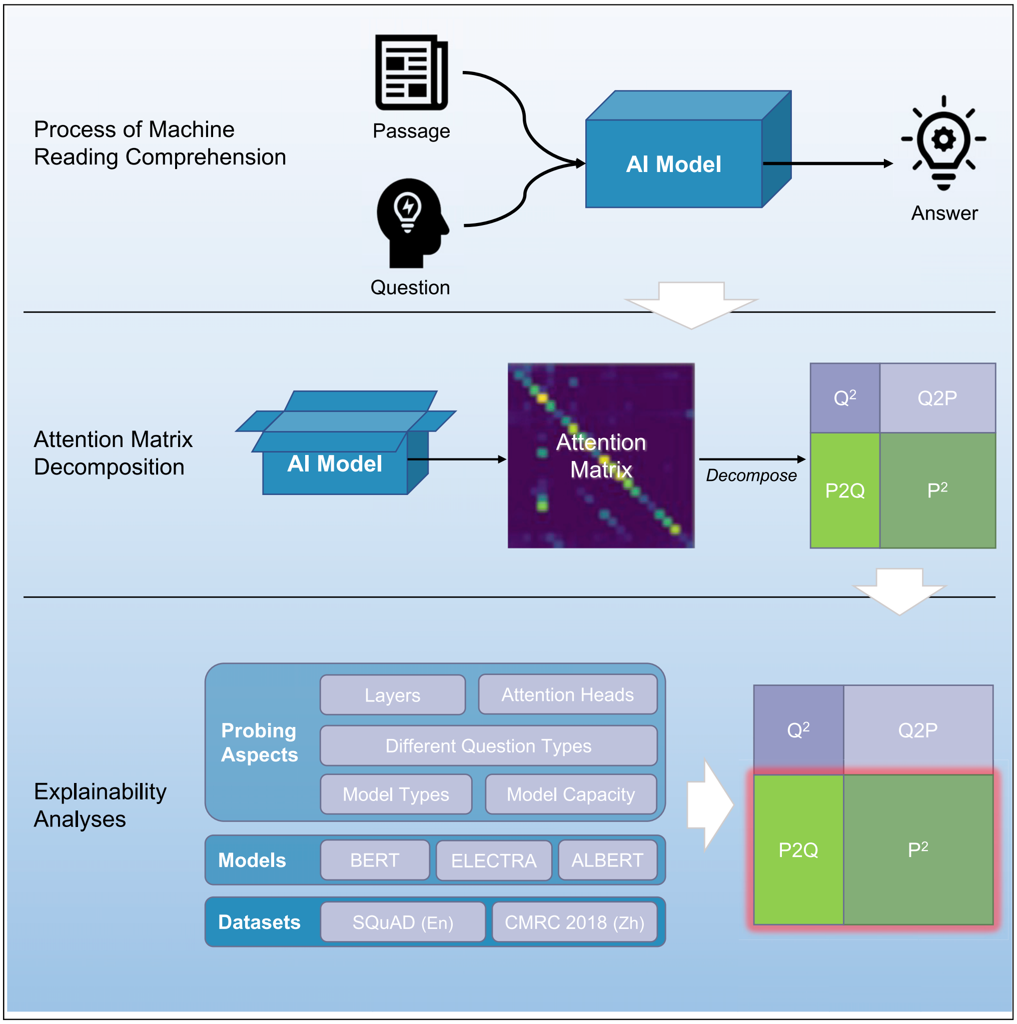
Article

Multilingual multi-aspect explainability analyses on machine reading comprehension models



Yiming Cui, Wei-Nan Zhang, Wanxiang Che, Ting Liu, Zhigang Chen, Shijin Wang

ymcui@ir.hit.edu.cn (Y.C.)
tliu@ir.hit.edu.cn (T.L.)

Highlights

What are the important components that account for explainability of MRC models?

Robust explainability investigation through multilingual and multi-aspect analyses

The focus of attention varies in different layers and heads

Passage-to-question and passage understanding play important roles in MRC

Cui et al., iScience *25*, 104176
May 20, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104176

iScience

CellPress
OPEN ACCESS

Article

Multilingual multi-aspect explainability analyses on machine reading comprehension models

Yiming Cui,[1,2,4,*] Wei-Nan Zhang,[1] Wanxiang Che,[1] Ting Liu,[1,*] Zhigang Chen,[2] and Shijin Wang[2,3]

SUMMARY

Achieving human-level performance on some of the machine reading comprehension (MRC) datasets is no longer challenging with the help of powerful pre-trained language models (PLMs). However, the internal mechanism of these artifacts remains unclear, placing an obstacle to further understand these models. This paper focuses on conducting a series of analytical experiments to examine the relations between the multi-head self-attention and the final MRC system performance, revealing the potential explainability in PLM-based MRC models. To ensure the robustness of the analyses, we perform our experiments in a multilingual way on top of various PLMs. We discover that passage-to-question and passage understanding attentions are the most important ones in the question answering process, showing strong correlations to the final performance than other parts. Through comprehensive visualizations and case studies, we also observe several general findings on the attention maps, which can be helpful to understand how these models solve the questions.

INTRODUCTION

Teaching machines to read and comprehend human language is an important topic in artificial intelligence (AI). Machine reading comprehension (MRC) has been regarded as an important task to test how well the machine comprehends human languages. Machine reading comprehension task is to read and comprehend given passages and answer relevant questions, which is a type of Question Answering (QA) task but focuses more on text comprehension. In the earlier stage, as most of the MRC models (Dhingra et al., 2017; Kadlec et al., 2016; Cui et al., 2017) are solely trained on the training data of individual MRC datasets without much prior knowledge, their performances are not very impressive and are far from humans. In recent years, the pre-trained language model (PLM) has become a new way for text representation. Pre-trained language models utilize large-scale text corpora and self-supervised approaches to learn the text semantics. Various PLMs bring significant improvements to many natural language processing (NLP) tasks, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), ALBERT (Lan et al., 2020), MacBERT (Cui et al., 2021a), etc. With the development of PLMs, many MRC models could outperform human performance on a series of MRC benchmarks, such as SQuAD 1.1 (Rajpurkar et al., 2016) and SQuAD 2.0 (Rajpurkar et al., 2018), indicating that these models can comprehend human languages to a certain extent.

However, achieving human-level prediction performance is not the only goal in AI research. The decision process and the explanation of these AI models remain unclear, raising concerns about their reliability and placing obstacles to achieving controllable and reliable AI. In this context, explainable artificial intelligence (XAI) (Gunning, 2017) becomes more important than ever not only in the NLP field but also in various directions of AI. The goal of XAI is to produce more explainable machine learning (ML) models while preserving a high accuracy of the model prediction. XAI provides a way for humans to understand the intrinsic mechanism of AI models. To improve the AI system's explainability, one could seek decomposability of the conventional machine learning model, such as decision trees, rule-based systems, etc. Moreover, we can also use post-hoc techniques for deep learning models (Barredo Arrieta et al., 2020; Murdoch et al., 2019). However, most of the cutting-edge systems are developed on artificial neural networks, and investigating the explainability of these models is non-trivial. In this context, some researchers advocate using interpretable models instead of explaining black-box machine learning models (Rudin, 2019). Nonetheless, the community has made great efforts on explaining the neural network model's behavior by post-hoc approaches (Cui et al., 2022), probing tasks (Vulić et al., 2020), visualizations (Jain and Wallace, 2019), etc.

[1] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China

[2] State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Beijing 100083, China

[3] iFLYTEK AI Research (Central China), Wuhan 430000, China

[4] Lead contact

*Correspondence: ymcui@ir.hit.edu.cn (Y.C.), tliu@ir.hit.edu.cn (T.L.)

https://doi.org/10.1016/j.isci.2022.104176

However, understanding the intrinsic mechanism of the neural network is still a challenging issue. In the NLP field, most of the models rely on the attention mechanism (Bahdanau et al., 2014) to model the importance of the input text. Later, transformer-based PLMs are becoming a new paradigm to process NLP tasks, whose core component is the multi-head self-attention mechanism (Vaswani et al., 2017). While PLMs achieve excellent performance across various NLP tasks, it is necessary to know what is going on inside the multi-head self-attention mechanism.

As a representative PLM, Bidirectional Encoders from Transformers (BERT) (Devlin et al., 2019) has become a popular testbed for explainability studies. Some researchers conducted analyses to help us better understand the internal mechanism of BERT-based architecture. For example, Kovaleva et al. (2019) discovered that there are repetitive attention patterns across different heads in the multi-head self-attention mechanism indicating its over-parametrization in BERT. Among various research topics on explainability in NLP, perhaps the most trending one is *whether the attention can be treated as explanations*. Unlike the attention in computer vision area, such as using attention heatmap to visualize how machine understands chest radiograph (Preechakul et al., 2022), the explainability of the attention mechanism is still uncertain in NLP. Some researchers argue that attention could not be used as explanation. For example, Jain and Wallace (2019) verify that using completely different attention weights could also achieve the same prediction. However, on the contrary, some works hold positive attitudes about this topic, and they believe that the attention mechanism is a source of explainability (Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020). These works have brought us various views on the attention mechanism in PLMs, but there is still no consensus about this important topic as of now. Also, most of these works only investigate the text classification tasks, which require less reasoning skills and lack a comprehensive understanding of the long text.

Regarding the explainability studies in MRC tasks, Yang et al. (2018) proposed a multi-hop question answering dataset, called HotpotQA. However, unfortunately, most of its following works only focus on improving the system performance without specifically caring about the explainability. Cui et al. (2022) proposed an unsupervised approach to extract evidence span in the passage, which can be seen as a post-hoc explanation. Cui et al. (2021b) proposed a comprehensive benchmark for evaluating the explanations in MRC tasks, including span-extraction MRC and multi-choice MRC for both English and Chinese. However, most of these works mainly focus on the post-hoc explainability approaches, which lack a comprehensive understanding of the internal mechanism of the model itself. Wu et al. (2021) investigated several black-box attacks at the character, word, and sentence level for MRC systems. Overall, a comprehensive and robust explainability investigation of the MRC model is not well studied in the previous literature.

To increase the diversity in better understanding the attention mechanism in PLMs, in this paper, we present an explanatory study specifically for the MRC tasks. Except for the traditional attention visualizations in a layer-wise or head-wise view, we also provide a thorough view with extensive and robust experiments to better understand whether these observations can be generalizable to other PLMs and even for the PLMs in a different language or size. Our contributions are listed as follows.

- We specifically aim to investigate the attention mechanism of PLM-based MRC models in various aspects of the PLMs, including the language, model type, capacity, etc. As far as we know, this is the first work that analyzes the MRC model's explainability in a multilingual and multi-aspect way.
- Through massive analytical experiments, we find that *passage-to-question* and *passage understanding* attention are the most important zones in the attention map, which might be the sources for the model's explainability.
- Several interesting observations are discovered, including model-specific behaviors in attention map, etc., which can be useful in better understanding the internal mechanism of these MRC models when solving the questions.

RESULTS

A new view on attention map: attention zones

Before presenting our analyses, we first present a new view on the attention map in MRC models, which is a crucial component throughout this paper. Formally, MRC tasks consist of three essential parts: passage P, question Q, and answer A. Usually, we concatenate the passage and the question into the pre-trained language model, letting them interact with each other, and finally, the model outputs an answer.

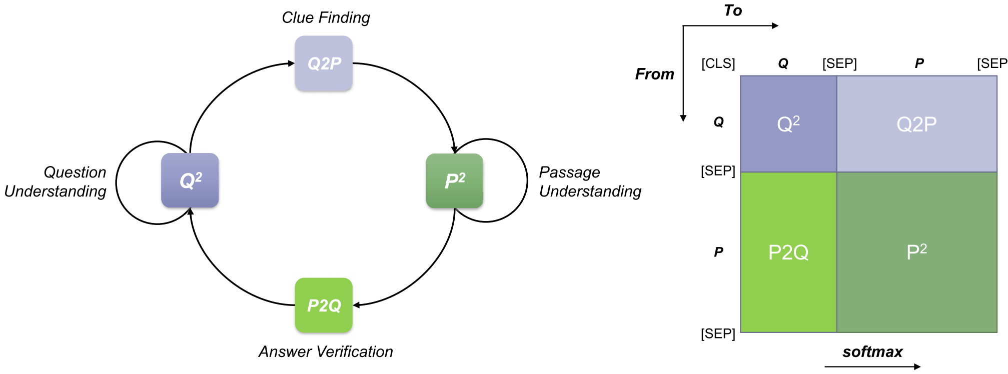


Figure 1. Intuitive explanation of four different attention zones for MRC tasks
Q: Question, P: Passage.

Specifically, the input is organized as follows, where the [CLS] represents the special starting token, and [SEP] represents the special separating token, respectively.

$$\text{[CLS] Question [SEP] Passage [SEP]}$$

Unlike previous works that regard the attention map as a whole, in this paper, we propose to decompose the attention map in a much more precise view, which is specifically designed for MRC tasks. To have a better understanding of the multi-head self-attention in MRC models, we divide the attention map $\boldsymbol{M} \in \mathbb{R}^{L \times L}$ into four areas (where L is the length of the input), namely *attention zones*, as shown in Figure 1. For each part, we give intuitive illustrations as follows (These illustrations may not represent the actual behavior in transformer model but can help us understand them intuitively).

- Q^2: The question is attended to itself, which can be seen as *question understanding* process.
- Q2P: It represents the distributions of passage words in terms of a specific question word, which can be seen as *finding clues using the question* process.
- P2Q: Similar to the Q2P, but in a reverse order, which can be seen as *answer verification* process.
- P^2: The passage text is attended to itself, which can be seen as *passage understanding* process.

In the following sections, we observe the behaviors in different attention zones rather than regard the attention map as a whole. This allows us to understand the attention mechanism in MRC models better.

Experimental setups

In this paper, we aim to analyze the span-extraction MRC, which is one of the most representative MRC tasks. The span-extraction MRC task is to read a passage and answer the relevant question, where the answer is an exact span in the passage. Specifically, we use SQuAD (Rajpurkar et al., 2016) dataset for English and CMRC 2018 (Cui et al., 2019) dataset for Chinese to simultaneously evaluate attention behaviors in both languages. To build MRC models, following previous works, we use BERT (Devlin et al., 2019) as a natural baseline for most of the experiments. We use BERT-base-cased model (https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip) for English and BERT-base (https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip) for Chinese for weight initialization.

Unlike most previous works that only report single-run experimental results, to make our observations more robust and reliable, all experiments are trained and evaluated five times (with different random seeds), and their average scores are reported.

Quantitative study: attention is conditional explanation

Firstly, we investigate the effect of masking some parts in the attention map. Recall that the PLMs show a strong pattern for special tokens ([CLS] and [SEP]) and diagonal tokens in the attention map (Kovaleva et al., 2019), as shown in Figure 2. To examine the effect of these tokens, we mask the special tokens or attention zones during the training phase to see their dependence on the model performance. Masking

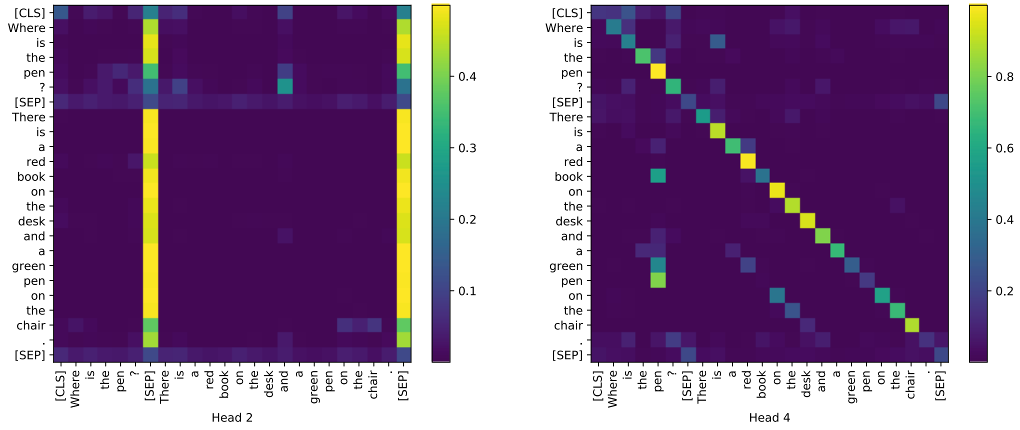


Figure 2. Attention maps of 2nd and 4th head in the last layer of fine-tuned BERT$_{\text{base}}$ on SQuAD

There are strong patterns in diagonal elements and the elements that are related to special tokens.

means to set the whole (or partial) attention map with all the same and large negative numbers (say −10000). After the softmax function, the resulting attention map loses its ability to "highlight" important relations in the respective area. The results are shown in Table 1.

We first look into the results of masking special tokens. Not surprisingly, removing all three special tokens yields a decline in the performance, where it hurts more on SQuAD than CMRC 2018. However, when removing these tokens individually, we did not see a significant drop and even noticed a rise in the performance, such as removing [CLS] in CMRC 2018. Also, when removing diagonal elements, we see a relatively consistent improvement on both datasets, where there is +0.752 EM on CMRC 2018. This indicates that removing higher attention values does not necessarily result in a significant performance drop. When masking specific attention zones, we can see that their performances vary a lot. Removing P^2 and P2Q zones hurts the performance most, while there is no significant drop for the Q2P and Q^2 zones. This demonstrates that the "*from passage to X*" attentions are relatively more important than "*from question to X*" in MRC models. Further discussion will be presented in the next section.

Secondly, we also present the baseline performance when removing the whole or all top-10 elements in each attention zones for all layers. This experiment examines whether there is a significant performance drop for a regular baseline system when a certain attention zone is disabled. As we can see from Table 2, removing the P2Q zone (partial or whole) hurts performance the most, indicating that the key to answer the questions mostly resides in this attention zone. On the contrary, the Q2P zone hurts performance least. This is in line with the observations in Table 1.

Lastly, as most of the previous works use attention scores to present where the model emphasizes, we wonder whether there is a high correlation between the attention score and system performance. In this experiment, we mask top-k^{th} value in different attention zone and calculate the Pearson correlation between its performance and the rank $k \in \{1 \ldots 10\}$.

As we can see from Table 3, not all attention zones correlate well to the system performance. We can see a consistent higher correlation in P2Q and P^2 zones while lower in Q2P zones. This strengthens our claim that a higher attention score does not necessarily contribute more to the performance. This also indicates that rather than performing a rough analysis on the whole attention map, it is necessary to conduct experiments on different attention zones, especially those with higher correlations. Through the experiments above, we conclude that the *attention is conditional explanation* in MRC models. Based on these observations, we proceed with further and deeper analyses on different attention zones to examine their behaviors individually in the rest of the paper.

P2Q and P^2 zones matter most in MRC models

Based on the observations in the previous section, we analyze the attention behavior in different zones in terms of different aspects. We use the baseline system and experimental setups in the previous

Table 1. Results on masking part of attention map

	SQuAD		CMRC 2018	
	EM	F1	EM	F1
Baseline	80.687	88.129	63.796	84.789
No [CLS]	80.802	88.276	64.119	84.858
No Mid [SEP]	80.689	88.082	63.896	84.626
No End [SEP]	80.522	87.959	64.299	84.866
No All	78.956	86.414	63.659	83.945
No Diagonal	80.645	88.241	64.548	84.908
No Q^2	76.395	84.195	60.100	80.625
No Q2P	79.941	87.352	64.517	84.592
No P2Q	12.763	16.355	15.070	18.466
No P^2	34.441	51.792	16.278	42.906

section and decode them under different settings. To make the visualization results comparable, we get the decoding performance (only EM scores are considered) when masking a certain attention zone and calculate the difference to the baseline score for all experiments. Then, we observe the attention behavior in different layers and attention heads in terms of different languages, model's capacity, etc.

Firstly, we look into general situations that disable a certain attention zone in a specific attention head or layer. The layer-wise analysis is depicted in Figure 3. Surprisingly, we find that though the visualizations are made with different datasets and pre-trained language models, two figures look similar in their performance distributions, where we conclude as follows.

- Disabling Q^2 and Q2P does not show significant drops to the overall performance.
- Passage understanding (P^2) starts from the first layer and shows a strong reaction after masking.
- Removing the top-most layer (layer-12 in this case) does not show significant performance drops. This is in accordance with the findings that using the representation of the second most layer for fine-tuning results in a better performance on downstream tasks (Xiao, 2018).

We move onto the head-wise analysis, which is depicted in Figure 4. Except for the strong dependence on P^2 and P2Q zones, the head-wise view does not show a consistent pattern in SQuAD and CMRC 2018, and thus we focus on layer-wise analyses in the following parts.

Overall, based on the visualizations of layer-wise and head-wise analyses, we induce that the model first pays more attention to modeling the passage itself (P^2) to fully understand the text. Moreover, the question information also flows to the passage (Q2P) to indicate where to attend.

Next, a natural question would be *why removing P2Q attention results in a severe performance drop than Q2P?* Both P2Q and Q2P are in the shape of $L_p \times L_q$, and thus it is nothing to do with the area. However, if we take a closer look into their positions in the attention map (Figure 1), we might possibly understand its reason intuitively. The most important thing to keep in mind is that the softmax function is applied in a row-wise manner. Disabling Q2P indicates that "*from question to passage*" attention is removed. The question can only be attended to itself, and the passage can be attended to both passage itself and the question. On the contrary, disabling P2Q indicates that the question can be attended to both question itself and passage, but the passage can only be attended to itself. In MRC tasks, the length of the question (dozens of words) are typically shorter than the length of the passage (several hundreds of words or more). In this context, discarding the P2Q zone harms the performance a lot due to the fact that a large amount of the softmax functions cannot be applied to both the question and passage, resulting in an insufficient interaction between them, which is a crucial process in machine reading comprehension.

Table 2. Results on removing top-*k* and all attention scores

	SQuAD		CMRC 2018	
	All	Top-10	All	Top-10
All	–	66.539	–	57.813
Q^2	40.464	65.272	27.145	58.652
Q2P	77.533	79.743	56.390	63.324
P2Q	4.354	45.790	1.634	43.939
P^2	6.923	78.412	28.565	63.175

Only EM scores are reported.

Harder questions require a deeper understanding in question

We further analyze the attention behavior for different types of questions, which can help us understand their behaviors in a linguistic view. We select the seven most frequent question types in SQuAD: *what, how, who, when, which, where*, and *why*. The visualizations are shown in Figure 5.

As we can see, the attention patterns for different types of questions are quite similar and are similar to the overall attention pattern (Figure 3). Regarding the attention map for "why" questions, all zones show stronger impacts on the performance than other types of questions. Especially, it puts more emphasis on the Q^2 and P^2 attention zones. As "why" questions are relatively harder than the others, the visualization indicates that when solving harder questions, the model focuses more on the question understanding (Q^2) and passage understanding (P^2), which is in line with problem solving process in human view.

Furthermore, CMRC 2018 provides an additional challenge set, which contains the questions that need comprehensive reasoning over multiple sentences. We can also compare the attention map between the normal development set and the challenge set. The results are shown in Figure 6. As we can see, the two figures are quite similar, where the P^2 and P2Q are the most important attention zones. We also discover a stronger focus in the Q^2 zone of the challenge set compared to the counterpart. This observation is similar to Figure 5 ('*why*' questions in SQuAD). Through the visualizations of both languages, the results strengthen our claims that these hard questions (and longer question text (The average length of challenge question is 18 compared to 15 in dev set, described in Cui et al. (2019)) require a deeper understanding of the question in both English and Chinese MRC tasks.

PLM-specific attention behaviors

In the previous analyses, we have observed several interesting and consistent findings. However, *are these observations generalizable to other PLMs as well*? To investigate this question, we perform layer-wise decomposition on another two popular PLMs: ELECTRA (Clark et al., 2020) and ALBERT (Lan et al., 2020). Besides, we also carry out experiments on their large-level model (~ 340M parameters) to compare with their base-level model (~ 110M params).

- ELECTRA (Clark et al., 2020) employs a new generator-discriminator framework that is different from most of the previous PLMs. The generator is typically a small masked language model (MLM) that learns to predict the original words of the masked tokens. The discriminator is trained to discriminate whether the input token is replaced by the generator. In the fine-tuning stage, only the discriminator is used.
- ALBERT (Lan et al., 2020) mainly focuses on designing a compact PLM by introducing two techniques of parameter reduction. The first is the factorized embedding parameterization, which decomposes the embedding matrix into two small matrices. The second one is the cross-layer parameter sharing in the transformer, which significantly reduces the number of parameters. Besides, they also proposed the sentence order prediction (SOP) task to replace the next sentence prediction (NSP).

The results are shown in Figure 7. Overall, the base-level models are much sensitive to the elimination of specific attention zones in several layers. On the contrary, all large-level models yield minor performance loss (depicted in lighter colors) than the counterpart, which indicates that the large-level models are more

Table 3. Pearson correlation of masking top-k^{th} attention score

	SQuAD	CMRC 2018
Q^2	0.624 ± 0.083	−0.316 ± 0.370
Q2P	0.159 ± 0.435	0.134 ± 0.531
P2Q	0.765 ± 0.017	0.778 ± 0.118
P^2	0.534 ± 0.216	0.291 ± 0.299

We report five-run average and its standard deviations.

robust, and the learning of the model is not concentrated to a few attention zones. A possible guess is that with a larger capacity for large-level PLMs, there is redundant knowledge stored in the model. In this way, when a specific attention zone is disabled, the model can still recover such knowledge in other relevant zones, and thus the final performance is not affected that much. By comparing different PLMs, the importance for different zones is as follows.

- BERT: $P^2 > P2Q > Q^2 \approx Q2P$
- ELECTRA: $P^2 > P2Q > Q^2 > Q2P$
- ALBERT: $P^2 > P2Q \approx Q2P \approx Q^2$

As we can see, P^2 and P2Q are the most important attention zones across different sizes and types of PLMs. However, these PLMs show different attention patterns, indicating their distinct ways of processing the text. For BERT and ELECTRA, P^2 is the most important attention zone, followed by P2Q. While for ALBERT, it can be seen that the importance of attention is evenly distributed in each attention head and layer. The main difference between ALBERT and other PLMs is that the parameter of each transformer layer is shared. Thus, the learning for each attention zones is amortized, as changing the parameter in one layer will also change the attention behavior in other layers. In this way, the model could not focus on learning a specific feature at a particular layer or attention head and must be amortized through all layers. Apart from the observations above, we also notice that

- For base-level PLMs, the passage understanding is mostly learned from the bottom layer, but it still progressively learns in the following layers.
- Disabling all attentions in the top layers yields no performance drop and even a minor gain, indicating that there are redundant attention heads that can be pruned without hurting the system performance.
- Disabling all attentions does not necessarily result in worse performance compared to disabling a specific attention zone, and vice versa, such as in the 6th layer of $\text{ELECTRA}_{\text{base}}$ and the 14th and 16th layer of $\text{ELECTRA}_{\text{large}}$. This indicates the interaction complexities between different attention zones.

DISCUSSION

In the previous sections, we perform quantitative analyses on the proposed attention zones to explore their behaviors in MRC models. To further investigate how these attention zones affect the machine reading process, in this section, we come back to visualize the attention map and look into specific examples to analyze the potential behavior of MRC models. Based on our findings in the previous sections, we visualize the multi-head self-attention to explicitly discover how the model processes the MRC example. To make the visualization clear, we discard attention values that connect to [CLS] and two [SEP] special tokens, which have great attention values but do not provide helpful hints on understanding the explainability of the MRC model.

Here, we use a simple example to examine the attention behavior. The passage is "*There is a red book on the desk and a green pen on the chair.*", and the question is "*Where is the pen?*" We omit the full picture of attention maps in all transformer layers and only show the 11th and 12th layer of $\text{BERT}_{\text{base}}$ trained on SQuAD, as shown in Figure 8. It can be seen that the attention patterns are not fixed for a specific head. For example, the 12th head shows a strong "*all-to-question*" pattern in layer 11, where the majority of

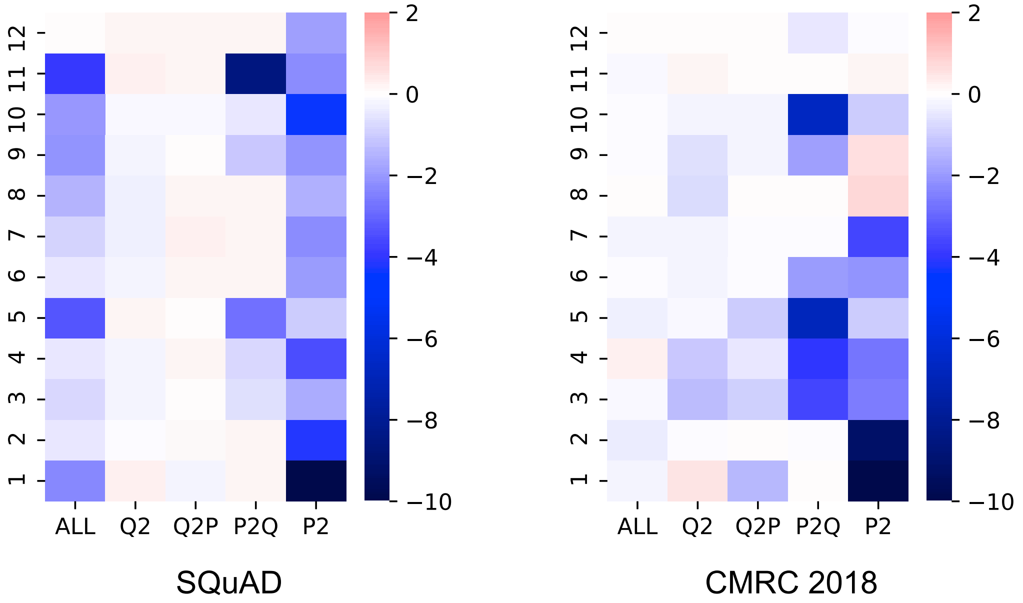


Figure 3. Layer-wise analyses in different attention zones for SQuAD and CMRC 2018

The lighter color means the performance is near the baseline, while darker color means a bigger gap to the baseline (red: above baseline, blue: below baseline).

the lines are connected to the top right. In contrast, it shows an "*identity-mapping*" in layer 12, where there are many horizontal lines, meaning the words are connected to themselves. Thus, it is not feasible to select a fixed set of attention heads for explainability evaluation across different layers.

In this context, to get a closer look, we manually select the 3-4-8-9th heads for layer 11 and 3-4-6-12th heads for layer 12 to present how the model solves the question in MRC. We visualize the attention distribution in terms of the question word "*why*" in attention source and target, which is depicted in Figure 9. Through observations, we found a strong indication of explainability connected to the question word. By comparing the attentions in the 11th and 12th layer, we can see that the question word "*why*" shows strong attention to other words in layer 11 while it gets weak in layer 12. However, in layer 12, we can see that the phrase "*on the chair*" in the passage attends to the question word "*where*" (P2Q), and the word "*where*" attends to both "*on the desk*" and "*on the chair*" (Q2P) (Here, we mean a relatively higher attention value than the others). We induce that the answer is obtained by taking both P2Q and Q2P attention zones into account, and thus the model champions the phrase "*on the chair*" as the final answer. This is also observed in the previous section that removing attentions in layer 12 does not yield significant performance drops, as a similar pattern can also be observed in layer 11.

Besides, another interesting observation is that the attention values are not only higher for the start and end position of the answer span but also the words between them. It can be seen that the word "*why*" does not only attend to the word "*on*" and "*chair*", but the whole phrase "*on the chair*" in both layer 11

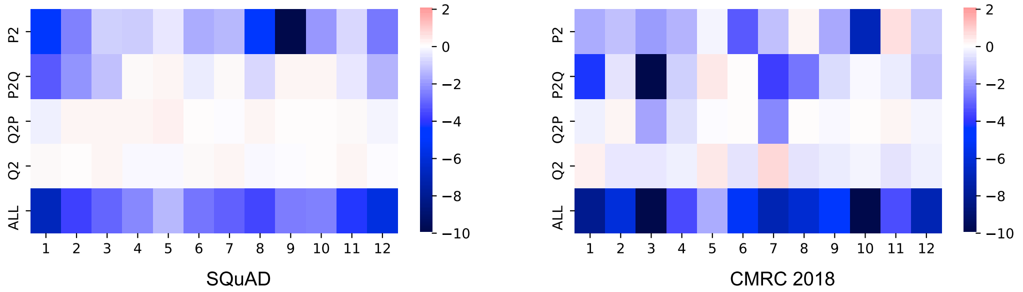


Figure 4. Head-wise analyses in different attention zones for SQuAD and CMRC 2018

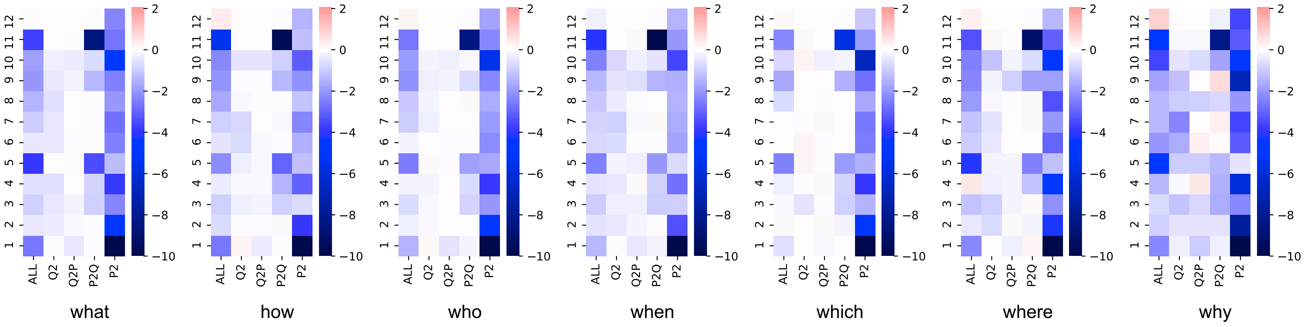


Figure 5. Analyses of different question types for SQuAD

The number of each question type (in order): what (6073), how (1389), who (1377), when (864), which (747), where (508), and why (158).

and 12. This implies that though the answer span is extracted by the start and end pointers, the MRC models are capable of considering the words between them to make final answer predictions, but not solely on the start and end tokens. Perhaps, this is why almost all span-extraction MRC systems are not modeled in a sequence tagging manner, as the words between start and end positions are already considered in the transformer.

Based on our findings, there might be two directions for future works. First, we will try to find many-to-many mappings in the attention map, which is much important to the questions that need comprehensive reasoning. Also, we will find a way to automatically discard the attention head that contributes less to the final system performance, as not all attention heads are important in transformer models.

Limitations of the study

In this paper, we have discussed the potential explainability within machine reading comprehension models. Though we have strived to make our analyses as comprehensive as possible, there may have several limitations that need to be studied in future work.

- Explainability for other languages: We have studied the explainability in both English and Chinese models, which is a step forward to increase the language diversity, as these two languages belong to different language genres. The conclusions made in this paper might have good generalizations

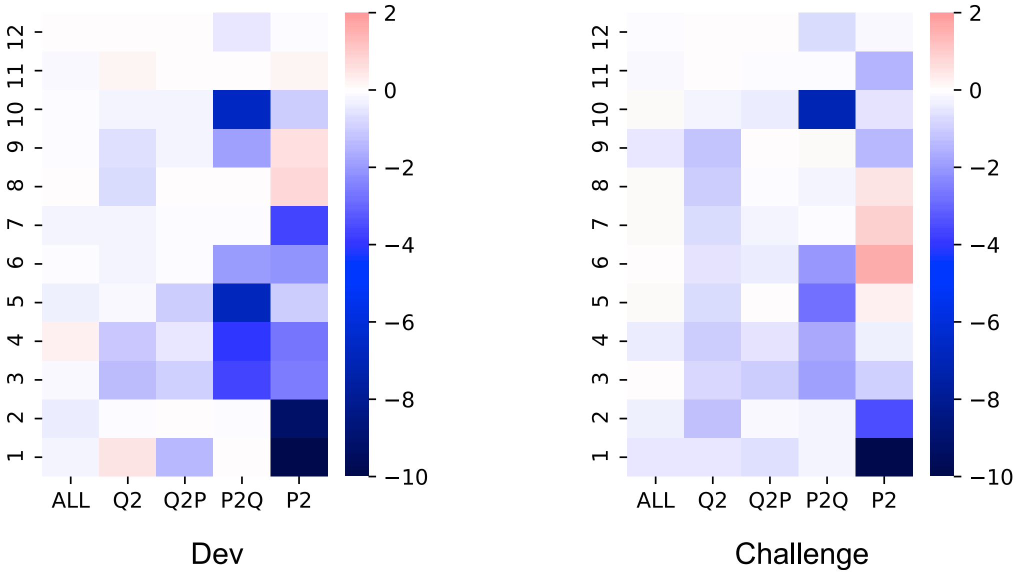


Figure 6. Comparison of the development and challenge set for CMRC 2018

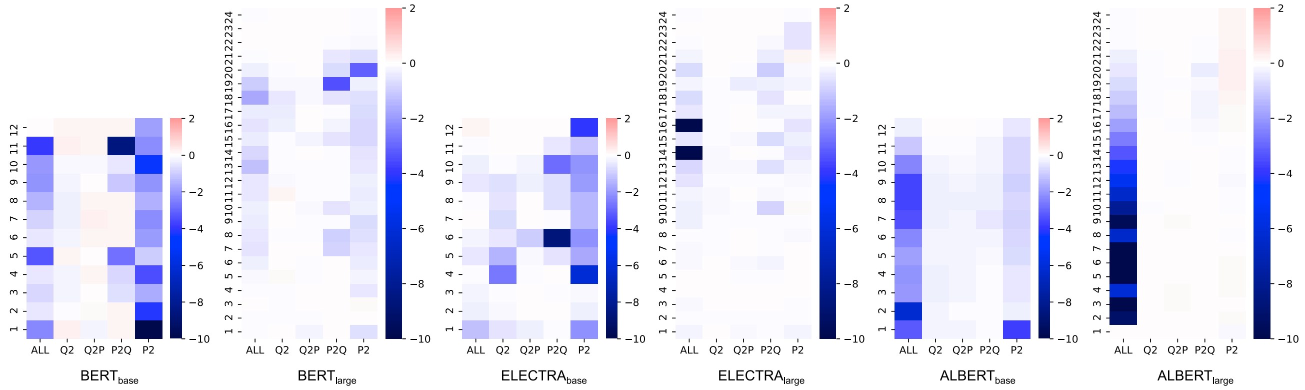


Figure 7. Analyses of base-level and large-level BERT, ELECTRA, and ALBERT for SQuAD

than monolingual experiments. However, it is not sure whether our analytical conclusions are generalizable to other genres of languages, such as Arabic.

- Explainability for other models: Though we have found several common phenomena as shown in the visualizations, different pre-trained language models exhibit different patterns in the attention map, especially for those with different neural architecture (such as BERT v.s. ALBERT). It is interesting to see how other PLMs perform in a similar context.
- Different ways to examine the attention mechanism: Using attention values or importance scores have been a normal way to visualize the attention map. This paper provides a different way to examine the attention map by using system performance. With the development of XAI studies, it is promising to have a more efficient way to analyze the attention map.

As the explainability of machine learning approaches is still an ongoing research topic, we hope that such limitations can be further studied in future work to help us better understand the internal mechanism of machine reading comprehension models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability

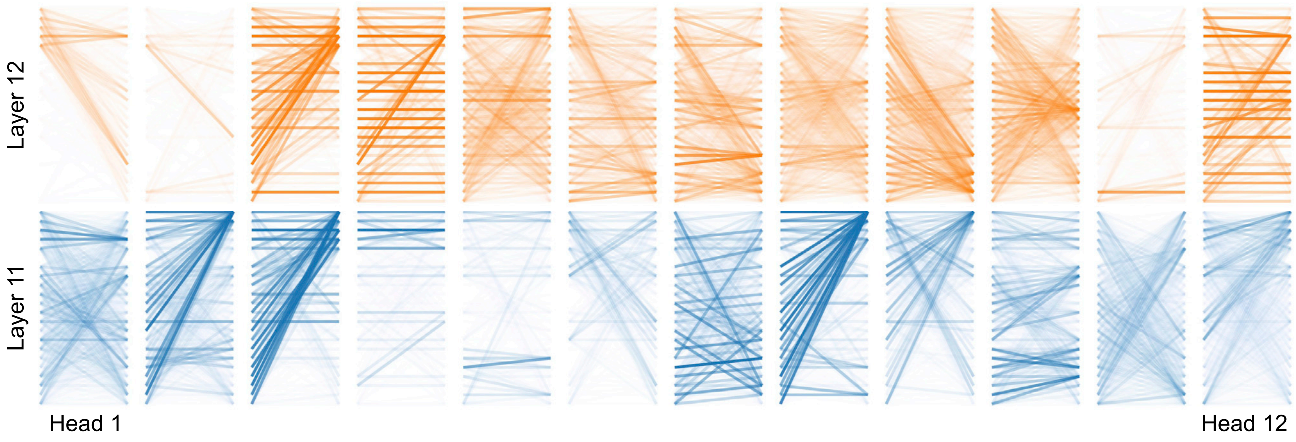


Figure 8. Attention maps for different attention heads in layer 11 and 12

Recall that the input is created by the concatenation of "[CLS] Q [SEP] P [SEP]". The darker line means a strong connection between two words.

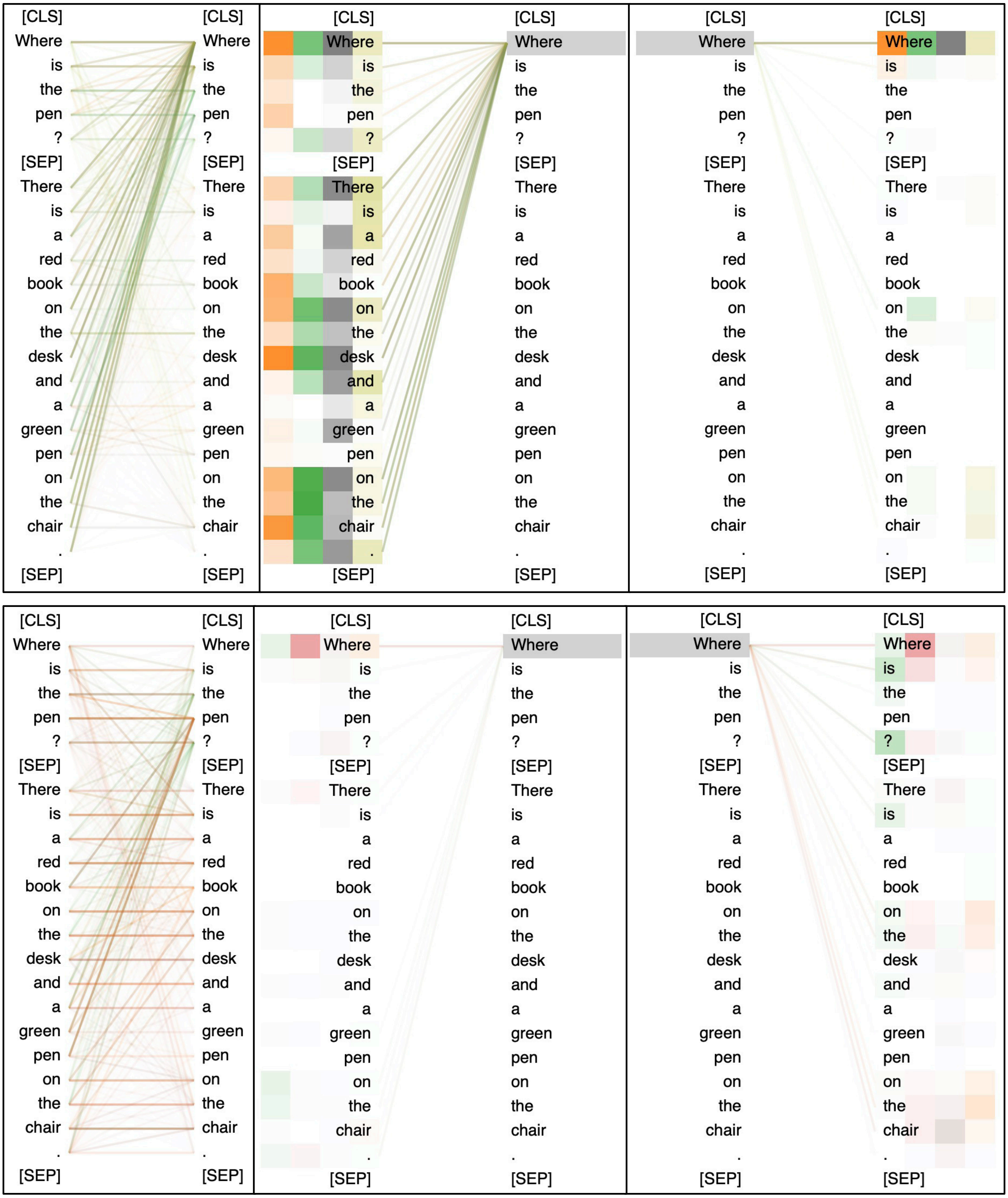


Figure 9. Visualization of specific attention heads in layer 11 (upper) and 12 (lower) of SQuAD model

The four boxes behind each word represent the values of each attention head (the higher value the darker).

ACKNOWLEDGMENTS

We would like to thank Dr. Kanudha Sharda (Associate Editor) and three anonymous reviewers for their constructive comments to improve our paper. This work is supported by the National Key Research and Development Program of China via grant No. 2018YFB1005100. Yiming Cui would like to thank TPU Research Cloud (TRC) program for Cloud TPU access.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.C.; Methodology, Y.C.; Formal Analysis, Y.C., W.Z., W.C., S.W.; Writing - Original Draft, Y.C.; Writing - Review & Editing, Y.C., W.Z., W.C., S.W.; Visualization, Y.C.; Supervision, T.L.; Funding Acquisition, Z.C., S.W.; All authors read and approved the submission of this paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 2, 2021
Revised: March 1, 2022
Accepted: March 25, 2022
Published: May 20, 2022

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 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Preprint at arXiv. https://doi.org/10.48550/arXiv:1409.0473.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf. Fusion *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012. https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Bastings, J., and Filippova, K. (2020). The elephant in the interpretability room: why use attention as explanation when we have saliency methods? In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Online, pp. 149–155. https://doi.org/10.18653/v1/2020.blackboxnlp-1.14. https://www.aclweb.org/anthology/2020.blackboxnlp-1.14.

Clark, K., Luong, M.T., Le, Q.V., and Manning, C.D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In ICLR. https://openreview.net/pdf?id=r1xMH1BtvB.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021a). Pre-training with whole word masking for Chinese bert. IEEE/ACM Trans. Audio, Speech, Lang. Process. *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365.

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), Association for Computational Linguistics, pp. 593–602. https://doi.org/10.18653/v1/P17-1055. https://aclanthology.org/P17-1055.

Cui, Y., Liu, T., Che, W., Chen, Z., and Wang, S. (2021b). ExpMRC: explainability evaluation for machine reading comprehension. Preprint at arXiv. https://doi.org/10.48550/arXiv:2105.04126.

Cui, Y., Liu, T., Che, W., Chen, Z., and Wang, S. (2022). Teaching machines to read, answer and explain. IEEE/ACM Trans. Audio Speech Lang. Process. https://doi.org/10.1109/TASLP.2022.3156789.

Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., Wang, S., and Hu, G. (2019). A span-extraction dataset for Chinese machine reading comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp. 5886–5891. https://doi.org/10.18653/v1/D19-1600.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dhingra, B., Liu, H., Yang, Z., Cohen, W., and Salakhutdinov, R. (2017). Gated-attention readers for text comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp. 1832–1846. https://doi.org/10.18653/v1/P17-1168. https://aclanthology.org/P17-1168.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI) (Defense Advanced Research Projects Agency (DARPA), nd Web 2).

Jain, S., and Wallace, B.C. (2019). Attention is not Explanation. In Proceedings 2019 Conference North America Chapter Association Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3543–3556. https://

doi.org/10.18653/v1/N19-1357. https://www.aclweb.org/anthology/N19-1357.

Kadlec, R., Schmid, M., Bajgar, O., and Kleindienst, J. (2016). Text understanding with the attention sum reader network. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, pp. 908–918. https://doi.org/10.18653/v1/P16-1086. https://aclanthology.org/P16-1086.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. (2020). Captum: a unified and generic model interpretability library for pytorch. Preprint at arXiv. https://doi.org/10.48550/arXiv:2009.07896.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp. 4365–4374. https://doi.org/10.18653/v1/D19-1445. https://www.aclweb.org/anthology/D19-1445.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: a lite bert for self-supervised learning of language representations. In International Conference on Learning Representations (ICLR 2020). https://openreview.net/forum?id=H1eA7AEtvS.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: a robustly optimized bert pretraining approach. Preprint at arXiv. https://doi.org/10.48550/arXiv:1907.11692.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. U S A. *116*, 22071–22080. https://doi.org/10.1073/pnas.1900654116. https://www.pnas.org/doi/abs/10.1073/pnas.1900654116.

Preechakul, K., Sriswasdi, S., Kijsirikul, B., and Chuangsuwanich, E. (2022). Improved image classification explainability with high-accuracy heatmaps. iScience *25*, 103933. https://doi.org/10.1016/j.isci.2022.103933.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, pp. 784–789. https://doi.org/10.18653/v1/P18-2124.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 2383–2392. https://doi.org/10.18653/v1/D16-1264.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Machine Intelligence *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, pp. 37–42. https://doi.org/10.18653/v1/P19-3007. https://www.aclweb.org/anthology/P19-3007.

Vulić, I., Ponti, E.M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp. 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586.

Wiegreffe, S., and Pinter, Y. (2019). Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp. 11–20. https://doi.org/10.18653/v1/D19-1002. https://www.aclweb.org/anthology/D19-1002.

Wu, W., Arendt, D., and Volkova, S. (2021). Evaluating neural model robustness for machine comprehension. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, pp. 2470–2481. https://doi.org/10.18653/v1/2021.eacl-main.210. https://aclanthology.org/2021.eacl-main.210.

Xiao, H. (2018). bert-as-service. https://github.com/hanxiao/bert-as-service.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C.D. (2018). HotpotQA: a dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, pp. 2369–2380. https://doi.org/10.18653/v1/D18-1259. https://www.aclweb.org/anthology/D18-1259.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
SQuAD Training Set	https://rajpurkar.github.io/SQuAD-explorer/dataset/train-v1.1.json	Version 1.1
SQuAD Dev Set	https://rajpurkar.github.io/SQuAD-explorer/dataset/dev-v1.1.json	Version 1.1
CMRC 2018 Training Set	https://github.com/ymcui/cmrc2018	N/A
CMRC 2018 Dev Set	https://github.com/ymcui/cmrc2018	N/A
Software and algorithms		
Python	https://www.python.org	Version 3.8
TensorFlow	https://tensorflow.org	Version 1.15
Matplotlib	https://matplotlib.org	Version 3.4.0
Captum	https://github.com/pytorch/captum	Version 0.4.0
Bertviz	https://github.com/jessevig/bertviz	Version 1.1.0
MRC Model Analysis	https://github.com/ymcui/mrc-model-analysis	N/A
Other		
Chinese BERT-base	https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip	N/A
English BERT-base-cased	https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip	N/A
English ALBERT-base	https://storage.googleapis.com/albert_models/albert_base_v1.tar.gz	N/A
English ALBERT-large	https://storage.googleapis.com/albert_models/albert_large_v1.tar.gz	N/A
English ELECTRA-base	https://storage.googleapis.com/electra-data/electra_base.zip	N/A
English ELECTRA-large	https://storage.googleapis.com/electra-data/electra_large.zip	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and/or reagents should be directed to and will be fulfilled by the lead contact, Yiming Cui (ymcui@ir.hit.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Code: The source codes for the main experiments are publicly available on GitHub at https://github.com/ymcui/mrc-model-analysis.
- Dataset: All datasets used in this paper are publicly available, listed in the key resources table.
- Additional information: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

METHOD DETAILS

Datasets

In this paper, we mainly conduct our experiments on two span-extraction machine reading comprehension datasets.

- **SQuAD** (Rajpurkar et al., 2016): This is the first span-extraction MRC dataset with over 100K samples. The dataset is constructed by English Wikipedia pages. SQuAD has been a leading benchmark in MRC research.
- **CMRC 2018** (Cui et al., 2019): This is also a span-extraction MRC dataset but in Chinese. The dataset is constructed by Chinese Wikipedia and with 10K human-annotated questions. Besides traditional train/dev/test splits, CMRC 2018 also contains a challenge set consisting of hard questions.

Probing method

Pre-trained language model, such as BERT (Devlin et al., 2019), mainly comprises stacked multi-head self-attention layers with several dense layers. Given a hidden representation $H\in\mathbb{R}^{n\times d}$ (n for length of input and d for hidden dimension), the model first uses three dense layers to transform H into query, key, and value representations.

$$Q = HW^{Q}\ , W^{Q}\in\mathbb{R}^{d\times d} \quad \text{(Equation 1)}$$

$$K = HW^{K}\ , W^{K}\in\mathbb{R}^{d\times d} \quad \text{(Equation 2)}$$

$$V = HW^{V}\ , W^{V}\in\mathbb{R}^{d\times d} \quad \text{(Equation 3)}$$

Then we calculate the dot product of query and key representations and apply softmax function to get the attention map $M\in\mathbb{R}^{n\times n}$ (d_a is the dimension of each attention head), indicating the correlations between each input token.

$$M' = \frac{1}{\sqrt{d_a}}QK^{\top} \quad \text{(Equation 4)}$$

$$M = \mathrm{softmax}(M') \quad \text{(Equation 5)}$$

Finally, the dot product of attention matrix M and value representation V is calculated as the final self-attended representation H'.

$$H' = MV \quad \text{(Equation 6)}$$

To examine the effect of each attention zone, we perform masking on the attention matrix (before softmax activation) M'. For example, if we choose to mask Q^2 (upper left part in Figure 1), the values in Q^2 zone will be set to a big negative value (in this paper, we set as -10000). After the softmax function, these negative values will be normalized to values close to zero, demonstrating that this area is disabled.

Evaluation metrics

For span-extraction MRC tasks, there are two evaluation metrics: exact match (EM) and F1.

- **EM**: This is to measure the exact match between the prediction and the ground truth. An exact match will give a score of 1, otherwise 0.
- **F1**: This is to measure the text overlap between the prediction and the ground truth. If there are more words overlapping, F1 will be close to 1, otherwise 0.

Hyperparameters and detailed setups

The implementation is performed on the official fine-tuning script based on TensorFlow (Abadi et al., 2016)(https://github.com/google-research/bert). All models are trained for three epochs with a universal initial learning rate of 3e-5 and batch size of 64. We set other hyper-parameters as default. For visualizations, we use bertviz (Vig, 2019) and captum (Kokhlikyan et al., 2020). All experiments are carried out on Cloud TPUs v2 (64G HBM) or v3 (128G HBM), depending on the magnitude of the model.

QUANTIFICATION AND STATISTICAL ANALYSIS

Each experiment was repeated five times with different random seeds (and led to different weight initializations) to make the analysis more robust. We mainly report the average scores for five runs and report its standard deviation whenever necessary.