Rception: Wide and Deep Interaction Networks for Machine Reading Comprehension (Student Abstract)

Xuanyu Zhang, Zhichun Wang
School of Artificial Intelligence
Beijing Normal University, Beijing, China
xyz@mail.bnu.edu.cn, zcwang@bnu.edu.cn

Abstract

Most of models for machine reading comprehension (MRC) usually focus on recurrent neural networks (RNNs) and attention mechanism, though convolutional neural networks (CNNs) are also involved for time efficiency. However, little attention has been paid to leverage CNNs and RNNs in MRC. For a deeper understanding, humans sometimes need local information for short phrases, sometimes need global context for long passages. In this paper, we propose a novel architecture, i.e., Rception, to capture and leverage both local deep information and global wide context. It fuses different kinds of networks and hyper-parameters horizontally rather than simply stacking them layer by layer vertically. Experiments on the Stanford Question Answering Dataset (SQuAD) show that our proposed architecture achieves good performance.

Introduction

Machine reading comprehension (MRC) aims to teach machines to comprehend a passage and answer corresponding questions about the passage. Most of the previous models use RNN as the main skeleton to synthesizes global information. While CNN is rarely explored in MRC after QANet (Yu et al. 2018). Although the larger receptive field of CNNs can be obtained by deeper networks or more sophisticated techniques, CNN is still not ideal enough to synthesize global context compared to RNN. Besides, lots of classical architectures are proposed for CNNs in computer vision, such as Inception (Szegedy, Liu, and Jia 2015). And for RNNs, there are few special designs. RNNs are usually used layer by layer in depth. Moreover, little attention has been paid to the trade-off on local and global interactive information in MRC. Intuitively, local information grasped by CNNs is useful for short answers and phrases, and global understanding grasped by RNNs is important for long passages. In this paper, we propose a novel architecture, i.e., Rception, to leverage both deep local and wide global information and apply it to the interaction of MRC. It is not a simple combination of CNNs and RNNs in depth. We arrange CNNs and RNNs with different parameters into different branches and aggregate them in the framework like Inception, creatively. For wide part, RNNs are aggregated horizontally rather than stacked layer by layer. For deep part, we use depthwise separable convolution with dilation mechanism to fuse original information and increase the receptive field, capturing features of different dimensions over phrases rather than words.

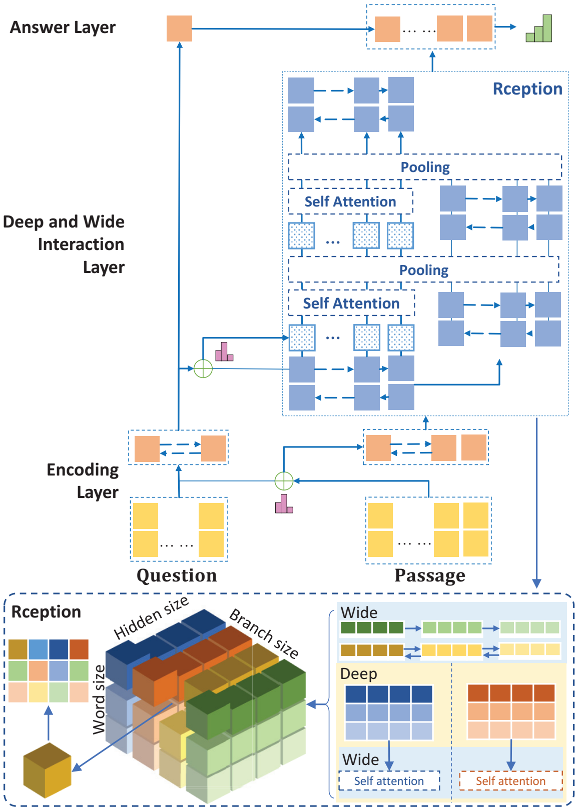

Figure 1: The architecture of our model and Rception.

Methodology

Suppose we are given a question with m tokens $Q = \{w_t^Q\}_{t=1}^m$ and a passage with n tokens $P = \{w_t^P\}_{t=1}^n$. Our goal is to predict the answer A in P. Generally, the answer is a continuous span of the text from the passage. As shown in Figure 1 (top), there are three layers in our model: encoding layer, deep and wide interaction layer and answer layer. Our proposed Rception is used in the second layer.

Encoding Layer and Answer Layer For each word in the question or passage, we obtain the contextualized embedding by collapsing all hidden layers of the weight-fixed BERT (Devlin et al. 2019) into a vector following SDNet and MC^2 (Zhang 2019). The representation of the passage is $r_t^P = [e_t^P; r_t^{POS}; r_t^{NER}; r_t^{attn}]$, where $\{e_t^P\}_{t=1}^n$ is the embedding of the passage, r_t^{POS}, r_t^{NER} and r_t^{attn} are embeddings of the part-of-speech, named entity recognition tags and attention scores with questions $\{e_t^Q\}_{t=1}^m$. And for the question, we use a RNN to generate new representation r_j^Q according to e_j^Q. And answer layer is the top one of our model. We use bilinear functions to obtain the start and end probability of each token in the passage like other models.

Deep and Wide Interaction Layer This layer plays a significant role in our model. As shown in Figure 1 (bottom), our proposed Rception is composed of deep parts and wide parts in the interaction layer. The yellow area represents the deep part, and the blue area is the wide one, including RNN and self-attention. They were interlaced with each other to capture the deep and wide interactive information. Different from traditional models, we utilize multiple branches information like Inception (Szegedy, Liu, and Jia 2015). Different branches have different network structures and hyper parameters. Self-attention and RNN are also used in branches to obtain more comprehensive information besides CNN, so we call this structure *Rception*. Specifically, we use three CNNs with different kernel size as our CNN branches. Then they are fed to self-attention layer, separately. And for RNN branches, we use a unidirectional RNN and a bidirectional RNN as our RNN branches. Then we integrate 5 different branches by the max pooling operation to choose the most important branch for each hidden state of the word. It is like humans tend to choose the most reasonable understanding after thinking from different perspectives. And this motivation can also be applied to every CNN branch. We use 1D dilated depth-wise separable convolution. It first performs a depth-wise spatial convolution, which acts on each input channel, i.e., each hidden state of the word separately. Then a point-wise convolution is used to mix together the output channels above for each word.

As shown in Figure 1 (top), after the passage u_t^P is obtained by RNN, we can obtain question-aware passage representation by attention mechanism $s_j^t = {r_j^Q}^{\mathrm{T}} u_t^P$, $a_i^t = exp(s_i^t)/\sum_{j=1}^m exp(s_j^t)$, $r_t^{att} = \sum_{i=1}^m a_i^t e_i^Q$. We concatenate it with the passage representation $h_t^P = [u_t^P; r_t^{att}]$. Then it is refined to $\{\bar{h}_t^P\}_{t=1}^n$ by Rception. The final states for answer layer is $z_t^P = \mathrm{BiRNN}(z_{t-1}^P, [h_t^P; \bar{h}_t^P])$.

Experiments

We evaluate our model on the Stanford Question Answering Dataset (SQuAD 1.1) (Rajpurkar and Liang 2016). As shown in Table 1, Rception achieves 84.4% in EM and 90.9% in F1 score, which is comparable with BERT$_{\mathrm{LARGE}}$ fine-tuning models. And even if we replace weight-fixed BERT with ELMo, it still outperforms other models using ELMo, like BiDAF+Self-Attention+ELMo (Peters et al. 2018). The results indicate that the good performance of the model is not just because of BERT.

Model	**Dev Set**		**Test Set**	
	EM	F1	EM	F1
BIDAF	67.7	77.3	68.0	77.3
R-NET	72.3	80.6	72.3	80.7
QANet (+data×3)	75.1	83.8	76.2	84.6
BiDAF+Self-Att+ELMo	77.9	85.6	78.6	85.8
R.M.Reader	78.9	86.3	79.5	86.6
Ours (-BERT,+ELMo)	**80.5**	**87.6**	-	-
BERT$_{\mathrm{BASE}}$	80.8	88.5	-	-
BERT$_{\mathrm{LARGE}}$ (-TriviaQA)	84.1	90.9	-	-
Ours	**84.6**	**90.9**	**84.4**	**90.6**
Ours (w/o deep part)	83.5	90.4	-	-
Ours (w/o wide part)	83.5	90.0	-	-
Ours (w/o Rception)	82.4	89.0	-	-

Table 1: The performance on the SQuAD dataset.

We also explore the effect of each component of our model by ablation study in Table 1. We can observe that both the deep part and the wide part of Rception make important contributions to the performance of our model. Furthermore, we evaluate our model on two adversarial datasets (Jia and Liang 2017). Our model achieves EM 40.0%, F1 45.6% in AddSent and EM 52.0%, F1 58.2% in AddOneSent, which also outperforms most of models and shows our model is robust and has good generalization ability.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No.2017YFC0804004).

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *ACL*, 2021–2031.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *ACL*, 2227–2237.

Rajpurkar, P., and Liang, P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *ACL*.

Szegedy, C.; Liu, W.; and Jia, Y. 2015. Going deeper with convolutions. In *CVPR*.

Yu, A. W.; Dohan, D.; Luong, M.; Zhao, R.; and Chen, K. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

Zhang, X. 2019. MCˆ2: Multi-perspective convolutional cube for conversational machine reading comprehension. In *ACL*, 6185–6190.