MaxSimE: Explaining Transformer-based Semantic Similarity via Contextualized Best Matching Token Pairs

Eduardo Brito
Henri Iser
eduardo.alfredo.brito.chacon@iais.fraunhofer.de
henri.iser@iais.fraunhofer.de
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Sankt Augustin, Germany
Lamarr Institute for Machine Learning and Artificial Intelligence
Sankt Augustin, Germany

ABSTRACT

Current semantic search approaches rely on black-box language models, such as BERT, which limit their interpretability and transparency. In this work, we propose MaxSimE, an explanation method for language models applied to measure semantic similarity. Our approach is inspired by the explainable-by-design ColBERT architecture and generates explanations by matching contextualized query tokens to the most similar tokens from the retrieved document according to the cosine similarity of their embeddings. Unlike existing post-hoc explanation methods, which may lack fidelity to the model and thus fail to provide trustworthy explanations in critical settings, we demonstrate that MaxSimE can generate faithful explanations under certain conditions and how it improves the interpretability of semantic search results on ranked documents from the LoTTe benchmark, showing its potential for trustworthy information retrieval.

CCS CONCEPTS

• **Information systems** → **Similarity measures**; **Query representation**; **Document representation**; • **Computing methodologies** → Neural networks.

KEYWORDS

explainable search, semantic similarity, ad-hoc explanations, neural models, trustworthy information retrieval

ACM Reference Format:
Eduardo Brito and Henri Iser. 2023. MaxSimE: Explaining Transformer-based Semantic Similarity via Contextualized Best Matching Token Pairs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592017

1 INTRODUCTION

Modern ranking systems often depend on pre-trained language models to compute representations for queries and documents [12]. Because of the black-box nature of the large deep neural networks they mostly rely on, these models are not suitable when the user requires some explanation to trust the system or to correct it when its output is erroneous [1]. Among the recent Transformer-based [20] approaches, ColBERT [8] introduces a late interaction mechanism to a pre-trained BERT model [5]. This additional layer is used to calculate a similarity score between a query and a document by matching each token vector representation from the query to the closest document token representations, summing them all into a global similarity score. This sum of similarity scores over query terms is similar to more standard ranking methods such as BM25 [16] and we can exploit it to generate explanations about the similarity score. Under the hood, this so-called *MaxSim* operation matches each query token to the most semantically similar document token within their respective contexts. Since we can compute the cosine similarity between any two token representations, we can show the matched tokens by decreasing order of similarity i.e., by decreasing contribution to the global similarity score, so that we can visualize why a retrieved document is (not) similar to the input query. Since the BERT tokens are mostly (sub)words, the matched token pairs can be interpretable terms that are found to be similar.

In this work, we provide examples of where these matches seem informative and discuss the limitations of their interpretability. Additionally, we extend this approach to more 'standard' BERT-based models and compare the resulting explanations to those obtained from ColBERTv2. Our contribution on MaxSim-based Explanations (MaxSimE)[1] is twofold:

(1) We propose an explainability method for Transformer-based semantic similarity, whose fidelity is maximal when applied to models fine-tuned via late interaction such as ColBERTv2 [17]. Visualizing the contextualized best matching tokens can help to confirm a highly ranked document or to hint at some model failure e.g., paired tokens wrongly contributing to a high similarity score.
(2) We intrinsically measure the correctness of MaxSimE taking ColBERTv2 as a proxy for the ground truth to discuss the settings where our explanations are most informative while considering their limitations as well.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592017

[1] Source code available on https://github.com/fraunhofer-iais/MaxSimE

2 RELATED WORK

From the wide spectrum of available explanation methods [1, 3, 11], feature attribution aims to identify the important features or terms that contribute to a particular result. Among them, Local Interpretable Model-agnostic Explanations (LIME) [15] is a popular method that has been adapted for information retrieval tasks [18, 21]. More recent approaches focus on generating explanations that consider not only individual retrieved documents but also the context of the entire search result list to provide more coherent and diverse explanations [23]. While all these approaches provide post-hoc explanations, whose fidelity to the ranker cannot be guaranteed, we focus instead on an explainable by architecture approach. Formal et al. [6] report how BERT-based representations implicitly capture term importance and how the ColBERT fine-tuning approach amplifies this effect, improving the retrieval results. Our approach explicitly exploits this fact to generate explanations highlighting the matched terms and their contribution to the similarity score. Some frameworks focus on inspecting ranking models by evaluating on diagnostic datasets to detect global properties of the tested ranking models [4, 10, 13]. They progress towards a better understanding of why contextualized word embeddings outperform traditional term-based IR methods. Our approach does not aim to analyze model behavior as a whole like them but rather explain a similarity score i.e., to provide local explanations. Calculating semantic similarity based on token embeddings is not a new idea and it has been explored to rank documents [22]. However, we do not aim to build a ranking model from the computed similarity scores but to explain existing models instead.

3 MAXSIME

MaxSimE is a method to generate local explanations for document retrieval systems using language models from which the semantic similarity between two tokens can be measured by the cosine similarity between their vector representations. Its purpose is to provide insights into why a document was retrieved given a query by highlighting the tokens in both the query and the document that contribute the most to their similarity score. We adopt the notation from Santhanam et al. [17] and define a similarity function $S_{q,d}$ between a query q of N tokens and a document d of M tokens as the summation of query-side MaxSim operations, namely, the maximum cosine similarity between each query token embedding and all document token embeddings (implemented as dot-products assuming normalized embeddings):

$$S_{q,d} := \sum_{i=1}^{N} \max_{j=1}^{M} Q_i \cdot D_{d_j}^T, \tag{1}$$

where Q is a matrix of N vectors encoding q and D a matrix of M vectors encoding d, being each vector an embedding of a token.

We match each query token to the most similar document token (given a context) according to the MaxSim operation, as displayed in Figure 1. Formally, given a query embedding q_i, our matching function f_{match} returns the document token embedding d_j with the highest dot product to q_i:

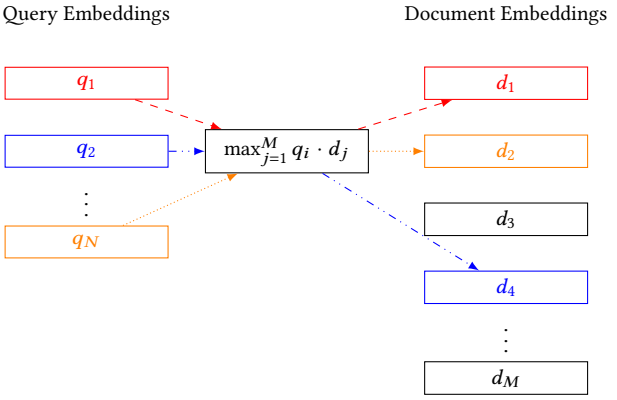

Figure 1: Visualization of the MaxSim operation. Each embedding represents a token created by the BERT tokenizer. Given a query q and a document d, for a query embedding q_i, MaxSim selects the closest document embedding d_j. When the represented query token is an interpretable term, this is equivalent to finding the most semantically similar term appearing in d, represented by the document embedding d_j.

$$f_{match}(q_i) := \arg\max_{d_j} q_i \cdot d_j, \tag{2}$$
$$i \in [\![1..N]\!], j \in [\![1..M]\!]$$

Applying our matching function to all embeddings from a query results in a list of token pairs with the highest similarity according to the cosine similarity of their respective embeddings. These token pairs with their respective similarity scores (computed from their dot product) construct an explanation about "why" document d_j was retrieved given q_i as a query.

4 EXPERIMENTS

4.1 Data

Our experiments are performed on the LoTTE benchmark, a collection of questions and answers sourced from StackExchange. The benchmark covers a wide range of topics, including writing, recreation, science, technology, and lifestyle [17]. To pair documents, we use ColBERTv2 to rank the documents, and we select the top-1-ranked document for each question.

4.2 Fully Faithful Explanations from ColBERT-based Models

We apply our approach first to a ColBERTv2 model to generate explanations. The first observed explanations seem to be informative from a qualitative point of view, as seen in the example from Table 1. The fidelity of the explanations is maximal because ColBERTv2 scoring is directly reliant on the sum of query side MaxSim scores, and the similarity function has been optimized through fine-tuning, thereby giving more significance to the best matching token pairs. In addition, these explanations come at no cost, since the MaxSim scores for each query token are already computed in the retrieval process. Considering that ColBERTv2 approaches state-of-the-art

Table 1: Matched tokens from the query "Why do kittens love packets?" and first ranked document by the pretrained ColBERTv2 model. MaxSim was performed on ColBERTv2 and S-BERT$_{base}$, sorted by descending similarity score.

Query	ColBERTv2		S-BERT$_{base}$	
Token	Token	Score	Token	Score
why	**because**	0.874	**because**	0.911
kitten	[D]	0.809	cats	0.891
##s	**they**	0.756	**they**	0.874
[CLS]	[CLS]	0.728	[CLS]	0.843
do	which	0.722	to	0.848
love	**love**	0.694	**love**	0.912
packets	boxes	0.485	dart	0.787
?	boxes	0.466	means	0.843

level according to most of the metrics from the BEIR benchmark for dense retrieval [19], we assume these explanations to be our "gold standard" for further experiments.

4.3 Explanations from Other BERT-based Models

We generate explanations with our approach from other BERT-based models that were not fine-tuned with a late interaction mechanism like ColBERT. We aim to confirm if these explanations are trustworthy and we thus compare the resulting explanations with those extracted from ColBERTv2 as in Section 4.2, assuming the latter as the reference. As shown in the example from Table 1, the matched tokens partially coincide with those obtained from the ColBERTv2 model although the contribution of the token pairs to the similarity score differs to a greater extent. Performance-wise, generating explanations for non-ColBERT architectures involves $N \cdot M$ cosine distance computations (see Equation 1).

4.4 Evaluation

Despite the absence of ground truth and user feedback, we aim to evaluate the correctness of our explanations extracted from several BERT-based models by comparing them with the ColBERTv2 explanations we generated in Section 4.2, which we take as a proxy for a "gold standard". Let T be the number of correctly retrieved document tokens, P the number of retrieved query/token pairs according to the gold standard, and N the number of query tokens. For each query document, we evaluate the following metrics on the Top-1 document retrieved by ColBERTv2:

(1) Token precision: $\frac{T}{N}$
(2) Matching accuracy: $\frac{P}{N}$
(3) Spearman's rank correlation of the matching token scores with the gold standard.

Notice that the matching accuracy is a stricter variant of the token precision since the token precision just measures how many of the expected document tokens were retrieved (independent from the query tokens they were matched to), whereas the matching accuracy only counts the matches where the tokens are correct both from the query and the document side. The Spearman's rank correlation is intended to capture the similarity in terms of ranking query tokens.

We compare the explanations from two variants of model architectures: Cross-Encoders, which use a regression head to compute the similarity of two input texts directly; and Bi-Encoders, which produce one embedding per document either by Mean/Max Pooling token embeddings or by selecting the `[CLS]` token embedding so that the similarity of two texts is measured by the cosine similarity of the respective embeddings. Bi-Encoders therefore also use a late-interaction mechanism for similarity estimation whereas Cross-Encoders are fully attention-based. We analyze the effect this has on the generated explanations. For Cross-Encoders we choose the MSMARCO pretrained TinyBERT and MiniLM-L6 model, provided by the `sentence-transformers` library [14]. For Bi-Encoders we compare the S-BERT$_{base}$ model with its distilled variant DistilBERT and with the MiniLM-L6 model.

4.5 Results

We first analyze the explanations generated by both ColBERTv2 and the Bi-Encoder S-BERT$_{base}$. Table 1 shows token matching pairs of both models. Qualitatively, we can observe that both explanations match similar document tokens to the query. Partially these matches coincide between the two models. From Figure 2 we can observe a noticeable difference in absolute score values, especially in the ranking of the matching token pairs. In comparison, S-BERT$_{base}$ yields higher scores for query tokens ranked lower by ColBERTv2. Furthermore, the score values produced by ColBERTv2 exhibit a greater degree of variance, especially for these lower-ranked tokens. We assume that this is due to the fine-tuning of ColBERTv2 token representations with the MaxSim late-interaction mechanism, which forces the model to also perform a fine-grained ranking on the token level.

When evaluating the correctness of our explanations on non-ColBERT models, we observe that token precision is generally high across most models (as displayed in Table 2). All metrics have high variance, which suggests that the quality of the explanations is highly dependent on the query sentence. Especially the matching accuracy and ranking of the tokens are inconsistent throughout the dataset. For the smaller model MiniLM-L6, we see that the Bi-Encoder variant provides explanations closer to our gold standard. This could be explained by the fact that the late-interaction mechanism used in sentence transformers (especially with mean pooling) is more similar to the MaxSim operation than the regression head in Cross-Encoders.

4.6 Discussion

We observe that, although the non-ColBERT models were not trained using the MaxSim operations, the generated explanations largely align with those of ColBERTv2, as demonstrated by the example in Table 1. The similarity between the explanations suggests that they similarly capture term importance, in line with previous white box analysis on ColBERT [6]. Considering that the ranking performance does differ, we guess that the different similarity value distributions assigned to the matches have a noticeable impact on the global similarity score and thus on the ranked documents. The distributions in Figure 2 illustrate how ColBERTv2 weights with

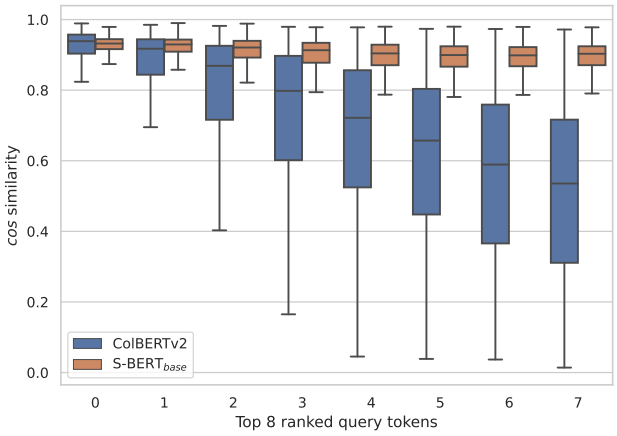

Figure 2: Cosine similarity distribution of the top 8 ranked query tokens for each query from the LoTTE dataset.

Table 2: Similarity of explanations from BERT-based models to our ColBERTv2 gold standard measured by token precision (TP), match accuracy (MA), and Spearman's rank correlation (SR).

Model	TP	MA	SR
	Bi-Encoders		
S-BERT$_{base}$	0.730 ± 0.153	**0.471** ± 0.213	0.427 ± 0.380
DistilBERT	0.740 ± 0.163	0.444 ± 0.212	0.349 ± 0.386
MiniLM-L6	0.664 ± 0.149	0.411 ± 0.200	**0.473** ± 0.376
	Cross-Encoders		
TinyBERT	**0.749** ± 0.158	0.446 ± 0.204	0.391 ± 0.343
MiniLM-L6	0.387 ± 0.233	0.307 ± 0.192	0.270 ± 0.284

significantly higher similarity scores for the most semantic relevant terms than the rest of the tokens, whereas the similarity score difference among embeddings coming from BERT$_{base}$ is clearly less differentiated. Despite this, the high token precision (displayed in Table 2) implies that non-ColBERT models frequently match the same tokens as ColBERT.

Although we demonstrated how we can generate meaningful explanations for both ColBERT and other BERT-based models using the MaxSim operation, we acknowledge two main limitations of our approach: the limited faithfulness to the model for non-ColBERT architectures and the limited interpretability of some explanations because of the contribution of the `[MASK]` tokens to the similarity score.

First, our explanations from non-late-interaction-based models i.e., Bi- and Cross-encoders [14], cannot guarantee faithfulness to their respective ranking models because their computed similarity usually comes from either a regression head or from the cosine similarity of `[CLS]` or mean pooled embeddings. Although Cross-Encoder models may achieve better evaluation scores, their computational cost is much higher, becoming impractical for most setups. Hence, we favor late interaction models for ranking not only because of their efficiency on ranking tasks but also because we can extract fully faithful explanations from the underlying language model.

Second, Khattab and Zaharia [8] use `[MASK]` tokens within the ColBERTv2 model for query expansion. These non-interpretable tokens are also included in the late-interaction scoring mechanism, leading to best-matching token pairs that cannot be explained in a meaningful way. Depending on the length of the query, these `[MASK]` tokens make up for up to 62% of the final score of the retrieved document. Nonetheless, Lassance et al. [9] show that these special tokens can be safely removed without affecting model performance in a significant way.

Finally, we could only evaluate the correctness of the explanations extracted from the different models by comparing them to our ColBERTv2 gold standard, which we consider confirmed when they correlate but we cannot discard otherwise. Other explainability aspects such as plausibility [7] are yet to be assessed as well. Despite the limitations, we find our first exploratory results promising and we hope to motivate more work towards trustworthy information retrieval.

5 CONCLUSION AND FUTURE WORK

We leveraged the MaxSim operation from the ColBERT approach to generate explanations for the documents retrieved by the ranking system, based on the most relevant document tokens that match those of the query. We also demonstrated that our method can be applied to other BERT-based models, although we cannot guarantee its fidelity for those models. The correlation between the explanations generated by the different models confirms that our proposed method can provide insights into the underlying model, and can be used as a proxy to evaluate explanation correctness. Our presented method enables "explanations for free" i.e., without needing to learn any explanation model, from similarity functions constructed upon BERT-based language models. Our proposed approach may have applications beyond information retrieval e.g., text classification use cases where unfaithful explanations from black-box models are not acceptable and where a similarity-based classifier can be used without a dramatic performance loss compared to the best-performing black-box deep learning model [2]; or even less related areas where Transformer-based models can deal with a concept of semantic similarity such as computer vision [24]. In future work, we aim to systematically compare the ranking performance of different BERT-based models with our evaluation results, including additional evaluation criteria and benchmark datasets where ColBERTv2 was not fine-tuned. From a more applied perspective, we also plan to apply our approach to domain-specific settings e.g., information retrieval on legal texts to support lawyers finding previous similar legal cases when facing a new one, which is an opportunity to assess the plausibility of our explanations.

ACKNOWLEDGMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence. We thank Christian Bauckhage, Katharina Beckh, and Stefan Rüping for their feedback prior to our paper submission.

REFERENCES

[1] Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. 2023. Harnessing Prior Knowledge for Explainable Machine Learning: An Overview. In *First IEEE Conference on Secure and Trustworthy Machine Learning*. https://openreview.net/forum?id=1KE7TlU4bOt

[2] Eduardo Brito, Vishwani Gupta, Eric Hahn, and Sven Giesselbach. 2022. Assessing the Performance Gain on Retail Article Categorization at the Expense of Explainability and Resource Efficiency. In *KI 2022: Advances in Artificial Intelligence*, Ralph Bergmann, Lukas Malburg, Stephanie C. Rodermund, and Ingo J. Timm (Eds.). Springer International Publishing, Cham, 45–52.

[3] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[4] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 605–618.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A White Box Analysis of ColBERT. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 257–263.

[7] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? https://doi.org/10.48550/ARXIV.2004.03685

[8] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[9] Carlos Lassance, Maroua Maachou, Joohee Park, and Stéphane Clinchant. 2021. A Study on Token Pruning for ColBERT. https://doi.org/10.48550/arXiv.2112.06540 arXiv:2112.06540 [cs].

[10] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the Behavior of Neural IR Models. *Transactions of the Association for Computational Linguistics* 10 (2022), 224–239. https://doi.org/10.1162/tacl_a_00457

[11] Christoph Molnar. 2020. *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

[12] Gerhard Paaß and Sven Giesselbach. 2023. *Foundation Models for Natural Language Processing–Pre-trained Language Models Integrating Media*. Springer Nature 2023. https://arxiv.org/abs/2302.08575

[13] David Rau and Jaap Kamps. 2022. How Different Are Pre-Trained Transformers For Text Ranking?. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 207–214. https://doi.org/10.1007/978-3-030-99739-7_24

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. https://doi.org/10.48550/ARXIV.1606.05386

[16] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[17] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[18] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search Using Local Model Agnostic Interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, Melbourne VIC Australia, 770–773. https://doi.org/10.1145/3289600.3290620

[19] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=wCu6T5xFjeJ

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[21] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1281–1284. https://doi.org/10.1145/3331184.3331377

[22] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-Hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/3077136.3080809

[23] Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards Explainable Search Results: A Listwise Explanation Generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 669–680. https://doi.org/10.1145/3477495.3532067

[24] Chao Zhang, Stephan Liwicki, and Roberto Cipolla. 2022. Beyond the CLS Token: Image Reranking using Pretrained Vision Transformers. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0080.pdf