HIT-SCIR

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 RESEARCH (HFL), BEIJING, CHINA

JAN 30, 2019

AAAI 19, HAWAII, USA

OUTLINE

- Introductions & Preliminaries
- Convolutional Spatial Attention Model (CSA)
- Experimental Results
- Quantitative Analysis
- Conclusions & Future Works

INTRODUCTION

- *Machine Reading Comprehension (MRC)* is to read and comprehend a given article and answer the questions based on it, which has become enormously popular in recent few years.
- Type of MRC
  - Cloze-style: CNN / Daily Mail [Hermann et al., 2015], CBT [Hill et al., 2015]
  - Span-extraction: SQuAD [Rajpurkar et al., 2016]
  - Choice selection: MCTest [Richardson et al., 2013], RACE [Lai et al., 2017]
  - Conversational MRC: CoQA [Reddy et al., 2018], QuAC [Choi et al., 2018]
- In this paper, we focus on solving the RC problem with multiple-choice questions

Introduction

- *RC with multiple-choice question*
  - *Document*
    - Pre-requisites for answering the questions
  - *Question*
  - *Candidates*
  - *Answer*

Document:
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got in to lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Question: What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

Answer: C) James

INTRODUCTION

- *RC with multiple-choice question*
  - *Document*
  - *Question*
    - A natural question based on the documents
  - *Candidates*
  - *Answer*

Document:
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got in to lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Question: What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

Answer: C) James

Introduction

- *RC with multiple-choice question*
  - *Document*
  - *Question*
  - *Candidates*
    - Candidate answers for the question
  - *Answer*

Document:
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got in to lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Question: What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

Answer: C) James

INTRODUCTION

- *RC with multiple-choice question*
  - *Document*
  - *Question*
  - *Candidates*
  - *Answer*
    - Choose the correct one as the answer

Document:
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got in to lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Question: What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

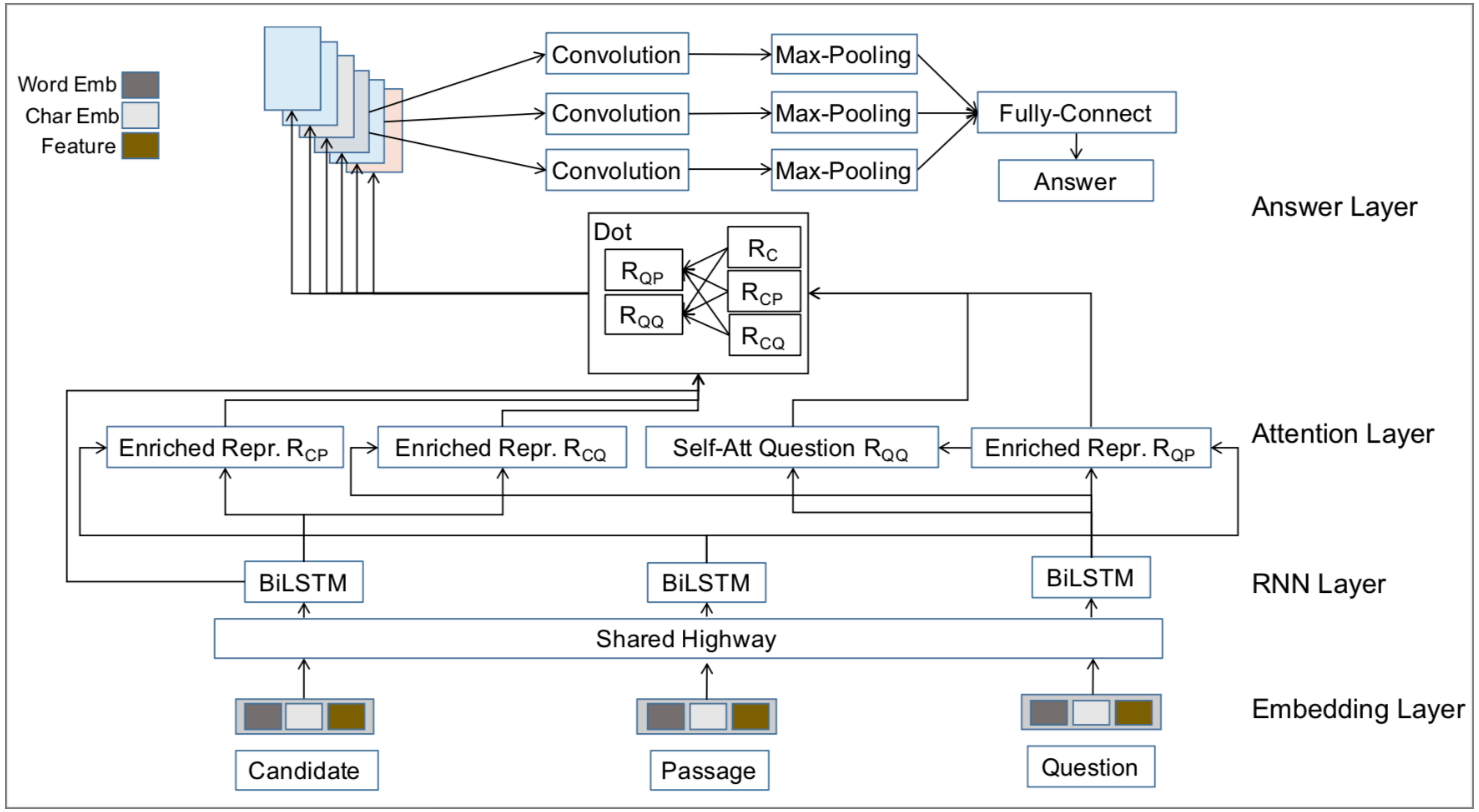
Answer: C) James

CSA MODEL

- **Contributions**
  - Focus on modeling different semantic aspects of **candidate answers**
  - Propose **C**onvolutional **S**patial **A**ttention (CSA) to simultaneously extract the attentions between various representations
  - Experimental results on RACE and SemEval 2018 Task 11 show that the proposed model achieves state-of-the-art performance.

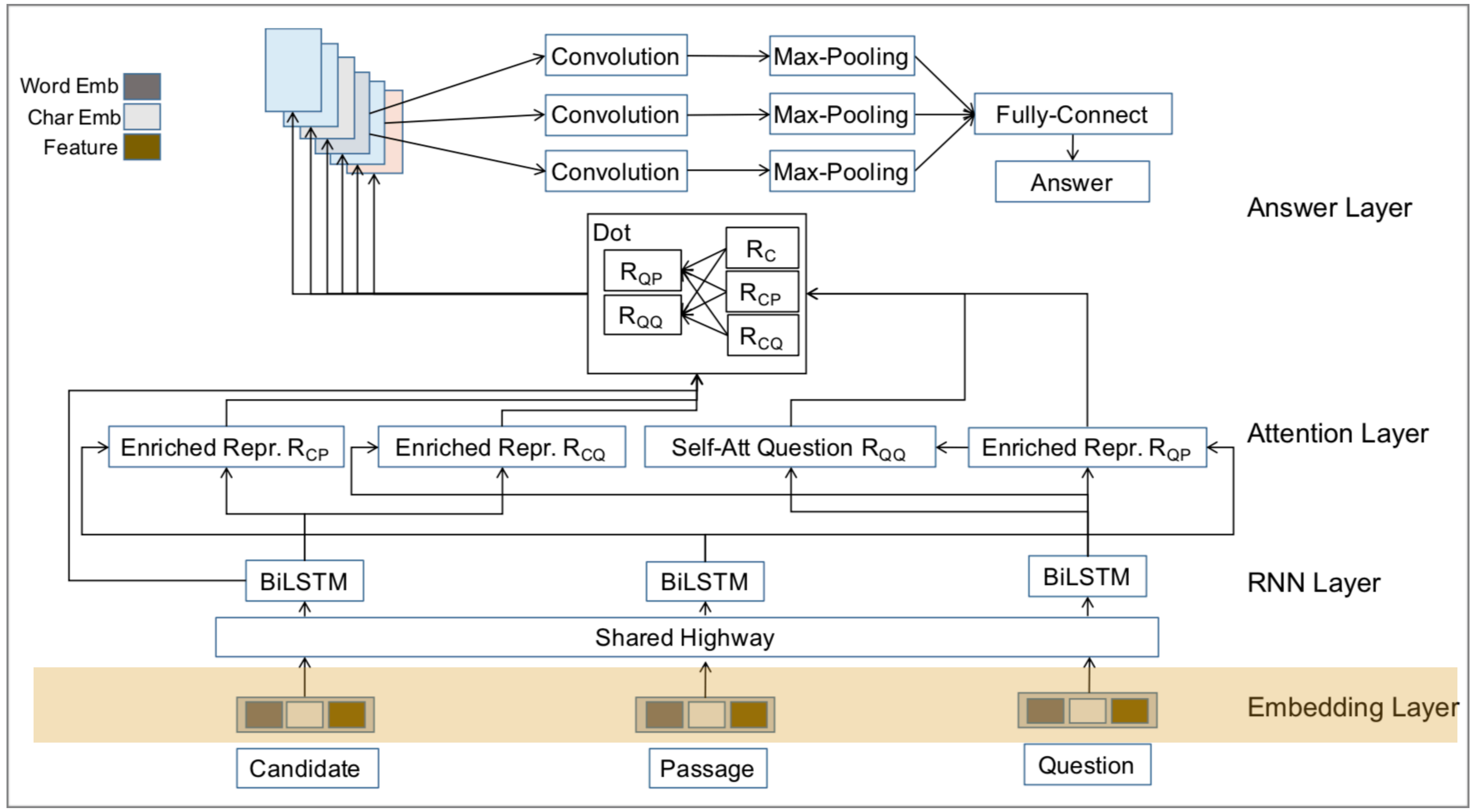
CSA Model

- **Formal Definition of the Task**
    - Inputs: Document, Question, Candidate
    - Output: Candidate score of being the answer
- **Basic Components**
    - Embedding Layer
    - LSTM Layer
    - Enriched Representation Layer
    - Convolutional Spatial Attention Layer
    - Answer Layer



CSA Model

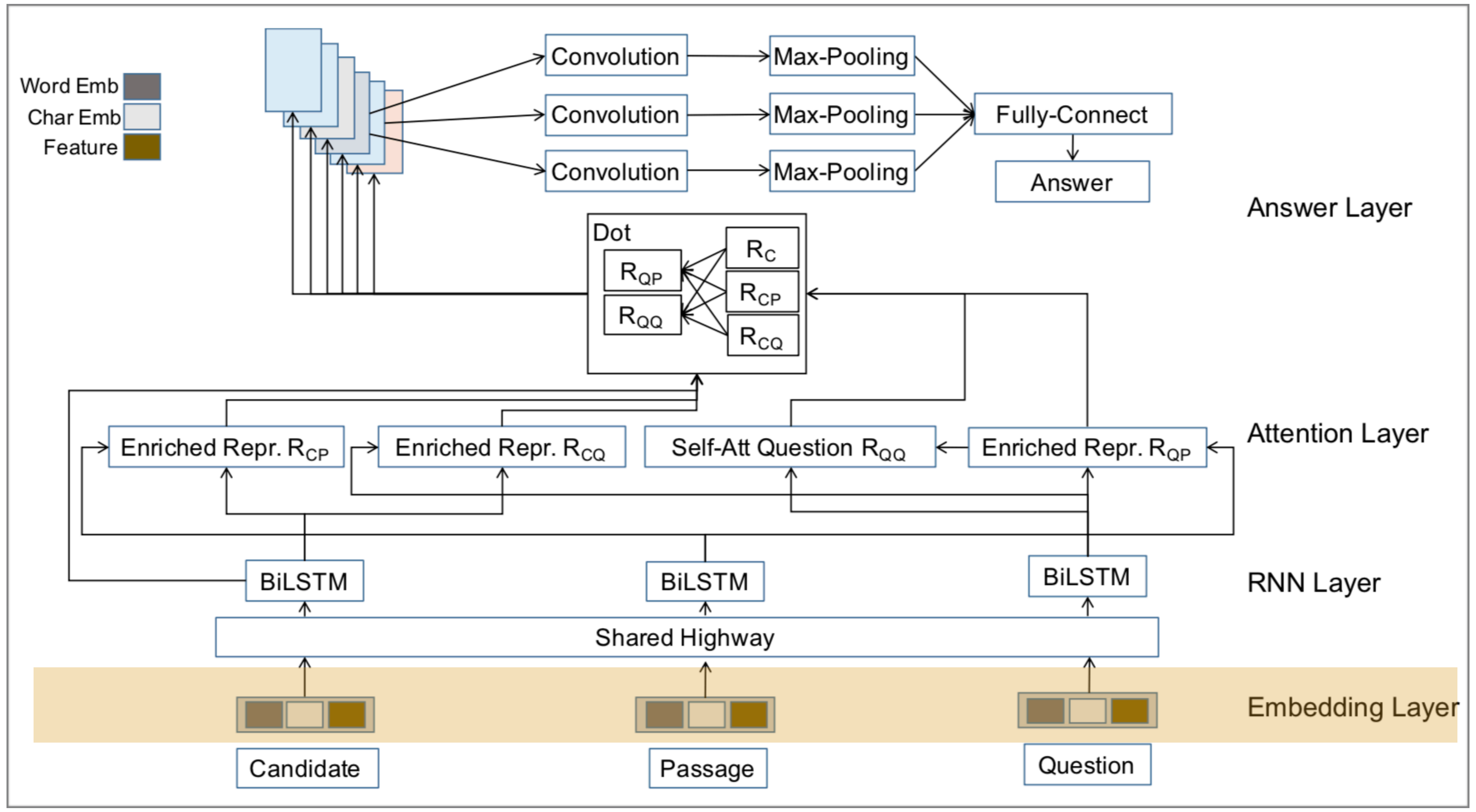
- **Embedding Layer**
  - GloVe Word Embedding [Pennington et al., 2013]
  - ELMo [Peters et al., 2018]
  - POS-tag Embedding
  - Exact Word Matching
  - Fuzzy Word Matching
- Concatenate all the features above



CSA Model

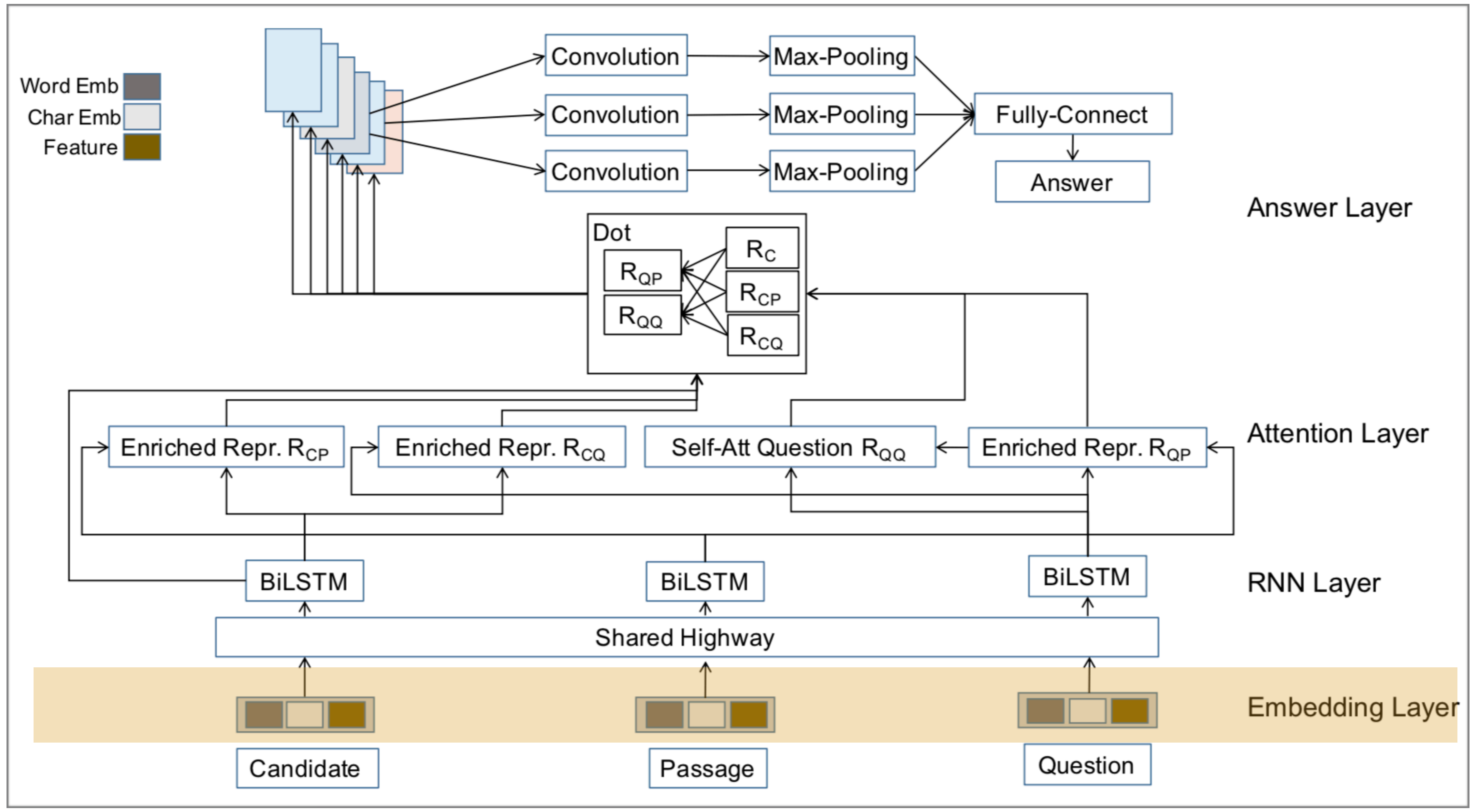
- **LSTM Layer**
  - Apply highway layer to better mix various types of embeddings
  - Place an ordinary Bi-LSTM layer after embedding to obtain contextual representation

$$\widetilde{H} = \sigma(\text{2-Highway}(E))$$
$$H = \text{Bi-LSTM}(\widetilde{H})$$



CSA Model

- **Enriched Representation Layer**
  - Using 'enriched representation algorithm' to get various attention-guided representations.
  - **R_{CQ}**: question-aware candidate representation
  - **R_{CP}**: passage-aware candidate representation
  - **R_{QP}**: passage-aware question representation
  - **R_{QQ}**: self-attended question representation



CSA MODEL

- **Algorithm for Enriched Representation**

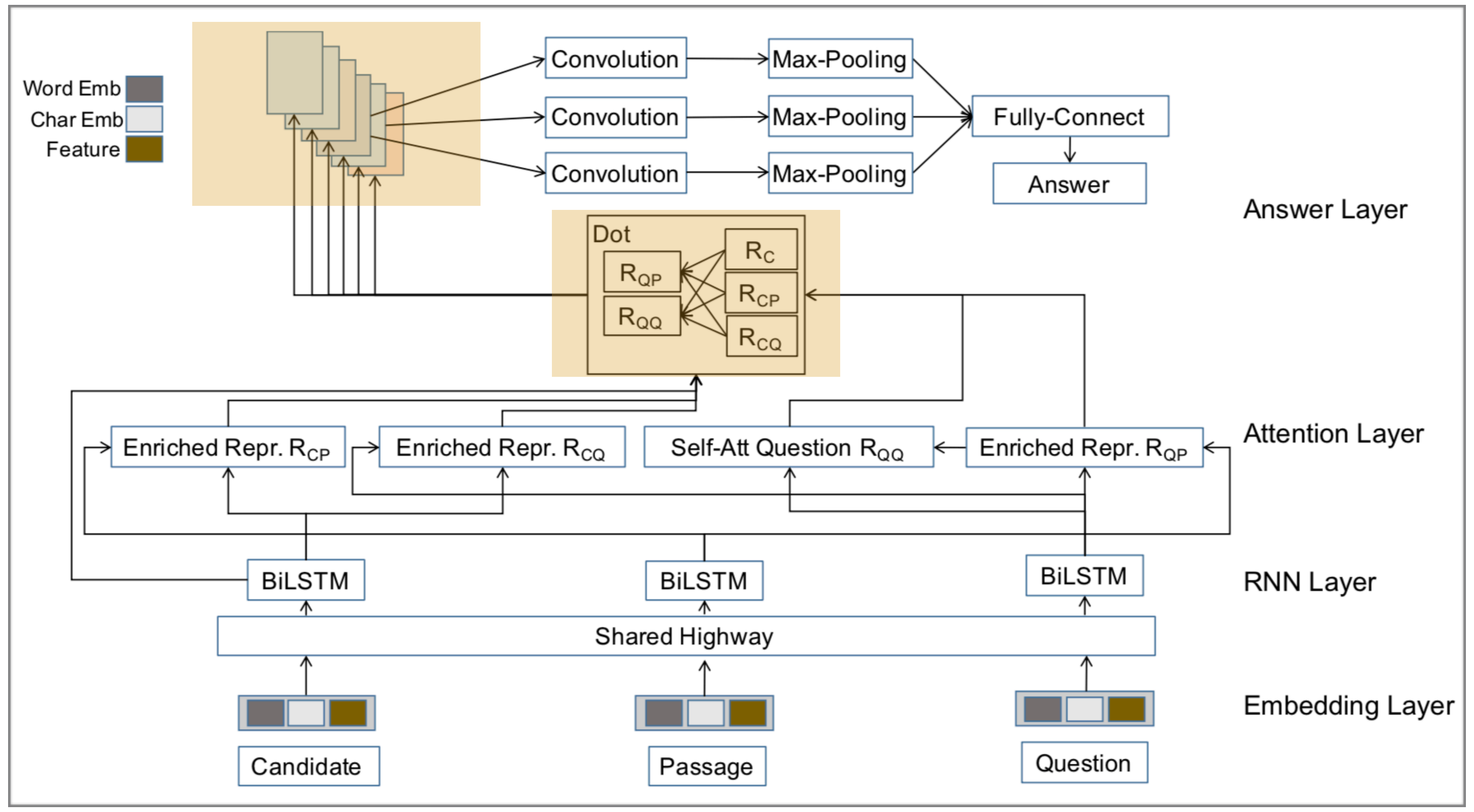
- Two Key Points
  - Adopt symmetric attention mechanism [Huang et al., 2017]
  - Apply element-wise weight to the attention matrix

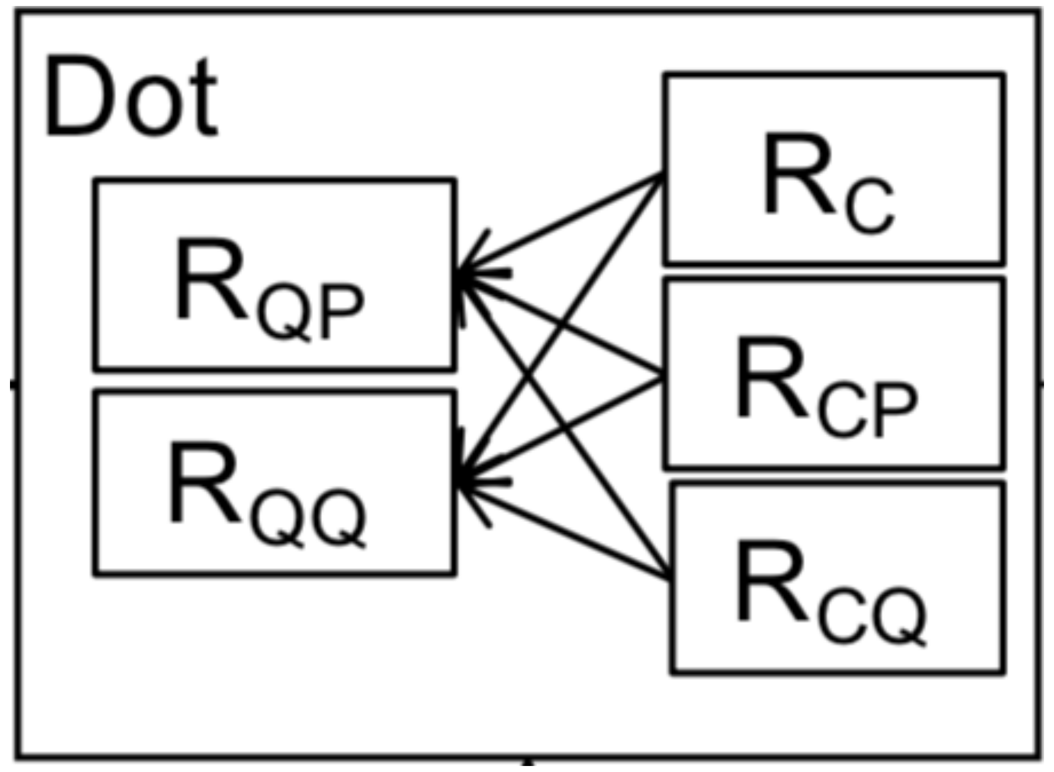
Algorithm 1 Enriched Representation.

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:
   $Y = \text{Bi-LSTM}([X_1; Y'])$
6: **return** Y

CSA MODEL

- **Convolutional Spatial Attention Layer**
  - Candidate information is important
  - We calculate dot attentions between three candidate representations and two question representations
  - Concatenate 2*3=6 attention matrices, forming an attention cuboid **M** with shape [6, candidate_len, question_len]

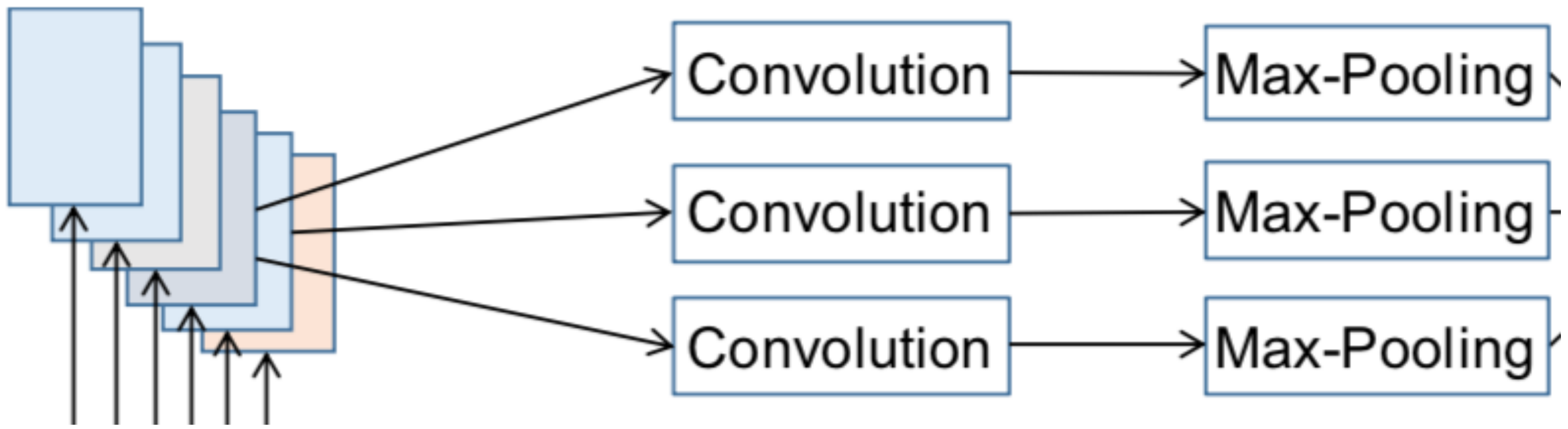




CSA Model

- **Convolutional Spatial Attention Layer**
  - The resulting matching cuboid **M** can be seen as a 2D-image with 6-channels
  - We use Convolution-MaxPooling operation to dynamically extract high-level features with kernel size 5, 10, 15

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

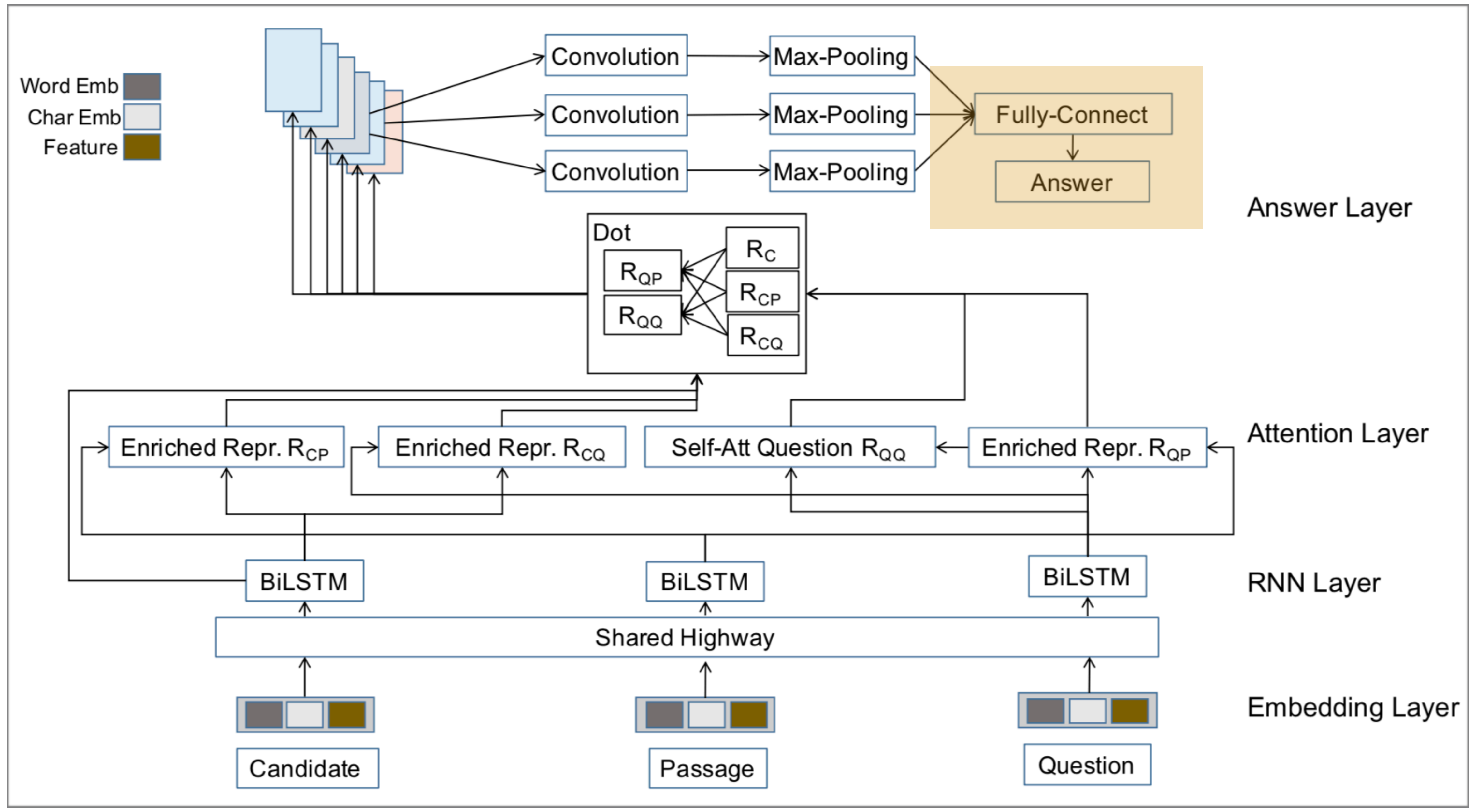


CSA Model

- **Answer Layer**
  - Concatenate all three feature vectors
  - Pass through a fully-connected layer to get a scalar score

$$s_i = \mathbf{w}^\mathbf{T} \cdot [O_1; O_2; O_3]$$
$$Pr(A|P,Q,C) = softmax([s_1; ...; s_N])$$

- Prediction: choose the candidate that has the largest score as the answer



EXPERIMENTS

- **Dataset**
  - RACE: English examinations of Chinese middle and high school students. (4 candidate selections)
  - SemEval 2018 Task 11: Machine Comprehension using Commonsense Knowledge (2 candidate selections)
- **Hyper-parameters**
  - Passage/Question/Candidate max length: 300 / 20 / 10
  - Word Embedding: 200-dim
  - Bi-LSTM hidden size: 250-dim
  - ELMo: 1024-dim
- **Implementation**: Keras + TensorFlow

EXPERIMENTS

- **Results on RACE**
  - Shows state-of-the-art performance, especially on RACE-H (high school)
  - Incorporating ELMo yields another significant improvements

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**

EXPERIMENTS

- **Results on SemEval 2018**
  - Baselines are the top two teams in SemEval 2018 Task 11.
  - CSA model shows marginal but consistent improvements on single/ ensemble settings.
  - With the help of ELMo, there is another boost in performance.

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**

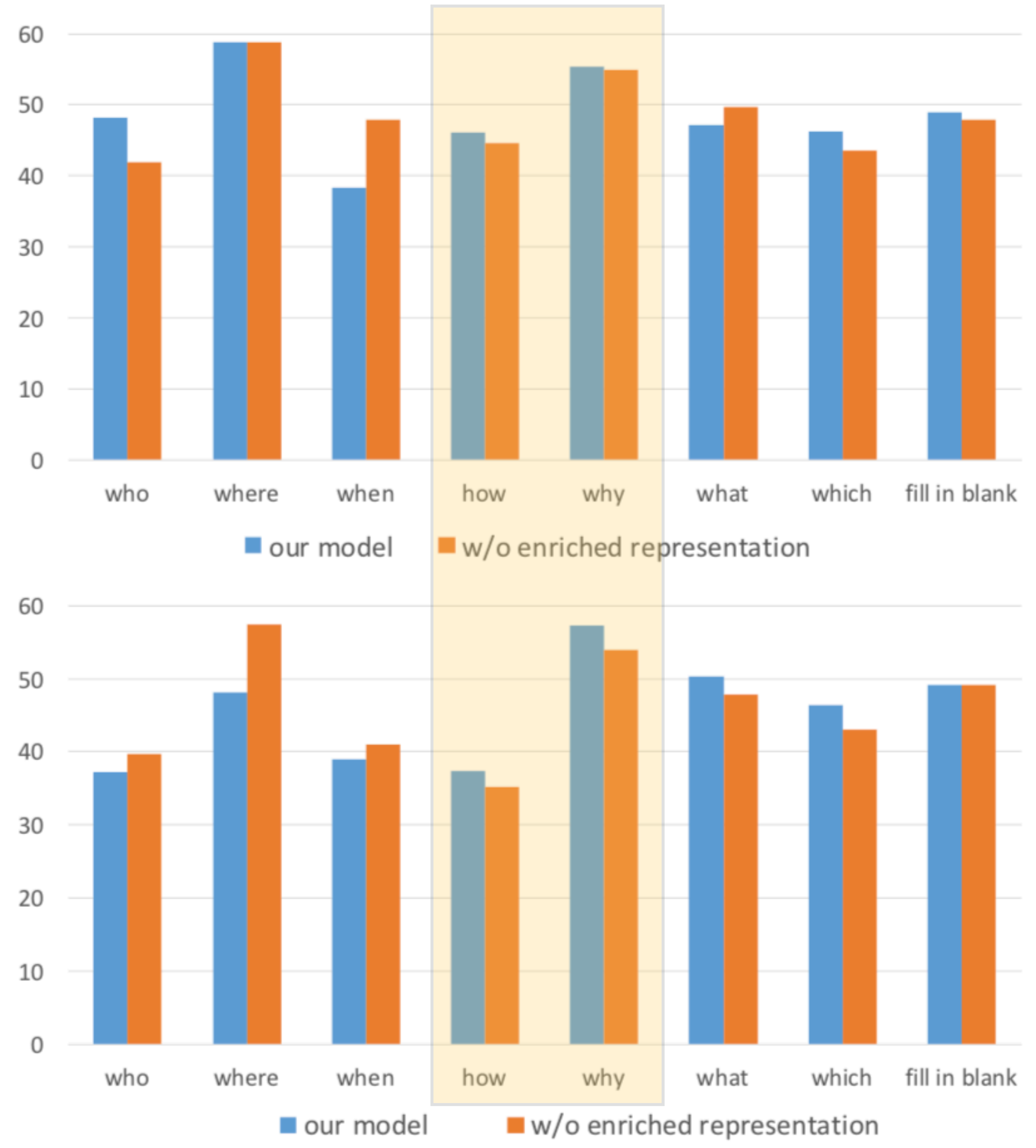
Ablation Study

- **Ablation Results on RACE**
  - w/o attention weight: do not apply element-wise weight on attention
  - w/o enriched repr: only use LSTM outputs
  - w/o CSA: using two fully connected layer to achieve dimensionality reduction of the 3D-attention
- Importance: CSA > enriched repr > att weight

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

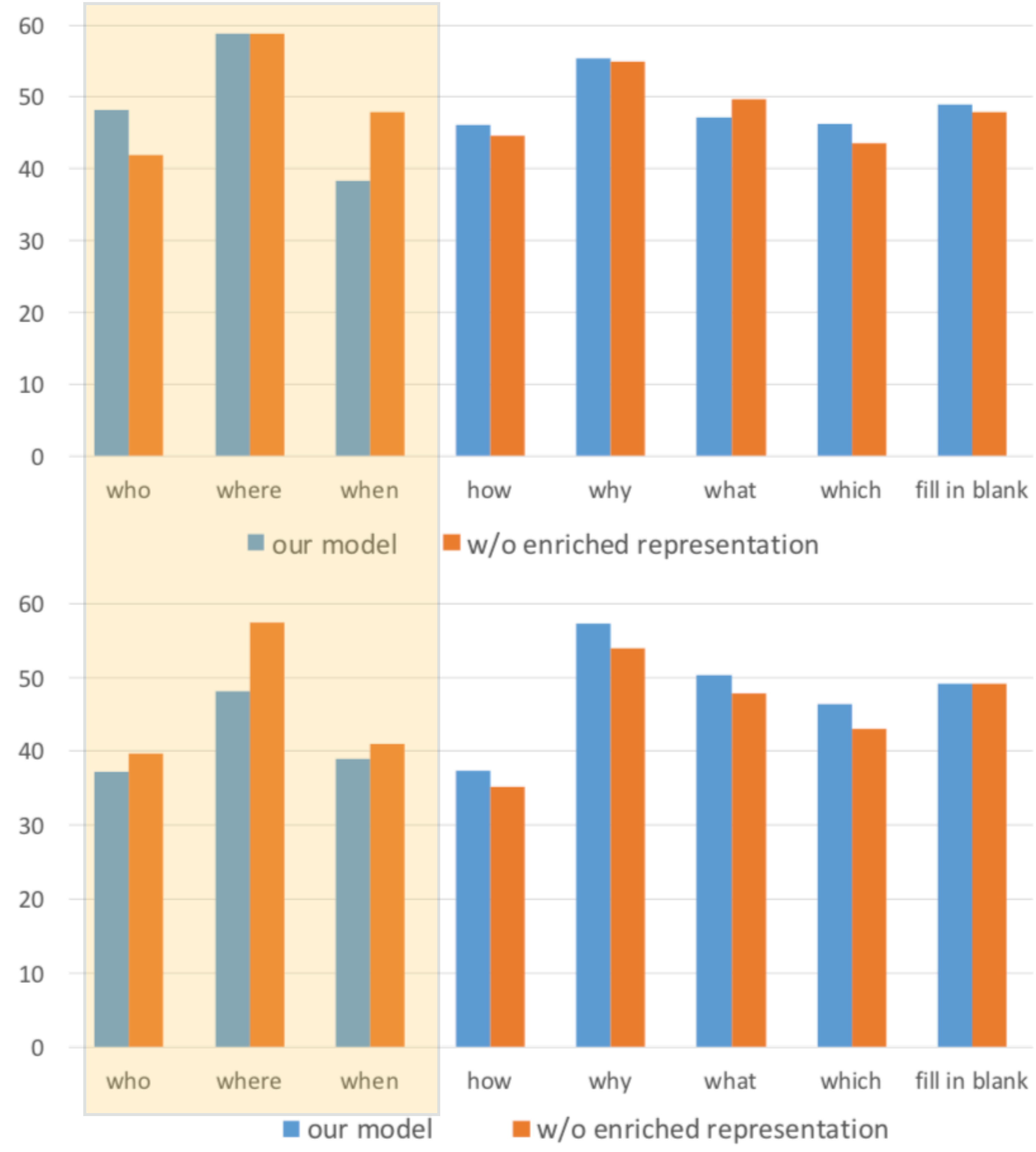
ANALYSIS

- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  - [+] CSA model is good at handling 'how' and 'why' questions, which needs comprehensive reasoning on the document
  - [-] On the contrary, CSA model shows inferior performance on 'who', 'when', 'where' questions
- Further efforts should be made on balancing the word-level attention and highly abstracted attention.



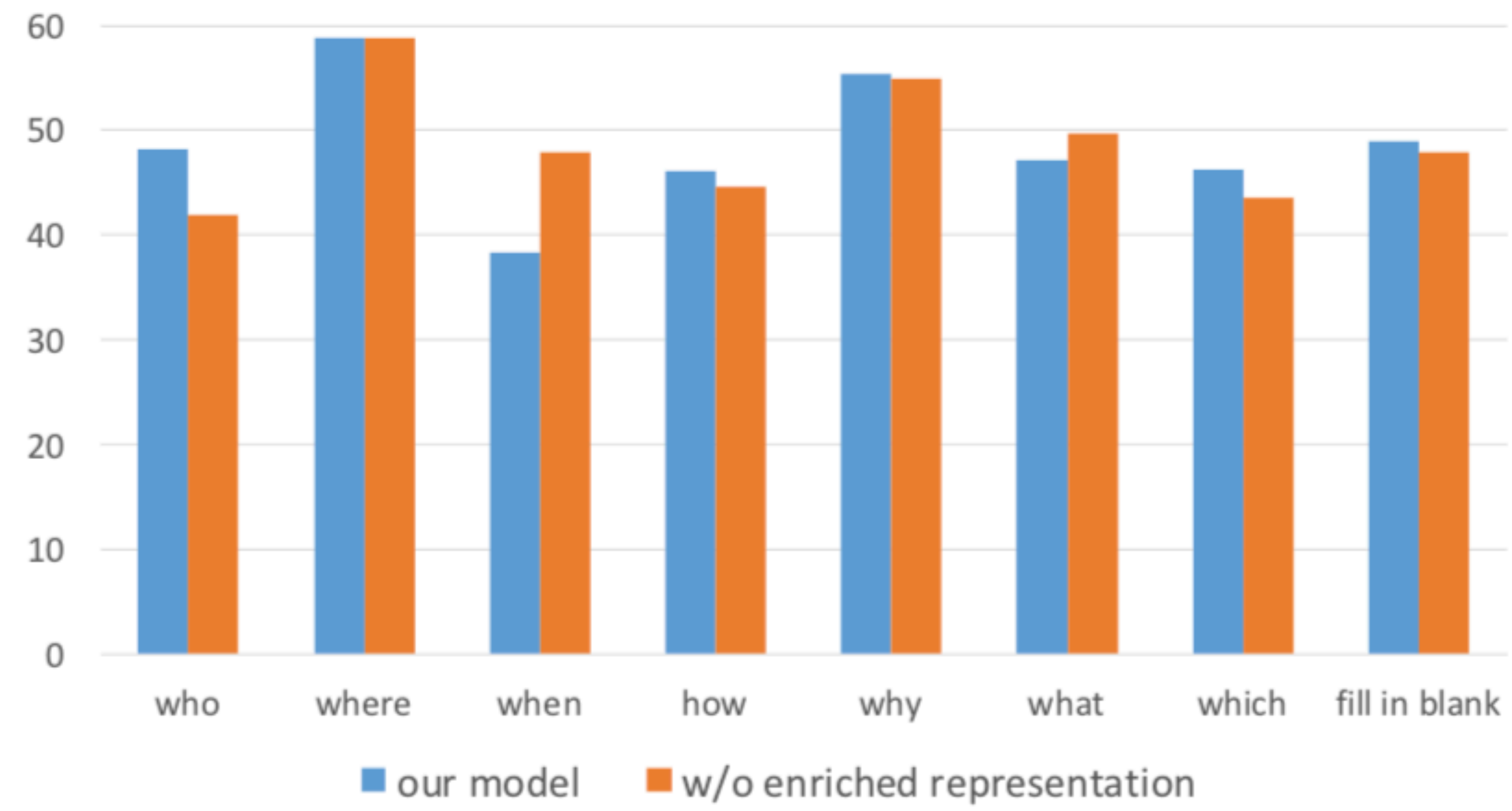
ANALYSIS

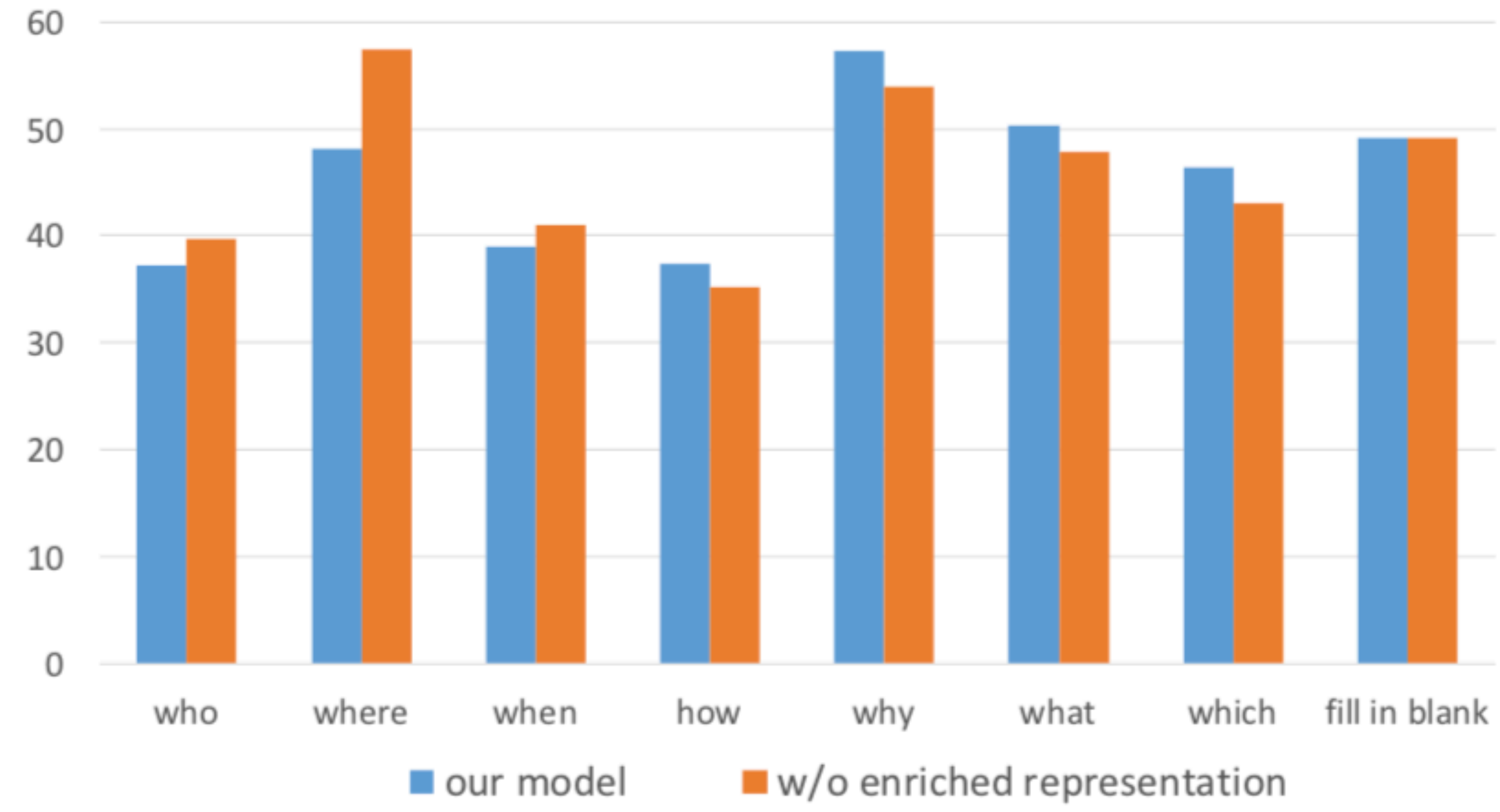
- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  - [+] CSA model is good at handling 'how' and 'why' questions, which needs comprehensive reasoning on the document
  - [-] On the contrary, CSA model shows inferior performance on 'who', 'where', 'when' questions
- Further efforts should be made on balancing the word-level attention and highly abstracted attention.



ANALYSIS

- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  - [+] CSA model is good at handling 'how' and 'why' questions, which needs comprehensive reasoning on the document
  - [-] On the contrary, CSA model shows inferior performance on 'who', 'where', 'when' questions
- Further efforts should be made on balancing the word-level attention and highly abstracted attention.





Conclusions & Future Work

- **Conclusion**
  - Propose Convolutional Spatial Attention model for RC with multiple-choice questions
  - The proposed model done well on hard problems types, such as 'how' and 'why'
  - Experimental results show significant improvements on RACE and SemEval 2018 datasets
- **Future Work**
  - Integrate CSA model into BERT
  - Further exploiting the relations between the document, question, and candidates

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.
- 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 the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 593–602.
- 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.
- 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.; Kocˇisky´, 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.

REFERENCES

- 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.
- 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.
- 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.
- 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.

REFERENCES

- 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.
- 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 ACL 2017, 189– 198.
- 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. https://www.aaai.org/ocs/index.php/AAAI/AAAI18 /paper/view/16331.

THANK YOU !

ENJOY YOUR TIME IN HAWAII !

CONTACT: ymcui [at] iFLYTEK [dot] com

download slides