABLE XI
BI-DIRECTIONAL VS. UNI-DIRECTIONAL ATTENTIONS ON DREAM DATASET

model	dev	test	avg
ALBERT$_{\text{base}}$	64.51	64.43	64.47
P-to-Q	64.61 (+0.10)	64.72 (+0.29)	64.67 (+0.20)
Q-to-P	66.76 (+2.25)	66.29 (+1.86)	66.53 (+2.06)
Both(DUMA)	67.06 (**+2.55**)	67.56 (**+3.13**)	67.31 (**+2.84**)

G. *Cooperation With PrLMs*

Although the proposed DUMA is supposed to enhance state-of-the-art PrLMs such as ALBERT and ELECTRA, we claim that it is generally effective for less advanced models. Thus,

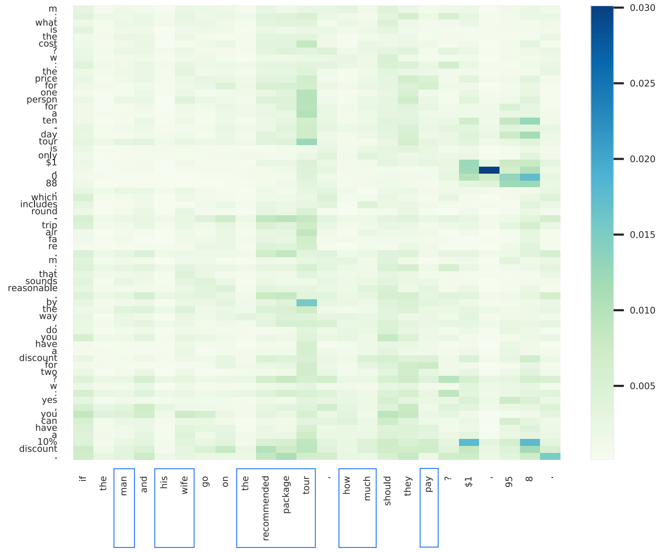

Fig. 5. Attention heatmap of the ALBERT$_{\text{base}}$ baseline passage-to-question part.

TABLE XII
RESULTS USING BERT AS ENCODER ON DREAM DATASET. RESULTS OF BERT$_{\text{base}}$ ARE OUR RE-RUNNING

model	dev	test	avg
ALBERT$_{\text{base}}$	64.51	64.43	64.47
+DUMA	67.06 (+2.55)	67.56 (+3.13)	67.31 (+2.84)
BERT$_{\text{base}}$	61.18	61.54	61.36
+DUMA	64.82 (+3.64)	64.03 (+2.49)	64.43 (+3.07)

TABLE XIII
RESULTS WITH AND WITHOUT SELF-ATTENTION ON DREAM DATASET. "SA" MEANS SELF-ATTENTION AND "CA" MEANS CO-ATTENTION. "SA+CA" MEANS STRAIGHTFORWARDLY USING CA TO REPLACE SA IN ALBERT

model	dev	test
ALBERT$_{\text{base}}$ (SA)	64.51	64.43
ALBERT$_{\text{base}}$ (SA) + DUMA (CA)	**67.06**	**67.56**
ALBERT$_{\text{base}}$ (SA+CA)	41.18	40.08

we simply replace the adopted ALBERT with its early variant BERT to examine the effectiveness of DUMA. Table XII shows the results. We see that our model can be easily transferred to other PrLMs; thus, it can be seen to be an effective module for modeling relationships among the passage, question, and answer for multi-choice MRC.

H. Effect of Self-Attention

Our overall architecture can be split into two steps from the standpoint of attention, of which the first is self-attention (ALBERT) and the second is co-attention (DUMA). To examine whether the structure can be further simplified, that is, where only co-attention is used, we straightforwardly change all of the Multi-head Self-attention of the ALBERT model to our Dual Multi-head Co-attention while still using its pre-trained parameters. The results are shown in Table XIII, indicating that putting co-attention directly into the ALBERT model may lead to much poorer performance compared to the original ALBERT and our ALBERT+DUMA integration method. To conclude, a better way for modeling is our PrLM plus DUMA model, where one first builds a global relationship using the self-attention of the well-trained encoder and then further enhances the relationship between passage and question-answer and distills more matching information using co-attention.

TABLE XIV
RESULTS OF ANSWER OPTION COMPARISON MODULE ON DREAM DATASET

model	dev	test	avg
ALBERT$_{\text{base}}$	64.51	64.43	64.47
+comparison	65.20 (+0.69)	64.72 (+0.29)	64.96 (+0.49)
+DUMA	67.06 (+2.55)	67.56 (+3.13)	67.31 (+2.84)
+comp.&DUMA	67.11 (+2.6)	67.66 (+3.23)	67.39 (+2.92)

I. Answer Options Comparison

According to [38], comparing different answer options helps select the best option. Though our work aims at studying the interaction among question, passage, and answer options, we further study the significance of answer option comparison based on powerful enough models. Thus, we employ an answer option comparison module upon our model and conduct experiments.

1) Module Details: For a set of answer options, we adopt the self-attention mechanism to model the interaction among them. The architecture is the same as the Transformer Block in [33], which is multi-head self-attention followed by a feed-forward neural network, in addition to residual connection and layer normalization, as shown in Fig. 3. Positional encoding is not necessary because only the contents of the answer options are needed. Thus, a set of representations of answer options O_i from (2) are input into this module to obtain new representations after interaction among them for predicting the final answer.

2) Results and Analysis: We study the effect of the answer option comparison module based on ALBERT-base or DUMA on the DREAM dataset. As shown in Table XIV, the module is able to boost ALBERT-base with slight improvements (+0.49), which is much less than that for DUMA (+2.84). In addition, for DUMA, the performance is quite comparable with or without the answer comparison module (only a 0.08 performance difference). In fact, interaction among answer options may be more significant in cases where obvious conflicts exist among different answer options while enhancing interaction among question, passage, and answer options, which DUMA focuses on, is a more general method.

VII. UNDERSTANDING DUMA BY CASES

Table XV shows a hard example that needs to capture important relationships and matching information. Benefiting from well-modeled relationship representations, DUMA can better distill important matching information between the passage and question-answer.

To understand why DUMA can model matching information better than self-attention, we visualize and compare the attention heatmap results. The heatmap figures show different importance of words in the x-axis to a given word in y-axis,

TABLE XV
PREDICTIONS OF DIFFERENT MODELS WHICH ARE SAME AS IN TABLE VIII. "SF ATT" MEANS SOFT ATTENTION

<table>
<tr><td>Passage</td><td colspan="4">M: <i>Could you give me some information on your European tours?</i>
W: <i>Our pleasure. We have several package tours you may choose, from ten days to three weeks in Europe.</i>
M: <i>I would be interested in a ten-day trip around Christmas time.</i>
W: <i>I have one ten-day tour that is still available. It will depart from New York on December 24.</i>
M: <i>What is the cost?</i>
W: <i>The price for one person for a ten-day tour is only $1,088, which includes round-trip airfare.</i>
M: <i>That sounds reasonable. By the way, do you have a discount for two?</i>
W: <i>Yes, you can have a 10% discount.</i></td></tr>
<tr><td>Question</td><td colspan="4"><i>If the man and his wife go on the recommended package tour, how much should they pay?</i></td></tr>
<tr><td>Answer options</td><td colspan="4">1) <i>$1,088</i>
2) <i>$1,958</i> √
3) <i>$2,176</i></td></tr>
<tr><td>ALBERT</td><td>+BiDAF</td><td>+Sf Att</td><td>+DCMN</td><td>+DUMA</td></tr>
<tr><td>Prediction</td><td>3)</td><td>1)</td><td>1)</td><td>2)√</td></tr>
</table>

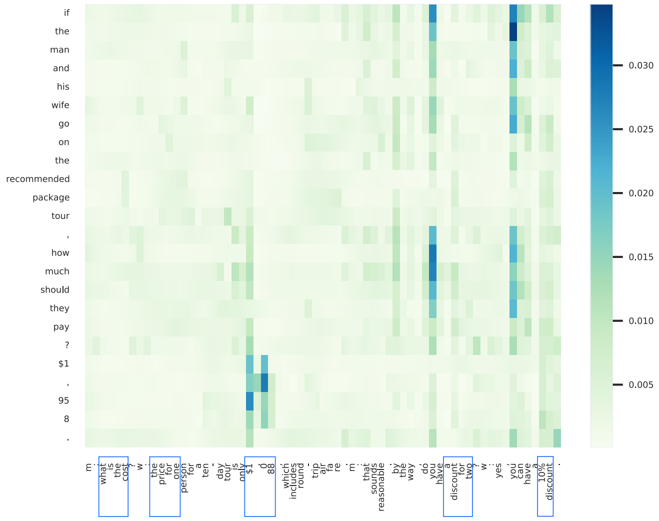

Fig. 6. Attention heatmap of the ALBERT$_{\text{base}}$ baseline question-to-passage part.

and important words in x-axis are highlighted. Fig. 4 indicates that for self-attention, a word tends to attend to itself and its adjacent words, and special tokens "[CLS]" and "[SEP]" always have higher attention weights. Following [7], [15], this feature comes from BERT's pre-training tasks. In addition, we split Fig. 4 into Figs. 5 and 6 to show the passage-to-question part and question-to-passage part, respectively, for comparison with DUMA's dual co-attention. The visualization results indicate that the self-attention of ALBERT may not sufficiently capture the matching information.

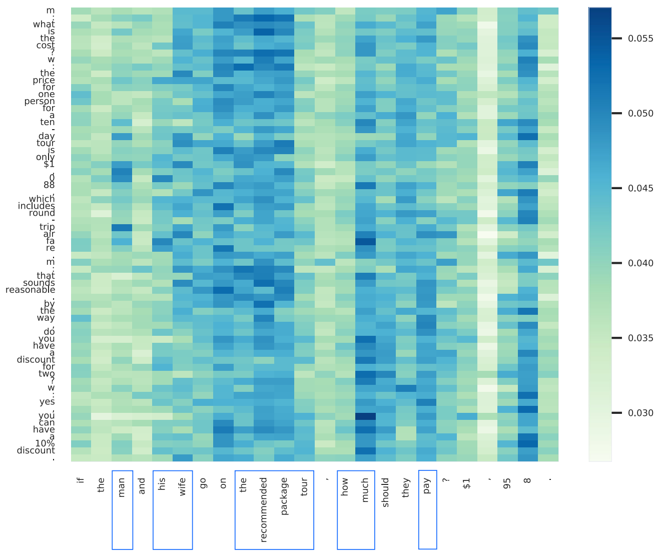

Fig. 7. Attention heatmap of the DUMA passage-to-question part.

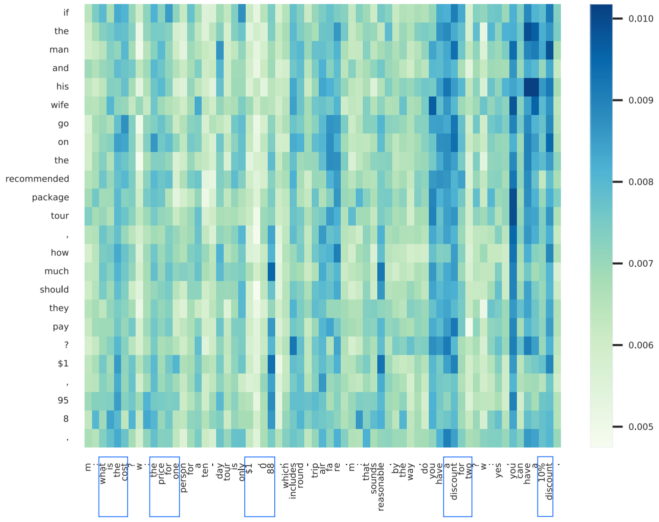

Fig. 8. Attention heatmap of the DUMA question-to-passage part.

However, as shown in Figs. 7 and 8, DUMA adopts bidirectional co-attention, which focuses on more fine-grained matching of vital information. In Fig. 7 of passage-to-question attention, *man*, *his wife*, *the recommended package tour*, *how much* and *pay* take up the main attention weights, and in Fig. 8 of question-to-passage attention, *what is the cost*, *the price for one*, *$1,088*, *a discount for two* and *10% discount* are more important. This helps the model capture more fine-grained matching information, which confirms our original motivation, that is, to better model interaction behavior among question, passage, and answer options based on PrLMs, to accomplish "re-consideration".

VIII. CONCLUSION

In this paper, we simulate the human transposition thinking experience when solving MRC problems and propose **DU**al **M**ulti-head Co-**A**ttention (DUMA) to model the relationships among passage, question, and answer for multi-choice MRC tasks. Our method is able to cooperate with popular large-scale Pre-trained Language Models and leads to effective performance improvements. In addition, we investigate previous attention mechanisms and matching networks applied on top of PrLMs. In this context, our model is shown to be optimal through extensive experiments. In particular, it achieves the best performance with an intuitively motivated design structure. Our proposed DUMA enhancement has been verified to be effective on two benchmark multi-choice MRC tasks, DREAM and RACE, with higher performance being achieved over strong PrLM baselines.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[2] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *CoRR*, 2020, *arXiv:2001.01582*.

[3] GP Shrivatsa *et al.*, "Translucent answer predictions in multi-hop reading comprehension," *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 7700–7707.

[4] C. Chelba *et al.*, "One billion word benchmark for measuring progress in statistical language modeling," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2635–2639. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_2635.html

[5] Z. Chen, Y. Cui, W. Ma, S. Wang, and G. Hu, "Convolutional spatial attention model for reading comprehension with multiple-choice questions," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 6276–6283, 2019.

[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, *arXiv:2003.10555*.

[7] Y. Cui, W.-N. Zhang, W. Che, T. Liu, and Z. Chen, "Understanding attention in machine reading comprehension," *CoRR*, 2021, *arXiv:2108.11574*.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. Conf. North Amer. Chapter Assoc. Computat. Linguistics, Human Lang. Technol.*, vol. 1, pp. 4171–4186, Jun. 2019.

[9] Y. Gao, L. Bing, P. Li, I. King, and M. R. Lyu, "Generating distractors for reading comprehension questions from real examinations," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 6423–6430, 2019.

[10] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowl. Based Syst.*, vol. 212, 2019, doi: 10.1016/j.knosys.2020.106622.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[12] Y. Jiang *et al.*, "Improving machine reading comprehension with single-choice decision and transfer learning," *CoRR*, 2020, *arXiv:2011.03292*.

[13] D. Jin, S. Gao, J.-Y. Kao, T. Chung, and D. Hakkani-Tur, "MMM: Multi-stage multi-task learning for multi-choice reading comprehension," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 8010–8017.

[14] H. Kim and P. Fung, "Learning to classify the wrong answers for multiple choice question answering (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13843–13844.

[15] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, 2019, pp. 4365–4374. doi: 10.18653/v1/D19-1445.

[16] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale reading comprehension dataset from examinations," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 785–794.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.

[18] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, 2019, *arXiv:1907.11692*.

[19] M. Peters *et al.*, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 2227–2237.

[20] M. Peters *et al.*, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA, Jun. 2018b, pp. 2227–2237. doi: 10.18653/v1/N18-1202.

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," OpenAI., Tech. Rep., 2018.

[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000 questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.

[24] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, Australia, 2018, pp. 784–789, doi: 10.18653/v1/P18-2124.

[25] Q. Ran, P. Li, W. Hu, and J. Zhou, "Option comparison network for multiple-choice reading comprehension," *CoRR*, 2019, *arXiv:1903.03033*.

[26] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[27] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," *CoRR*, 2019, *arXiv:1909.08053*.

[28] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, "DREAM: A challenge dataset and models for dialogue-based reading comprehension," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 217–231, 2019. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1534

[29] K. Sun, D. Yu, D. Yu, and C. Cardie, "Improving machine reading comprehension with general reading strategies," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 2633–2643.

[30] M. Tang, J. Cai, and H. H. Zhuo, "Multi-matching network for multiple choice reading comprehension," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 7088–7095.

[31] Y. Tay, A. L. Tuan, and S. C. Hui, "Multi-range reasoning for machine comprehension," *CoRR*, 2018, *arXiv:1803.09074*.

[32] M. Tu, K. Huang, G. Wang, J. Huang, X. He, and B. Zhou, "Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 9073–9080.

[33] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[34] H. Wan, "Multi-task learning with multi-head attention for multi-choice reading comprehension," *CoRR*, 2020, *arXiv:2003.04992*.

[35] S. Wang, M. Yu, S. Chang, and J. Jiang, "A co-matching model for multi-choice reading comprehension," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 746–751.

[36] M. Yan *et al.*, "A deep cascade model for multi-document reading comprehension," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 7354–7361, 2019.

[37] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst. 32*, pp. 5754–5764, Dec. 2019.

[38] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, and X. Zhou, "DCMN : Dual co-matching network for multi-choice reading comprehension," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 9563–9570.

[39] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu, "Modeling multi-turn conversation with deep utterance aggregation," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3740–3752.

[40] H. Zhu, F. Wei, B. Qin, and T. Liu, "Hierarchical attention flow for multiple-choice reading comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6077–6085.

[41] Y. Zhu *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 19–27. doi: 10.1109/ICCV.2015.11.

Pengfei Zhu is currently working toward the master's degree with the Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University, Shanghai, China. He majors in computer science and is expected to get the master's degree in 2022. His research interests include natural language processing, machine reading comprehension, and open-domain question answering.

Zhuosheng Zhang received the bachelor's degree in the Internet of Things from Wuhan University, Wuhan, China, in 2016, and the M.S. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently working toward the Ph.D. degree in computer science with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University. From 2019 to 2020, he was an Internship Research Fellow with National Institute of Information and Communications Technology, Japan. His research interests include natural language processing, machine reading comprehension, dialogue systems, and machine translation.

Hai Zhao received the B.Eng. degree in sensor and instrument engineering and the M.Phil. degree in control theory and engineering from Yanshan University, Qinhuangdao, China, in 1999 and 2000, respectively, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2005. He was a Research Fellow with the City University of Hong Kong, Hong Kong, from 2006 to 2009, a Visiting Scholar with Microsoft Research Asia in 2011, and a Visiting Expert with National Institute of Information and Communications Technology, Japan, in 2012. Since 2009, he has been with Shanghai Jiao Tong University, where he is currently a Full Professor with the Department of Computer Science and Engineering. He is an ACM Professional Member and was an Area Co-Chair with ACL 2017, on Tagging, Chunking, Syntax, and Parsing, Senior Area Chairs in ACL 2018, 2019 on Phonology, Morphology and Word Segmentation. His research interests include natural language processing and related machine learning, data mining, and artificial intelligence.

Xiaoguang Li received the bachelor's and master's degrees from Wuhan University, Wuhan, China, and is employed with Huawei Noah's Ark Lab. His main research interests include natural language processing, information retrieval, question answering and dialogue systems.