Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Zhipeng Chen,[†] Yiming Cui,[‡†*] Wentao Ma,[†] Shijin Wang,[†] Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
{zpchen,ymcui,wtma,sjwang3,gphu}@iflytek.com

Abstract

Machine Reading Comprehension (MRC) with multiple-choice questions requires the machine to read given passage and select the correct answer among several candidates. In this paper, we propose a novel approach called Convolutional Spatial Attention (CSA) model which can better handle the MRC with multiple-choice questions. The proposed model could fully extract the mutual information among the passage, question, and the candidates, to form the enriched representations. Furthermore, to merge various attention results, we propose to use convolutional operation to dynamically summarize the attention values within the different size of regions. Experimental results show that the proposed model could give substantial improvements over various state-of- the-art systems on both RACE and SemEval-2018 Task11 datasets.

Introduction

Owing to the rapid release of various large-scale datasets, Machine Reading Comprehension (MRC) has become enormously popular in Natural Language Processing. For example, cloze-style MRC (such as CNN/DailyMail (Hermann et al. 2015), Children's Book Test (CBT) (Hill et al. 2015)), span-extraction MRC (such as SQuAD (Rajpurkar et al. 2016)), and multiple-choice MRC (such as MCTest (Richardson, Burges, and Renshaw 2013), RACE (Lai et al. 2017), SemEval-2018 Task11 (Ostermann et al. 2018)).

In this paper, we mainly focus on solving the reading comprehension with multiple-choice questions. At the beginning of the reading comprehension study, this type of reading comprehension task was not that popular because there is no large-scale dataset available and thus we cannot apply neural network approaches to solve them. To bring more challenges to reading comprehension task and mitigate the absence of large-scale multi-choice reading comprehension dataset, Lai et al.(2017) propose a new dataset called RACE. Compared to the earlier MCTest (Richardson, Burges, and Renshaw 2013), the RACE dataset is made from the English examinations for Chinese middle and high school students, consisting near 100,000 questions generated by human experts, and is far more challenging than the MCTest.

*Corresponding author
Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this paper, we propose a novel model called Convolutional Spatial Attention (CSA) to fully utilize the hierarchical attention information for reading comprehension with multiple-choice questions. The proposed model first encode the passage, question, and candidates into word representations which are enhanced by the additional POS-tag and matching features. Then we concentrate on enriching the representation of the candidates by incorporating the passage, question information, and further calculate the attentions between the passage, question, and candidates, forming the spatial attentions. To further extract the representative features in the spatial attentions, we propose to use convolutional neural network to dynamically conclude adjacent regions with different window size. We mainly test our CSA model on two multiple-choice reading comprehension datasets: RACE and SemEval-2018 Task11, and our model achieves state-of-the-art performances on both of them. The examples of each dataset are given in Figure 1. The main contributions of our paper can be summarized as follows.

- We focus on modeling different semantic aspects of the candidates, by integrating the passage and question information, forming the 3D spatial attention among the passage, question, and candidates.
- We propose a Convolutional Spatial Attention (CSA) mechanism to dynamically extract representative features from the spatial attentions.
- The proposed model gives substantial improvements over various state-of-the-art systems on both RACE and SemEval-2018 Task11 datasets, showing its generalization and extensibility to other NLP tasks.

Related Works

Massive progress has been made on machine reading comprehension field in recent years. The booming of the MRC can trace back to the release of the large-scale datasets, such as CNN/DailyMail (Hermann et al. 2015) and CBT (Hill et al. 2015)). After the release of these datasets, various neural network approaches (Chen, Bolton, and Manning 2016; Kadlec et al. 2016; Cui et al. 2017; Dhingra et al. 2017) have been proposed and become fundamental components in the future studies. Another representative dataset is SQuAD (Rajpurkar et al. 2016), which was difficult than the cloze-style reading comprehension and requires the machine to

RACE	SemEval-2018 Task11
Passage Is it important to have breakfast every day? A short time ago, a test was given in the United States. People of different ages, from 12 to 83, were asked to have a test. During the test, these people were given all kinds of breakfast, and sometimes they got no breakfast at all. ...	**Passage** I was thirsty so I decided to make a cup of tea. I looked through my box of teas and rifled through the assorted flavors. I settled on Earl Gray, which is a black tea flavored with bergamot orange. I filled the kettle with water and placed it on the stove, turning on the burner so that it would heat up and begin boiling. ...
Question What do the results show?	**Question** Why did they use a kettle?
Candidates **A *They show that breakfast has affected on work and study.*** **B** Breakfast has little to do with a person's work. **C** A person will work better if he only has fruit and milk. **D** They show that girl students should have less for breakfast.	**Candidates** **A** to drink from **B *to boil water***

Figure 1: Example of RACE and SemEval-2018 Task11 dataset. The correct answer is depicted in bold face.

generate a span in the passage to answer the questions. With rapid progress on designing effective neural network models (Xiong, Zhong, and Socher 2016; Seo et al. 2016; Wang et al. 2017; Hu, Peng, and Qiu 2017), recent works on this datasets have surpassed the average human performance, such as QANet (Yu et al. 2018) etc.

However, current machine reading comprehension models are still struggling with solving the questions that need reasoning over multiple sentences or even passage. To solve the reading comprehension with multiple-choice questions, various approaches have been proposed, and most of them are focusing on designing effective attentions or persuing enriched representations for prediction. When releasing the RACE dataset, Lai et al.(2017) also adopted and modified two models of the cloze-style reading comprehension: Gated Attention Reader (Dhingra et al. 2017) and Stanford Attentive Reader (Chen, Bolton, and Manning 2016). However, experimental results show that these models are not capable of this task. Parikh et al.(2018) introduced ElimiNet which use a combination of elimination and selection to get refined representation of the candidates. Xu et al.(2017) proposed the Dynamic Fusion Networks (DFN), which uses multi-hop reasoning mechanism for this task. Zhu et al.(2018) proposed the Hierarchical Attention Flow model, which leverage candidate options to model the interactions among passage, questions, and candidates.

Though various efforts have been made, we believe there is still a large room for designing effective neural networks for better characterizing the relations between the passage, questions, and candidates. To this end, we summarize the main differences between our model and existing models for this task into three aspects. First, we calculate various representations of the candidate to better characterize it for prediction. Second, when calculating attention, we apply additional trainable weights to dynamically adjust the attention values, which is more flexible. Third, unlike the previous works only utilize the final-level hierarchical attentions, we propose to use every attention in each hierarchy, and use convolutional neural network to capture features for predicting the answer.

Convolutional Spatial Attention Model

Task Definition

The RACE dataset (Lai et al. 2017) for reading comprehension with multiple-choice questions was proposed, which consists of 28,000+ passages and near 100,000 questions generated by human experts for the English examinations of Chinese middle and high school students. Different from the earlier MCTest dataset (Richardson, Burges, and Renshaw 2013), the RACE dataset is significantly larger, and thus we can apply deep learning approaches for this task. As all the questions and choices are generated by human experts, RACE dataset provides more comprehensive and realistic evaluation on machine reading comprehension than the other popular datasets such as CNN/DailyMail, SQuAD datasets, whose answer should appear in context. Also according to the analysis by Lai et al.(2017), a large portion of the questions in RACE need reasoning over various clues, which makes it more challenging and suitable to evaluate the ability of the reading comprehension systems. SemEval-2018 Task11 is closely the same with the RACE dataset but with two candidates and small size. The examples of each dataset are given in Figure 1.

The Model

In this section, we will give a detailed description on the proposed model. The main neural architecture of our model is depicted in Figure 2. Throughout this section, we will use P for representing the passage, Q for the question, C for the candidates. Note that, as the operations on each candidate are the same, for simplicity, we only take one of the candidates for illustration.

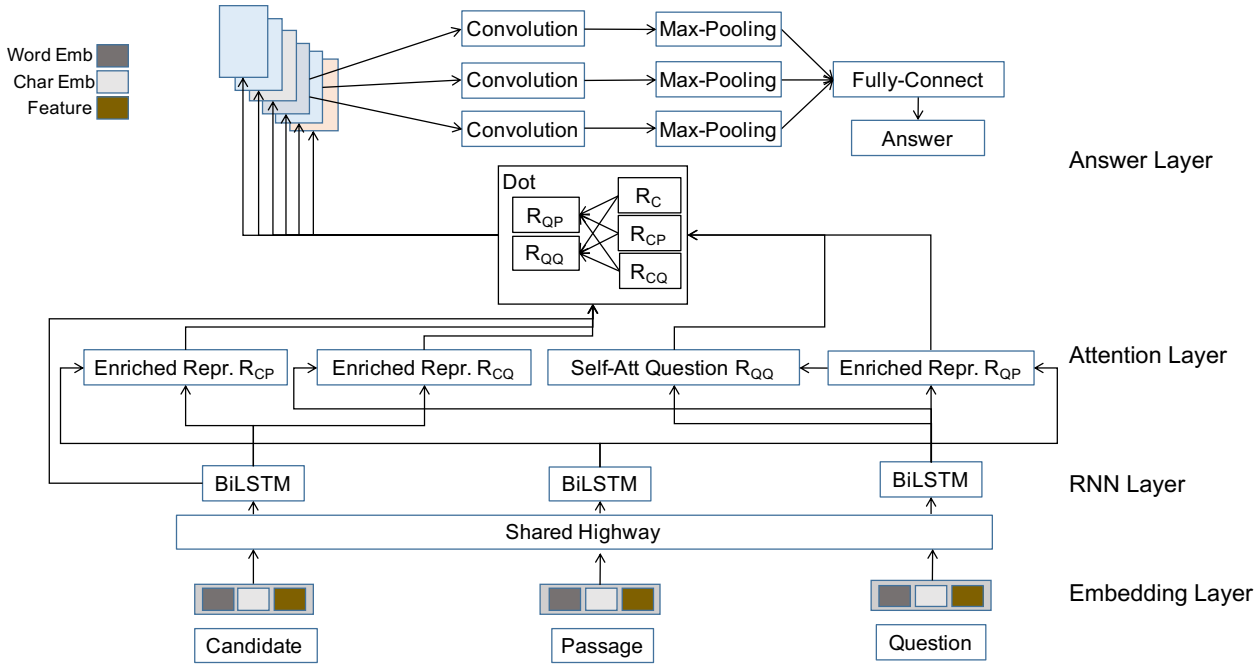

Figure 2: Main neural architecture of the Convolutional Spatial Attention (CSA) model.

Word Representation

We transform each word in the passage, question, and candidates into continuous representations. In this paper, there are three components in the embedding layer, which can be listed as follows.

- **Word Embedding** E_{word}: We use traditional pre-trained GloVe embedding for initialization (Pennington, Socher, and Manning 2014) and keep fixed during the training process.
- **ELMo Embedding** E_{elmo}: For this part, we use pre-trained ELMo (Peters et al. 2018) embedding.
- **Feature Embedding** E_{feat}: We also utilize three additional features to enhance the word representations.
  1. **POS-tag Embedding** E_{pos}: We use NLTK (Bird and Loper 2004) for part-of-speech tagging for each word. Similar to traditional word embeddings, we assign different trainable vectors for each part-of-speech tag.
  2. **Word Matching** F_{match}: Take the text as an example, if the word in text also appears in question or candidate, we set the value as one, otherwise set it as zero. In this way, we can also add this feature to the question and candidate.
  3. **Fuzzy Word Matching** F_{fuzzy}: Similar to the word matching feature, but we loosen the matching criteria as partial matching. For example, we regard 'teacher' and 'teach' as fuzzy matching, because the string 'teacher' is partially matched by 'teach'.

We concatenate three embedding components to form the final word representations for the text $E_P \in \mathbb{R}^{|P| \times e}$, question $E_Q \in \mathbb{R}^{|Q| \times e}$, and candidates $E_C \in \mathbb{R}^{|C| \times e}$, where $|P|, |Q|, |C|$ are the length of the passage, question, and candidates, e is the final embedding size (including all three components).

$$E = [E_{word}; E_{elmo}; E_{feat}] \quad (1)$$

$$E_{feat} = [E_{pos}; F_{match}; F_{fuzzy}] \quad (2)$$

After obtaining embedding representations, we further feed each word embedding into a shared highway network (Srivastava, Greff, and Schmidhuber 2015) with $\tanh$ output activation (denoted as σ). In this paper, we apply two consecutive highway networks with shared weights. Then we use Bi-Directional LSTM (Graves and Schmidhuber 2005) to model the contextual information, forming $H_P \in \mathbb{R}^{|P| \times h}$, $H_Q \in \mathbb{R}^{|Q| \times h}$, and $H_C \in \mathbb{R}^{|C| \times h}$ (h is hidden size of Bi-LSTM). Note that, we use different Bi-LSTMs for the passage, question, and candidates.

$$\widetilde{H} = \sigma(\text{2-Highway}(E)) \quad (3)$$

$$H = \text{Bi-LSTM}(\widetilde{H}) \quad (4)$$

Enriched Representation

Calculating attention and generating enriched representation play very important roles in machine reading comprehension. In our model, we will calculate various types of enriched representations for better characterizing the candidate and question, which are the essential components in this task. The procedure for generating enriched representation is illustrated in Algorithm 1.

For example, we wish to embed the passage information into the candidate representation to better aware the relevant part in the passage and obtain the passage-aware candidate

Algorithm 1 Enriched Representation.

Input:
Time-Distributed representation X_1
Time-Distributed representation X_2
Initialize:
Random weight matrix $W_1 \in \mathbb{R}^{h \times h_{att}}$
Random weight matrix $W_2 \in \mathbb{R}^{h \times h_{att}}$
Diagonal weight matrix $D \in \mathbb{R}^{h_{att} \times h_{att}}$
All-one weight matrix $W \in \mathbb{R}^{|X_1| \times |X_2|}$
Output: X_2-aware X_1 representation Y
1: Calculate attention matrix $M' \in \mathbb{R}^{|X_1| \times |X_2|}$: $M' = f(W_1 X^1)^T \cdot D \cdot f(W_2 X^2)$
2: Apply element-wise weight: $M = M' \odot W$
3: Apply softmax function to the last dimension of M: $M_{att} = softmax(M)$
4: Calculate raw representation $Y' \in \mathbb{R}^{|X_2| \times h}$: $Y' = {M_{att}}^T \cdot X_1$
5: Concatenate raw representation Y' and raw input X_1, then apply Bi-LSTM: $Y =$ Bi-LSTM$([X_1; Y'])$
6: **return** Y

representation R_{CP}. According to Algorithm 1, R_{CP} can be generated as follows.

- **[Line 1]** Given the Bi-LSTM representations of passage H_P and candidate H_C, we first calculate the attention matrix where each element indicate the matching information between them. In this paper, we adopt the attention mechanism used in FusionNet (Huang et al. 2017). where two representations are transformed by individual fully-connected layer with an output activation f. Also, a trainable diagonal weight matrix D is applied. The activation function f is defined as RELU throughout this paper.
- **[Line 2]** Then we apply an element-wise weight matrix W to the attention matrix M'. This is designed to let the model flexibly adjust the attention values.
- **[Line 3]** We apply *softmax* to the weighted attention matrix M to the last dimension of it, which calculates the passage-level attention vector w.r.t. each candidate word.
- **[Line 4]** After obtaining the normalized attention matrix M_{att}, we make a dot product between the M_{att} and the passage H_P to extract candidate-related passage.
- **[Line 5]** Finally, we concatenate candidate-related passage Y'_{CP} and Bi-LSTM candidate representation H_C, and feed them to a Bi-LSTM to fully integrate passage information into the candidate representations.

By applying the proposed algorithm $g(X_1, X_2)$, we can calculate the passage-aware and question-aware candidate representation R_{CP} and R_{CQ}. Besides, as the question information is also important, we also calculate passage-aware question representation R_{QP} and self-attended question representation $R_{self\text{-}Q}$. Note that, we use Bi-LSTM output B_Q and passage-aware question representation R_{QP} to obtain the (almost-)self-attended question representation, which combines different levels of representations.

$$R_{CQ} = g(B_C, B_Q) \tag{5}$$
$$R_{CP} = g(B_C, B_P) \tag{6}$$
$$R_{QP} = g(B_Q, B_P) \tag{7}$$
$$R_{self\text{-}Q} = g(B_Q, R_{QP}) \tag{8}$$

Convolutional Spatial Attention

With previously generated representations, we can calculate the matching matrix to measure the similarity between them. In this paper, we adopt simple dot product to obtain the matching matrix. As the candidate information is important to answer the question, firstly we calculate the matching matrix using various candidate representations to the self-attended question representation $R_{self\text{-}Q}$. The motivation is to use the question information as the key to extracting candidate information in different levels. We use question-aware candidate representation R_{CQ}, passage-aware candidate representation R_{CP}, and candidate Bi-LSTM representation H_C, as shown below.

$$M_{11} = R_{CQ} \cdot R_{self\text{-}Q} \tag{9}$$
$$M_{12} = R_{CP} \cdot R_{self\text{-}Q} \tag{10}$$
$$M_{13} = H_C \cdot R_{self\text{-}Q} \tag{11}$$

In a similar way, we can also replace the self-attended question representation $R_{self\text{-}Q}$ with passage-aware question representation R_{QP} in Equation 9, 10, 11, to obtain M_{21}, M_{22}, M_{23}. Then we concatenate all matrices on the channel dimension to form a *Spatial Attention Cube* $M \in \mathbb{R}^{6 \times |C| \times |Q|}$, which is similar to an 'image' with 6 channels.

$$M = [M_{11}; M_{12}; M_{13}; M_{21}; M_{22}; M_{23}] \tag{12}$$

In order to extract high-level features inside the spatial attention cube, we use CNN-MaxPooling operation to dynamically conclude adjacent attention information, which is similar to the traditional operation on the image. Formally, we first apply convolutional operation on the M to summarize different length of adjacent elements We adopt three convolutional kernels: 5, 10, and 15, along with the dimension of the question length.

After convolution operation, we could get three features maps w.r.t. different convolutional kernels. The procedure can be illustrated as the following equations.

$$O_1 = \text{Max-Pooling}_{1\times 3}\{CNN_{1\times 5}(M)\} \tag{13}$$
$$O_2 = \text{Max-Pooling}_{1\times 2}\{CNN_{1\times 10}(M)\} \tag{14}$$
$$O_3 = \text{Max-Pooling}_{1\times 1}\{CNN_{1\times 15}(M)\} \tag{15}$$

In this way, we used CNN-MaxPooling operation to obtain three feature vectors O_1, O_2, O_3 by using different convolutional kernels and max-pooling intervals.

Final Prediction

After obtaining three feature vectors, we flatten, concatenate, and feed them into a fully-connected layer to get a scalar value denoting the possibility of being the correct answer. Recall that, we have several candidates for a given

Model	RACE-M	RACE-H	RACE
Sliding Window (Lai et al. 2017)	37.3	30.4	32.2
Stanford AR (Lai et al. 2017)	44.2	43.0	43.3
GA Reader (Lai et al. 2017)	43.7	44.2	44.1
ElimiNet (Parikh et al. 2018)	N/A	N/A	44.5
Hierarchical Attention Flow (Zhu et al. 2018)	45.0	*46.4*	46.0
Dynamic Fusion Network (Xu et al. 2017)	*51.5*	45.7	*47.4*
CSA Model (single model)	51.0	47.3	48.4
CSA Model + ELMo (single model)	**52.2**	**50.3**	**50.9**
GA Reader (6-ensemble)	-	-	45.9
ElimiNet (6-ensemble)	-	-	46.5
GA + ElimiNet (12-ensemble)	-	-	47.2
Dynamic Fusion Network (9-ensemble)	*55.6*	*49.4*	*51.2*
CSA Model (7-ensemble)	55.2	52.4	53.2
CSA Model + ELMo (9-ensemble)	**56.8**	**54.8**	**55.0**

Table 1: Experimental results on RACE. The best previous results are in italics, and overall best results are in bold face.

question, so we will get N candidate scores in this stage. We apply softmax function to these scores to obtain the final probability distributions over the candidates.

$$s_i = \mathbf{w^T} \cdot [O_1; O_2; O_3] \quad (16)$$

$$Pr(A|P, Q, C) = softmax([s_1; ...; s_N]) \quad (17)$$

To train our model, we use traditional cross entropy loss to minimize the gap between the prediction and the ground truth.

Experiments

Experimental Setups

To evaluate our system, we carried out experiments on the following two public datasets.

- **RACE**: English examinations for Chinese middle and high school students. The questions are generated by human experts, which has four candidates for each question. The test set consists of 4,934 instances, where RACE-M (middle school) has 1,436 instances, and RACE-H (high school) for 3,498 instances.
- **SemEval-2018 Task11**: The dataset provided by the SemEval-2018 Task11 organizer, which mainly focus on solving commonsense reading comprehension. Each question has two candidates to choose from.

The data are tokenized and lower-cased by using Natural Language Toolkit (NLTK) (Bird and Loper 2004), and all punctuations are removed. The main hyper-parameters of our model are listed in Table 2. Note that, except for the candidate numbers, all hyper-parameters are identical among two datasets. The word embeddings are initialized by the pre-trained GloVe word vectors (Common Crawl, 6B tokens, 100-dimension) (Pennington, Socher, and Manning 2014), and keep fixed during training. The words that do not appear in the pre-trained word vectors are set to the `unk` token and initialized accordingly. We use Adam (Kingma and Ba 2014) for weight optimizations with an initial learning rate of 0.001. In order to prevent overfitting, we apply dropout of 0.35 to all the representation layers. The models are built on Keras platform (Chollet and others 2015) with Tensorflow backend (Abadi et al. 2016).

Symbol	Descriptions	Size
$\|P\|$	Passage max length	300
$\|Q\|$	Question max length	20
$\|C\|$	Candidate max length	10
e	Word embedding	200
h	Bi-LSTM hidden size	250
h_{att}	Attention hidden size	80
es	ELMo embedding size	1024
p	POS-tag embedding	16

Table 2: Hyper-parameter settings.

Overall Results

RACE. The experimental results are shown in Table 1. As we can see that, our CSA model shows significant improvements over various state-of-the-art systems by a large margin. We also compared our model to the recent unpublished work DFN, while our model gives an absolute gain of 1.0% in the overall test set, and further gains can be obtained by incorporating ELMo, demonstrating the effectiveness of our proposed model. Also, when compared to the Hierarchical Attention Flow model (Zhu et al. 2018), our CSA model show substantial improvements, indicating that utilizing the attentions in each hierarchy and using convolutional neural network to extract the most representative features from the spatial attentions are useful in this task.

When it comes to the ensemble results, though we only use 7 models in ensemble with majority voting approach, the overall results show an absolute gain of 2% over the previous state-of-the-art result by DFN. Also we observed that

Model	Dev	Test
HMA (Chen et al. 2018)	**84.48**	80.94
TriAN (Wang 2018)	83.84	81.94
CSA Model (single model)	83.63	82.20
CSA Model + ELMo (single model)	83.84	**83.27**
TriAN (ensemble)	85.27	83.95
HMA (ensemble)	**86.46**	84.13
CSA Model (ensemble)	84.05	84.34
CSA Model + ELMo (ensemble)	85.05	**85.23**

Table 3: Experimental results on SemEval-2018 Task11. The two top-ranked systems are listed as baselines.

our model shows slightly worse result in RACE-M but significantly better in RACE-H regardless of single model or ensemble. These results indicate that our model is more capable of solving difficult question (high school). While on the contrary, it may hurt the performance on the relatively easier question (middle school). We will give a detailed analysis for illustrating this phenomenon in the next section.

SemEval-2018 Task11. The distributions of question type is quite different between the RACE and SemEval-2018 Task11 datasets. To test if our model could generalize to other reading comprehension dataset, we also carried out experiments on the very recent SemEval-2018 Task11 dataset. The results are shown in Table 3. As we can see that, our CSA model could give moderate improvements over the top-ranked SemEval systems in both single model and ensemble, and set up new state-of-the-art performance on this task, demonstrating the proposed model is powerful and showing its potential possibility to generalize to other NLP tasks. Note that, we directly train the model using the hyper-parameter settings of the RACE experiments without finding other hyper-parameter combinations, indicating further improvements may be obtained through fine-tuning.

Ablation Study

We also carried out model ablations to further demonstrate the effectiveness of the proposed approaches. The results are shown in Table 4.

Model	RACE
CSA Model	48.52
w/o attention weight	48.18
w/o enriched representation	47.52
w/o convolutional spatial attention	47.30
CSA Model + ELMo	50.89
w/o attention weight	49.49
w/o enriched representation	49.78
w/o convolutional spatial attention	48.47

Table 4: Ablations on several model components.

As shown in the model description, we adopt the attention mechanism used in FusionNet (Huang et al. 2017), and applied an element-wise weight to the attention matrix. By applying additional trainable weights to the attention matrix could give moderate improvements suggesting that these weights are useful in dynamically adjusting the attention values. When we remove all the enriched candidate representations, there is an absolute drop of 1.0%. This suggests that it is necessary to incorporate various information (such as the passage and question) into the candidate representation. We also removed our convolutional spatial attention mechanism in the model. The results show a significant drop in performance by 1.22%, indicating that the proposed convolutional spatial attention is effective in extracting most representative values among the various attention matrices. The same results drop is shown by incorporating ELMo. Further more, whether we use ELMo or not, our convolutional spatial attention has improved significantly. That also proved that our model can get father improvement after some big data trained toolkit like ELMo or some other transformer.

The detailed settings of the ablation experiments are shown below.

- **w/o attention weight**: We change $M = M' \odot W($ in Alg 1 line 2) to $M = M'$, where W is the attention weight.
- **w/o enriched representation**: We remove interaction of candidate-to-question (R_{CQ} in Figure 2) and candidate-to-passage (R_{CP} in Figure 2) and only use candidate's LSTM output R_C.
- **w/o convolutional spatial attention**: We use two fully-connected layer to transform the matching matrix to a score. The first fully-connected layer squeeze the matching matrix to a vector along the question length dimension. The second fully-connected layer squeeze the vector to a scalar score. As every candidate has six matching matrices, so we can get six scores. Finally, we add the six scores as the final prediction.

Analysis and Discussion

Quantitative Analysis

In our CSA model, we adopt various enriched representations for better characterizing candidate information in different semantic aspects. To better understand our model, we compare the performance between the original CSA model and without enriched candidate representations, as shown in Figure 3.

As we can see that our model yields significant improvements on the question type 'how' and 'why', which are relatively difficult than the other type of questions requiring high-level reasoning within the context, demonstrating that our CSA model performs better on the relatively sophisticated questions. However, on the contrary, we find that our CSA model shows relatively inferior performance on the question type 'who', 'where', and 'when', which are often answered by a single word or entity name. This phenomenon suggests that further efforts should be made on balancing the word-level attention and highly abstracted attention to better solving both easy and hard questions in reading comprehension.

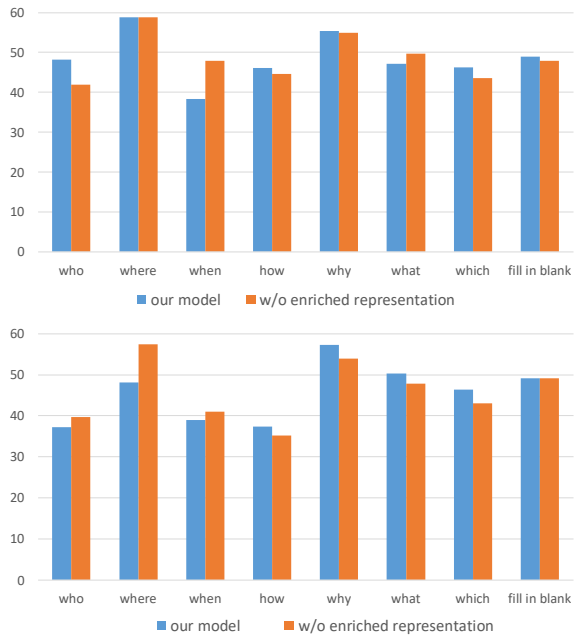

Figure 3: Quantitative analysis on different type of questions in RACE development and test set.

State	Total #	Right #	Accuracy
All	100	56	56.0%
First level	30	18	60%
Second level	30	20	66.7%
Third level	40	18	45.0%

Table 5: Case analysis on RACE dataset.

RACE Dataset	
Passage As is known to all, in daily ***conversation*** people often use simple words and simple sentences, especially ***elliptical*** sentences. Here is an interesting conversation between Mr Green and his good friend Mr Smith, a fisherman. Do you know what they are talking about?	
Question The ***text*** is mainly about _ .	
Candidates	
A how to catch fish.	**B** how to spend a Sunday
C ***ellipsis in conversations***	**D** joy in fishing

Figure 4: An example of RACE dataset. The correct answer is depicted in bold face.

These results also explain why our CSA model shows significant improvements on RACE-H (high school) subset while giving a relatively inferior performance on RACE-M (middle school) subset. As the sample number of RACE-H subset is three times the size of RACE-M, the overall performance of our model still shows significant improvements over various state-of-the-art systems.

Case Study

We also randomly sampled one hundred question from RACE test set, and classified them into three categories according to its difficulty. The results are shown in Table 5.

- The first level is the question that can be answered by matching a few words in the passage. If the model finds the right place in the text, the answer is easy to find. With the attention mechanism, the neural model is especially good at solving this kind of question.
- The second level is the question that can be answered by using a few sentences without reasoning, which is relatively difficult than the first level.
- The third level is the question that needed to comprehensive reasoning via multiple clues in the passage. This kind of question is more complicated. To answer this question, the intention of the questioner must be captured.

As we can see that the current model shows relatively good performance on the first and second level questions than the third one, indicating that more investigation should be made on solving those questions that need reasoning.

An interesting example is shown in Figure 4, where our model chooses the right answer "C". The reason why our model could pick out the right answer is that the candidate "C" has the biggest semantic overlap with passage. If we change the candidate "C" to *"ellipsis in talking"* or change the question to *"The conversation is mainly about _ ."*, the model could still pick out the candidate "C" with a high probability as the correct answer. Through this experiment, we can see that our model could handle part of the high level semantic matching which suggest that the proposed CSA model is effective and robust on modeling text with similar meanings.

Conclusion

We propose the Convolutional Spatial Attention (CSA) model to tackle the machine reading comprehension with multiple-choice questions. The proposed model could fully extract the mutual information among the passage, question, and candidates, to form the enriched representations using the modified attention mechanism that has trainable weights. To summarize attention matrices from various sources, we propose to use convolutional operation to dynamically summarize the attention values within the different size of regions, which is beneficial to capture diverse features. Experimental results show that the proposed CSA model could give substantial improvements over various state-of-the-art systems on both RACE and SemEval 2018 Task11 datasets. Also, the ablation studies verify the effectiveness of several proposed components in our model, and case analysis also shows that the CSA model is superior in solving relatively sophisticated questions (such as 'why' or 'how' questions).

In the future, we would like to investigate how to make a good balance between the word-level attention and highly abstracted attention information to better solving both easy and hard questions.

Acknowledgments

We would like to thank all three anonymous reviewers for their constructive comments to improve our paper. This work was supported by the National Key R&D Program of China via grant No. 2016YFC0800806.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 265–283.

Bird, S., and Loper, E. 2004. Nltk: the natural language toolkit. In *ACL 2004 on Interactive Poster and Demonstration Sessions*, 31.

Chen, Z.; Cui, Y.; Ma, W.; Wang, S.; Liu, T.; and Hu, G. 2018. Hfl-rc system at semeval-2018 task 11: Hybrid multi-aspects model for commonsense reading comprehension. *arXiv preprint arXiv:1803.05655*.

Chen, D.; Bolton, J.; and Manning, C. D. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of ACL 2016*, 2358–2367. Association for Computational Linguistics.

Chollet, F., et al. 2015. Keras. https://github.com/fchollet/keras.

Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; and Hu, G. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of ACL 2017*, 593–602. Association for Computational Linguistics.

Dhingra, B.; Liu, H.; Yang, Z.; Cohen, W.; and Salakhutdinov, R. 2017. Gated-attention readers for text comprehension. In *Proceedings of ACL 2017*, 1832–1846. Association for Computational Linguistics.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5-6):602–610.

Hermann, K. M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *International Conference on Neural Information Processing Systems*, 1693–1701.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Hu, M.; Peng, Y.; and Qiu, X. 2017. Reinforced mnemonic reader for machine comprehension. *CoRR, abs/1705.02798*.

Huang, H.-Y.; Zhu, C.; Shen, Y.; and Chen, W. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341*.

Kadlec, R.; Schmid, M.; Bajgar, O.; and Kleindienst, J. 2016. Text understanding with the attention sum reader network. In *Proceedings of ACL 2016*, 908–918. Association for Computational Linguistics.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of EMNLP 2017*, 785–794. Association for Computational Linguistics.

Ostermann, S.; Roth, M.; Modi, A.; Thater, S.; and Pinkal, M. 2018. Semeval-2018 task 11: Machine comprehension using commonsense knowledge.

Parikh, S.; Sai, A.; Nema, P.; and Khapra, M. M. 2018. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. https://openreview.net/forum?id=B1bgpzZAZ.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, 1532–1543. Association for Computational Linguistics.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*, 2227–2237. Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, 2383–2392. Association for Computational Linguistics.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of EMNLP 2013*, 193–203.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of ACL 2017*, 189–198. Association for Computational Linguistics.

Wang, L. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. *CoRR* abs/1803.00191.

Xiong, C.; Zhong, V.; and Socher, R. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Xu, Y.; Liu, J.; Gao, J.; Shen, Y.; and Liu, X. 2017. Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies. *arXiv preprint arXiv:1711.04964*.

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Zhu, H.; Wei, F.; Qin, B.; and Liu, T. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of AAAI-18*.