Dual Co-Matching Network for Multi-choice Reading Comprehension

Abstract

Multi-choice reading comprehension is a challenging task that requires complex reasoning procedure. Given passage and question, a correct answer need to be selected from a set of candidate answers. In this paper, we propose \textbf{D}ual \textbf{C}o-\textbf{M}atching \textbf{N}etwork (\textbf{DCMN}) which model the relationship among passage, question and answer bidirectionally. Different from existing approaches which only calculate question-aware or option-aware passage representation, we calculate passage-aware question representation and passage-aware answer representation at the same time. To demonstrate the effectiveness of our model, we evaluate our model on a large-scale multiple choice machine reading comprehension dataset (i.e. RACE). Experimental result show that our proposed model achieves new state-of-the-art results.

Introduction

Machine reading comprehension and question answering has becomes a crucial application problem in evaluating the progress of AI system in the realm of natural language processing and understanding BIBREF0 . The computational linguistics communities have devoted significant attention to the general problem of machine reading comprehension and question answering.

However, most of existing reading comprehension tasks only focus on shallow QA tasks that can be tackled very effectively by existing retrieval-based techniques BIBREF1 . For example, recently we have seen increased interest in constructing extractive machine reading comprehension datasets such as SQuAD BIBREF2 and NewsQA BIBREF3 . Given a document and a question, the expected answer is a short span in the document. Question context usually contains sufficient information for identifying

evidence sentences that entail question-answer pairs. For example, 90.2% questions in SQuAD reported by Min BIBREF4 are answerable from the content of a single sentence. Even in some multi-turn conversation tasks, the existing models BIBREF5 mostly focus on retrieval-based response matching.

In this paper, we focus on multiple-choice reading comprehension datasets such as RACE BIBREF6 in which each question comes with a set of answer options. The correct answer for most questions may not appear in the original passage which makes the task more challenging and allow a rich type of questions such as passage summarization and attitude analysis. This requires a more in-depth understanding of a single document and leverage external world knowledge to answer these questions. Besides, comparing to traditional reading comprehension problem, we need to fully consider passage-question-answer triplets instead of passage-question pairwise matching.

In this paper, we propose a new model, Dual Co-Matching Network, to match a question-answer pair to a given passage bidirectionally. Our network leverages the latest breakthrough in NLP: BERT BIBREF7 contextual embedding. In the origin BERT paper, the final hidden vector corresponding to first input token ([CLS]) is used as the aggregation representation and then a standard classification loss is computed with a classification layer. We think this method is too rough to handle the passage-question-answer triplet because it only roughly concatenates the passage and question as the first sequence and uses question as the second sequence, without considering the relationship between the question and the passage. So we propose a new method to model the relationship among the passage, the question and the candidate answer.

Firstly we use BERT as our encode layer to get the contextual representation of the passage, question, answer options respectively. Then a matching layer is constructed to get the passage-question-answer triplet matching representation which encodes the locational information of the question and the candidate answer matched to a specific context of the passage. Finally we apply a hierarchical

aggregation method over the matching representation from word-level to sequence-level and then from sequence level to document-level. Our model improves the state-of-the-art model by 2.6 percentage on the RACE dataset with BERT base model and further improves the result by 3 percentage with BERT large model.

Model

For the task of multi-choice reading comprehension, the machine is given a passage, a question and a set of candidate answers. The goal is to select the correct answer from the candidates. P, Q, and A are used to represent the passage, the question and a candidate answer respectively. For each candidate answer, our model constructs a question-aware passage representation, a question-aware passage representation and a question-aware passage representation. After a max-pooling layer, the three representations are concatenated as the final representation of the candidate answer. The representations of all candidate answers are then used for answer selection.

In section "Encoding layer" , we introduce the encoding mechanism. Then in section "Conclusions" , we introduce the calculation procedure of the matching representation between the passage, the question and the candidate answer. In section "Aggregation layer" , we introduce the aggregation method and the objective function.

Encoding layer

This layer encodes each token in passage and question into a fixed-length vector including both word embedding and contextualized embedding. We utilize the latest result from BERT BIBREF7 as our encoder and the final hidden state of BERT is used as our final embedding. In the origin BERT BIBREF7 , the procedure of processing multi-choice problem is that the final hidden vector corresponding to first

input token ([CLS]) is used as the aggregation representation of the passage, the question and the candidate answer, which we think is too simple and too rough. So we encode the passage, the question and the candidate answer respectively as follows:

$$\begin{split}
\textbf {H}^p=&BERT(\textbf {P}),\textbf {H}^q=BERT(\textbf {Q}) \\
&\textbf {H}^a=BERT(\textbf {A})
\end{split}$$ (Eq. 3)

where $\textbf {H}^p \in R^{P \times l}$, $\textbf {H}^q \in R^{Q \times l}$ and $\textbf {H}^a \in R^{A \times l}$ are sequences of hidden state generated by BERT. P , Q , A are the sequence length of the passage, the question and the candidate answer respectively. l is the dimension of the BERT hidden state.

Matching layer

To fully mine the information in a {P, Q, A} triplet , We make use of the attention mechanism to get the bi-directional aggregation representation between the passage and the answer and do the same process between the passage and the question. The attention vectors between the passage and the answer are calculated as follows:

$$\begin{split}
\textbf {W}&=SoftMax(\textbf {H}^p({H^{a}G + b})^T), \\
\textbf {M}^{p}&=\textbf {W}\textbf {H}^{a}, \\
\textbf {M}^{a}=\textbf {W}^T\textbf {H}^{p}, \\
\end{split}$$ (Eq. 5)

where $G \in R^{l \times l}$ and $b \in R^{A \times l}$ are the parameters to learn. $\textbf {W} \in R^{P \times A}$ is the attention weight matrix between the passage and the answer. $\textbf {M}^{p} \in R^{P \times l}$ represent how each hidden state in passage can be aligned to the answe rand $\textbf {M}^{a} \in R^{A \times l}$ represent how the candidate answer can be aligned to each hidden state in passage. In the same method, we can get $\textbf {W}^{\prime } \in R^{P \times Q}$ and $\textbf {M}^{q} \in R^{Q \times l}$ for the representation between the passage and the question.

To integrate the original contextual representation, we follow the idea from BIBREF8 to fuse $\textbf {M}^{a}$ with original $\textbf {H}^p$ and so is $\textbf {M}^{p}$. The final representation of passage and the candidate answer is calculated as follows:

$$\begin{split}
\textbf {S}^{p}&=F([\textbf {M}^{a} - \textbf {H}^{a}; \textbf {M}^{a} \cdot \textbf {H}^{a}]W_1 + b_1),\\
\textbf {S}^{a}&=F([\textbf {M}^{p} - \textbf {H}^{p}; \textbf {M}^{p} \cdot \textbf {H}^{p}]W_2 + b_2),\\
\end{split}$$ (Eq. 6)

where $W_1, W_2 \in R^{2l \times l}$ and $b_1 \in R^{P \times l}, b_2 \in R^{(A) \times l}$ are the parameters to learn. $[;]$ is the column-wise concatenation and $-, \cdot $ are the element-wise subtraction and multiplication between two matrices. Previous work in BIBREF9 , BIBREF10 shows this method can build better matching representation. F is the activation function and we choose $ReLU$ activation function there. $\textbf {S}^{p} \in R^{P \times l}$ and $\textbf {S}^{a} \in R^{A \times l}$ are the final representations of the passage and candidate answer. In the question side, we can get $\textbf {S}^{p^{\prime }} \in R^{P \times l}$ and $\textbf {S}^{q} \in R^{Q \times l}$ in the same calculation method.

Aggregation layer

To get the final representation for each candidate answer, a row-wise max pooling operation is used to $\textbf {S}^{p}$ and $\textbf {S}^{a}$. Then we get $\textbf {C}^{p} \in R^l$ and $\textbf {C}^{a} \in R^l$ respectively. In the question side, $\textbf {C}^{p^{\prime }} \in R^l$ and $\textbf {C}^{q} \in R^l$ are calculated. Finally, we concatenate all of them as the final output $\textbf {C} \in R^{4l}$ for each {P, Q, A} triplet.

$$\begin{split}
\textbf {C}^{p} = &Pooling(\textbf {S}^{p}),
\textbf {C}^{a} = Pooling(\textbf {S}^{a}),\\
\textbf {C}^{p^{\prime }} = &Pooling(\textbf {S}^{p^{\prime }}),
\textbf {C}^{q} = Pooling(\textbf {S}^{q}),\\
\textbf {C} &= [\textbf {C}^{p}; \textbf {C}^{a};\textbf {C}^{p^{\prime }};\textbf {C}^{q}]
\end{split}$$ (Eq. 9)

For each candidate answer choice i , its matching representation with the passage and question can be represented as $\textbf {C}_i$. Then our loss function is computed as follows:

$$\begin{split}
L(\textbf {A}_i|\textbf {P,Q}) = -log{\frac{exp(V^T\textbf {C}_i)}{\sum _{j=1}^N{exp(V^T\textbf {C}_j)}}},
\end{split}$$ (Eq. 10)

where $V \in R^l$ is a parameter to learn.

Experiment

We evaluate our model on RACE dataset BIBREF6 , which consists of two subsets: RACE-M and

RACE-H. RACE-M comes from middle school examinations while RACE-H comes from high school examinations. RACE is the combination of the two.

We compare our model with the following baselines: MRU(Multi-range Reasoning) BIBREF12 , DFN(Dynamic Fusion Networks) BIBREF11 , HCM(Hierarchical Co-Matching) BIBREF8 , OFT(OpenAI Finetuned Transformer LM) BIBREF13 , RSM(Reading Strategies Model) BIBREF14 . We also compare our model with the BERT baseline and implement the method described in the original paper BIBREF7 , which uses the final hidden vector corresponding to the first input token ([CLS]) as the aggregate representation followed by a classification layer and finally a standard classification loss is computed.

Results are shown in Table 2 . We can see that the performance of BERT $_{base}$ is very close to the previous state-of-the-art and BERT $_{large}$ even outperforms it for 3.7%. But experimental result shows that our model is more powerful and we further improve the result for 2.2% computed to BERT $_{base}$ and 2.2% computed to BERT $_{large}$.

Conclusions

In this paper, we propose a Dual Co-Matching Network, DCMN, to model the relationship among the passage, question and the candidate answer bidirectionally. By incorporating the latest breakthrough, BERT, in an innovative way, our model achieves the new state-of-the-art in RACE dataset, outperforming the previous state-of-the-art model by 2.2% in RACE full dataset.