An Analysis of BERT in Document Ranking

Jingtao Zhan
BNRist, DCST, Tsinghua University
jingtaozhan@gmail.com

Jiaxin Mao
BNRist, DCST, Tsinghua University
maojiaxin@gmail.com

Yiqun Liu*
BNRist, DCST, Tsinghua University
yiqunliu@tsinghua.edu.cn

Min Zhang
BNRist, DCST, Tsinghua University
z-m@tsinghua.edu.cn

Shaoping Ma
BNRist, DCST, Tsinghua University
msp@tsinghua.edu.cn

ABSTRACT

Although BERT has shown its effectiveness in a number of IR-related tasks, especially document ranking, the understanding of its internal mechanism remains insufficient. To increase the explainability of the ranking process performed by BERT, we investigate a state-of-the-art BERT-based ranking model with focus on its attention mechanism and interaction behavior. Firstly, we look into the evolving of the attention distribution. It shows that in each step, BERT dumps redundant attention weights on tokens with high document frequency (such as periods). This may lead to a potential threat to the model robustness and should be considered in future studies. Secondly, we study how BERT models interactions between query and document and find that BERT aggregates document information to query token representations through their interactions, but extracts query-independent representations for document tokens. It indicates that it is possible to transform BERT into a more efficient representation-focused model. These findings help us better understand the ranking process by BERT and may inspire future improvement.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Retrieval models and ranking*; Language models.

KEYWORDS

neural networks, explainability, document ranking

ACM Reference Format:
Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. An Analysis of BERT in Document Ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401325

1 INTRODUCTION

In recent years, neural models have been widely used in the field of information retrieval. As a widely-adopted neural language model, BERT [4] is also used in the document ranking task. After being pretrained on large corpus and finetuned on supervised data, BERT can achieve promising results in ranking tasks (e.g. [3, 10]). On the MS MARCO [9] Passage Ranking leaderboard, BERT is adopted by most top performers.

Since Jain et al. [7] argued that attention largely do not provide meaningful "explanations", researches [2, 5, 8, 11] have been trying to analyze BERT, which is based solely on attention mechanisms.

However, the explanation of BERT in the ranking task has not been fully studied. First, several studies [2, 8] observed a surprisingly large amount of attention focusing on "[CLS]", "[SEP]" and periods, which is not fully understood. Second, Qiao et al. [11] designed a representation-focused [6] BERT ranker. Because of its poor performance, they suggested that BERT shouldn't be used as a representation model. However, we believe this baseline is too simple, so whether and how BERT can learn good representations for queries and documents is not thoroughly investigated. To summarize, we want to address the following research questions:

- **RQ1:** How does BERT distribute attention to special tokens and periods in document ranking?
- **RQ2:** How does BERT represent queries and documents and model the interactions between them?

To address these research questions, we adopt three analysis methods. First, an attribution technique [12] is used to study the token importance in different layers. Second, several probing classifiers [1] are trained to study the relevance signal carried by the token representations. Third, we compare the performance of BERT when its attention matrix is masked in different ways to investigate the importance of interactions. Our contributions are as follows:

- We show that BERT dumps redundant attention weights on "[CLS]", "[SEP]" and periods due to their high document frequency, which is a potential threat to the model robustness.
- We demonstrate that BERT extracts representations for query and document in the beginning. Then it captures interaction signals to learn context-specific representations. In the last few layers, BERT relies heavily on the interactions to predict relevance.
- We show that the extracted representations for document tokens are largely query-independent. It reveals the possibility to transform BERT to a representation-focused model.

2 RELATED WORK

2.1 Ranking with BERT

BERT uses standard Transformer [13] architecture. Several studies (e.g. [3, 10]) investigated how to utilize BERT in ranking. A common approach is to construct the model input by concatenating the query

*Corresponding author

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401325

and document text. One "[CLS]" token is added to the beginning and two "[SEP]" tokens are added to mark the ends of query and document text. The output embedding of the "[CLS]" token is used as a representation of the query-document pair and is fed into a multi-layer perception (MLP) to predict the relevance. We refer readers to [4, 10, 13] for more details.

2.2 Analysis of BERT

Explainability of neural models has drawn much attention from researchers. Many methods have been proposed. Here we highlight relevant studies that try to explain how BERT works.

Clark et al. [2] found that BERT's certain attention heads correspond well to linguistic notions of syntax and coreference. Ethayarajh et al. [5] found that deeper layers produce more context-specific representations. Qiao et al. [11] showed that BERT is a strong interaction-focused seq2seq matching model.

The differences between our work and the previous studies are as follows. First, Clark et al. [2] argued that the attention to "[SEP]" token is a sort of "no-op". We generalize such conclusion to "[CLS]" and periods, and attribute it to their high document frequency. Moreover, we demonstrate such behavior may hurt the model robustness. Second, we find that the representations of document tokens are largely query-independent, which shows great promise to improve BERT's efficiency. This finding is, to some degree, different from what Qiao et al. [11] implied. Besides, we investigate interaction behavior in different layers, which, to the best of our knowledge, has not been studied before.

3 METHODS

We describe our experimental setup in Section 3.1 and surface attention pattern in Section 3.2. Then we elaborate the analysis methods and results in Section 3.3, 3.4, and 3.5 [1]. This paper uses "from A to B" or "A ⟶ B" to refer to the attention score that weights A representation for B's context vector. When it comes to the whole model, "model's attention towards A" refers to "the average attention score from A to any token".

3.1 Experiment Setup

3.1.1 Dataset. Our experiments are conducted on MS MARCO [9] Passage Ranking task, which is to rank the passages according to their relevance to a given query. It includes 8.8 million passages, 0.5 million training queries, and about 6800 queries for validation and evaluation, respectively. A standard BM25 model is run to produce 1000 candidate passages for each query. In order to maintain consistent terminology throughout this paper, we refer to these basic units of retrieval as "documents".

3.1.2 Model. Nogueira et al. [10] proposed a BERT-based ranking model and advanced state of the art result by 27% (relative) in MRR@10 on MS MARCO Passage Ranking task. We adopt their finetuned BERT-Base model.

3.1.3 Minor modifications. Several minor modifications are made to reduce the inference cost of BERT. We run BM25 [14] to select top 100 candidate documents per query and the recall of the candidates drops from 81.5% to 67.1%. We limit the input to 256 tokens instead of 512, resulting 3‰ of the documents truncated during evaluation. Despite the large drop in recall of the candidates, our final performance on validation set is 0.334 in MRR@10, which is still comparable to 0.347 reported by Nogueira et al. [10]. We believe such a difference in overall performance won't affect our investigation on the behavior of the BERT-based ranking model.

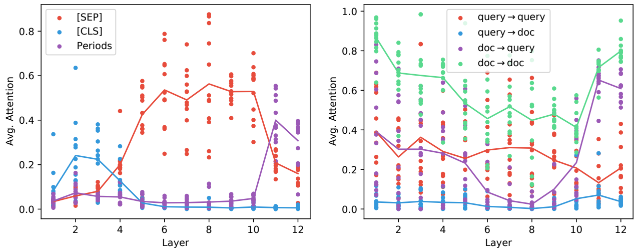

Figure 1: Average attention distribution in different layers.

3.2 Attention Patterns

Following Clark et al. [2], we calculate the average attention score from "[CLS]", "[SEP]", or periods to any token. Considering that the sum of attention scores is 1, they consume a major proportion of attention. We also investigate the average attention distribution of query and document token. The results are shown in Figure 1 [2]. Each point corresponds to a particular attention head.

3.3 Attribution

Attribution method aims to provide interpretation for the model prediction by attributing it to model's input features, e.g., words in a text classification task or pixels in an object recognition task. The attribution score indicates the contribution of each feature. We employ an attribution technique called Integrated Gradients [12].

Formally, $F : \mathbb{R}^{n \times dim} \longrightarrow [0, 1]$ is a function representing BERT-based ranking models, which takes input of $x = (x_1, ..., x_n) \in \mathbb{R}^{n \times dim}$ and outputs a relevance prediction. x_i is the embedding vector of token$_\mathrm{i}$. An attribution of the relevance prediction to input x relative to a baseline input x' is a vector $Attr_F(x, x') = (a_1, ..., a_n) \in \mathbb{R}^n$ where a_i is the contribution of x_i to the prediction $F(x)$ and $\sum_{i=1}^{n} a_i = F(x) - F(x')$.

Token representations are more context-specific in deeper layers, according to Ethayarajh et al. [5]. Thus, the contribution of tokens may vary as the layer gets deeper. We calculate attribution scores on different layers as follows. First, we run network F at the input x and baseline input x' to acquire the output token representations y_i and y_i' of the target i^{th} layer. Then, we think of what's behind this layer as a new model f_i and calculate $Attr_{f_i}(y_i, y_i{}')$. We define $Attr_{f_0}(y_0, y_0{}') := Attr_F(x, x')$.

Attributions are defined relative to an uninformative input called the baseline, which, in our experiments, are empty query baseline and empty document baseline. They are implemented by replacing query/document with padding tokens. Take empty document baseline for instance. We can evaluate how much document information is aggregated to query's representations by examining the attribution of query in different layers. We refer readers to [12] for detailed interpretation.

[1] The code is released at https://github.com/jingtaozhan/bert-ranking-analysis.

[2] The legend is wrong in the version we submitted to SIGIR2020

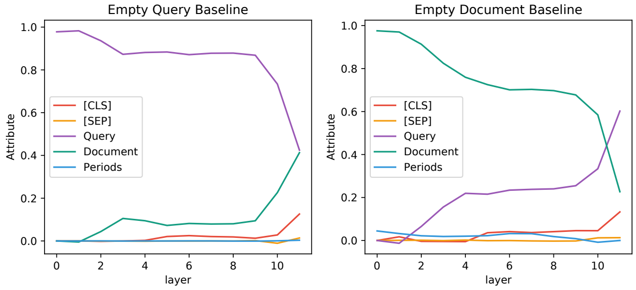

Figure 2: Attribution on different layers

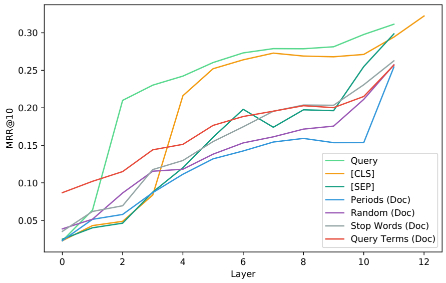

Figure 3: Ranking Performance for different layer's output vector representations of different kinds of tokens.

We perform the experiment on the relevant query-document pairs in the validation set, which should also be classified correctly by BERT ($P(relevant) > 0.5$). The results are shown in Figure 2.

3.4 Probing

Probing classifiers [1] are often used to investigate the internal vector representations of a model. They are simple neural networks that take the vector representations as input and are trained to do a supervised task. If a probing classifier achieves high accuracy, it suggests that the input representations contain much information for the task.

We implement our probing classifiers as simple multi-layer perception (MLP) networks and train them to do the same ranking task. We probe vector representations of special tokens, query tokens, and different kinds of tokens sampled from documents. Because the last layer is only trained to acquire a representation of the query-document pair, we only show the result of "[CLS]" token for the last layer. The embedding module is regarded as the 0^{th} layer. The ranking performance on validation set is shown in Figure 3.

3.5 Mask

BERT utilizes attention mechanism to model the interactions among tokens, as shown in equation 1 [13].

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T/\sqrt{d_k})V \qquad (1)$$

If some interactions are unimportant, masking corresponding attentions won't influence ranking performance much. Each mask experiment will generate a mask matrix $M \in \{0,1\}^{n\times n}$. $M_{i,j}$ corresponds to the attention from token_j to token_i and is 1 if the attention is to be masked, otherwise 0. We also design two mask methods as follows.

Table 1: Results of the Period Mask Experiment.

Method	Comma	Token Mask	Attention Mask	Original
MRR@10	0.299 (-10%)	0.308 (-8%)	0.328 (-2%)	0.334

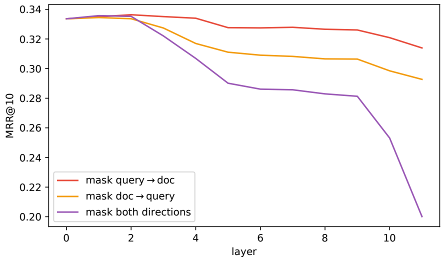

Figure 4: Ranking performance of Query-Doc Interaction Mask Experiment.

Token Mask. As shown in equation 2, the masked attention values are diverted to others by softmax. We use this method to mask tokens entirely, so it is called Token Mask.

$$\text{Attention}(Q,K,V) = \text{softmax}((QK^T - \text{INF}\cdot M)/\sqrt{d_k})V \qquad (2)$$

Attention Mask. As shown in equation 3, the values for the masked attentions are set to zero outside softmax function. So the sum of attention scores is less than 1 for the influenced tokens. The unmasked attentions are not directly affected. In this way, BERT can dump redundant attention weights on the target tokens but ignores information from them.

$$\text{Attention}(Q,K,V) = (1-M)\cdot\text{softmax}(QK^T/\sqrt{d_k})V \qquad (3)$$

3.5.1 Adversary Period Mask. We follow an adversary attack style by only processing the relevant query-document pairs. The mask matrix M is generated by masking attention from periods to any token. Apart from adopting above two mask methods, namely Token Mask and Attention Mask, we also conduct an experiment by replacing periods with commas. The results are shown in Table 1.

3.5.2 Query-Doc Interaction Mask. This experiment studies the importance of the interactions between query and document. The performance loss is an indicator of how important the removed interactions are.

The mask matrix M is generated by masking the attention from query to document, from document to query, or both directions. Correspondingly, attentions from special tokens are also masked in case query and document interact via these tokens. We adopt Attention Mask method to avoid performance loss due to redundant attention issue (Section 4.1.2). We do mask only in the first i layers. When $i == 0$, we report the original performance. The results on validation set are shown in Figure 4.

4 DISCUSSION

4.1 Attention to special tokens and periods

As Figure 1 shows, BERT distributes a significant amount of attention to special tokens ("[CLS]", "[SEP]") and periods. In this Section, we analyze this behavior to address **RQ1**.

4.1.1 Why this happens? Clark et al. [2] showed attention to "[SEP]" is a sort of "no-op" through qualitative analysis and gradient-based measures of feature importance. This observation may generalize to "[CLS]" and periods. We speculate that the attention towards "[CLS]", "[SEP]", and periods is model's redundant attention.

This speculation is supported by the Attribution experiment (Figure 2) and Probing experiment (Figure 3), where we find that the attention score is negatively correlated with the attribution score and probing performance. For instance, attention to periods is highest in the 11^{th} layer, compared to the locally minimal ranking performance of period representations output from 10^{th} layer. After more relevance signal is aggregated to periods in the next layer, the attention to periods drops. Attribution experiment shows the same trend. These results indicate that although "[CLS]", "[SEP]", and periods attract a large proportion of attention, they carry little relevance information.

4.1.2 What's the risk? We find BERT is not robust to some small changes with little effect on the semantics of input, which is caused by dumping redundant attention weights to a non-special token, period. We consider periods having two responsibilities, namely absorbing redundant attention and providing syntax information for the input. In the Adversary Period Mask experiment, Attention Mask keeps the former, replacing periods with commas keeps the latter, and Token Mask keeps neither. As shown in Table 1, result of Attention Mask setting significantly outperforms the others, indicating that the responsibility to store model's redundant attention is vital for BERT. It is reasonable to believe that BERT puts significant responsibility on these tokens because they can appear consistently in almost any input (high document frequency).

4.2 Representation and Interaction behavior

IR community [6] divides neural ranking models into two categories, namely representation-focused architecture and interaction-focused architecture. BERT adopts a hybrid way to enjoy the advantages of both by learning representations with attention across the query and document tokens. According to Qiao et al. [11], BERT is an interaction-focused model. However, we, to some extent, disagree with this conclusion and further investigate **RQ2** in the following perspectives.

4.2.1 Different behavior in different layers. We show the performance of BERT when interactions are removed layer by layer in Figure 4. We observe slight performance improvement in the previous layers. In the middle layers, the performance declines slowly. It drops rapidly in the last few layers. Similar trend can also be observed in the Attribution experiment, as shown in Figure 2. We believe in the beginning interactions are not important because BERT extracts representations for query and document, respectively. Then it captures interaction signals to learn more context-specific representations. In the last few layers, BERT relies heavily on the interactions between the high-level representations of query and document to predict their relevance.

4.2.2 How does BERT extract relevance signal? We believe that BERT predicts relevance via modeling interactions from document to query for the following reasons. In Figure 2, relative to the empty document baseline, query's attribution rises rapidly, which means that much document information is aggregated to the query token representations. According to the Probing experiment (Figure 3), query token representations contain strong relevance signal, even stronger than "[CLS]" token's representations. In the Mask experiment (Figure 4), removing the interactions from document to query results in significant performance loss.

4.2.3 Improving efficiency. The attention from query tokens to document tokens is little according to Figure 1, and removing it causes slight performance loss, as Figure 4 shows. It demonstrates that BERT extracts query-independent representations for document. Thus, the representations of document tokens can be pre-calculated offline to improve efficiency.

5 CONCLUSION

We investigate a BERT-based ranking model. We find that BERT dumps redundant attention weights on "[CLS]", "[SEP]" and periods due to their high document frequency. We show how BERT predicts relevance via modeling interactions between queries and documents. According to our findings, researchers should be careful with tokens with high document frequency, which may be assigned undeserved responsibilities. We also highlight the possibility of transforming BERT to a more efficient representation-focused model.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).

[2] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *ArXiv* abs/1906.04341 (2019).

[3] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR'19*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[5] Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations. In *IJCNLP 2019*.

[6] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *Information Processing & Management* (2019), 102067.

[7] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *NAACL-HLT*.

[8] Olga V. Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *EMNLP/IJCNLP*.

[9] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv* abs/1611.09268 (2016).

[10] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *ArXiv* abs/1901.04085 (2019).

[11] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *ArXiv* abs/1904.07531 (2019).

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *ICML*.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*.

[14] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *SIGIR '17*.