Interpretability Analysis of Domain Adapted Dense Retrievers

Göksenin Yüksel
goksenin.yuksel@student.uva.nl
University of Amsterdam
Amsterdam, Netherlands

Jaap Kamps
kamps@uva.nl
University of Amsterdam
Amsterdam, Netherlands

ABSTRACT

Dense retrievers have demonstrated significant potential for neural information retrieval; however, they exhibit a lack of robustness to domain shifts, thereby limiting their efficacy in zero-shot settings across diverse domains. Previous research has investigated unsupervised domain adaptation techniques to adapt dense retrievers to target domains. However, these studies have not focused on explainability analysis to understand how such adaptations alter the model's behavior. In this paper, we propose utilizing the integrated gradients framework to develop an interpretability method that provides both instance-based and ranking-based explanations for dense retrievers. To generate these explanations, we introduce a novel baseline that reveals both query and document attributions. This method is used to analyze the effects of domain adaptation on input attributions for query and document tokens across two datasets: the financial question answering dataset (FIQA) and the biomedical information retrieval dataset (TREC-COVID). Our visualizations reveal that domain-adapted models focus more on in-domain terminology compared to non-adapted models, exemplified by terms such as "hedge," "gold," "corona," and "disease." This research addresses how unsupervised domain adaptation techniques influence the behavior of dense retrievers when adapted to new domains. Additionally, we demonstrate that integrated gradients are a viable choice for explaining and analyzing the internal mechanisms of these opaque neural models.

1 INTRODUCTION

One of the main recent achievements in information retrieval (IR) is the development of neural dense retrieval methods. These methods are extremely fast and do not entail the memory and computational overhead associated with widely used other neural methods such as cross-encoder and late interaction models [21]. However, dense retrieval models face the challenging task of independently mapping inputs to a meaningful vector space, making them highly sensitive to domain shifts [15]. This non-robustness to domain shifts impedes their application in zero-shot settings [16, 18], posing challenges

Table 1: Retrieval effectiveness of domain adaptation.

Collection	NDCG@10			
	Baseline	GLP	Absolute	Percentage
TREC-COVID	0.6510	0.7160	+0.0650	+9.98%
FIQA	0.2670	0.3680	+0.1010	+37.83%

in real-world applications where access to extensive and domain-specific training data is limited [9]. Methods known as "domain adaptation" have been developed to address this issue.

To tackle the domain shift problem with dense retrievers, previous work has fine-tuned these models on target datasets using unsupervised and supervised learning objectives [10, 15, 17]. Supervised methods utilize labeled data to further fine-tune these models in novel domains, which is not possible for every task or domain of IR research due to the costs and difficulty of obtaining human-annotated labels [19]. In contrast, unsupervised methods assume only the availability of the target corpus [10, 17, 19], employing pre-training objectives [4, 6, 7] or pseudo-labeling [10, 17] to fine-tune the pre-trained models in a new domain without requiring labeled data. Notably, Wang et al. [17] found that pre-training objectives alone do not enhance the out-of-domain performance of the adapted model.

Prior work on domain adaption focuses on retrieval effectiveness (e.g. Table 1 discussed below), but has not addressed the interpretability of domain-adapted models. Moreover, no model-introspective analysis was conducted to understand changes in the models' inner workings before and after domain adaptation. Given the inherent opacity and lack of transparency of these neural models, interpretability analysis is crucial to understand the effect of domain adaptation on such dense retrievers. This paper conducts initial interpretability analysis experiments, and proposes to use of the Integrated Gradients (IG) [14] framework to develop an interpretability method for dense retrievers, providing both instance-based and ranking-based explanations. Subsequently, we apply this method to domain-adapted and non-domain-adapted dense retrievers to assess the differences in their behavior using input-based attributions. We set out to answer two research questions. *(i) How can Integrated Gradients be utilized in the dense retriever setting? (ii) How does domain adaptation influence the input attributions of the models?*

In this paper, we first demonstrate that IG is a viable approach for interpreting dense retrievers. We then utilize the proposed interpretability method to analyze input attribution differences in the FIQA and TREC-COVID datasets, which represent two distinct domains in IR: financial question answering and biomedical information retrieval. Our visualizations reveal that the domain-adapted model concentrates more on in-domain terminology, and title which the unadapted model tends to overlook.

2 RELATED RESEARCH

2.1 Unsupervised Domain Adaptation

Query Generation (QG): QG methods construct synthetic training data by using documents from the target domain to generate corresponding (pseudo) queries, aiming to augment the training data with queries that fit the target domain. QGen [10] trains an auto-encoder in the source domain to generate synthetic questions from a target domain document. They use binary-level relevancy labels to train the networks on generated query-document pairs. Similarly, GPL [17] generates synthetic queries with a pre-trained T5 model but uses cross-encoders to label the relevancy of generated query-document pairs. This method extends and over performs the QGen method by replacing binary relevance with continuous labels ranging from -inf to inf.

Knowledge Distillation (KD): KD is a commonly used strategy in the dense retriever setting, which utilizes a powerful model as the teacher model to improve the capabilities of the student models [5]. It has been found that such a technique can improve out-of-domain performance as well. GPL [17] and AugSBERT [15] use cross-encoders to annotate unlabeled synthetic query-doc pairs. Later, this knowledge is distilled into the dense retriever by training the model on generated labels. Different from the above methods, SPAR [3] proposes to distill knowledge from BM25 to the dense retriever model to integrate sparse retrieval.

2.2 Interpretability methods for IR

Interpretability of ranking models focuses on building models that can either be analyzed for interpretability in a post-hoc fashion or are interpretable by design.
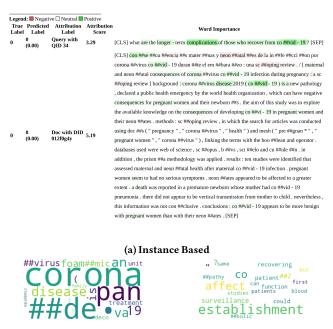
Post-hoc interpretability methods explain the decisions of already trained machine learning models. These approaches are either model-agnostic, where the interpretability approach has no access to the trained model parameters [8, 12], or model-introspective, with full access to the parameters of the underlying model [13, 14]. There exists a lot of work on model-introspective analysis of neural models for different tasks. These analyses use different methods like probing tasks, attention weights, or state activations [1]. A recent dominant class of model-introspective explanation outputs feature attributions. Most of these approaches utilize gradient-based attribution methods [1].

To our knowledge, only Zhan et al. [20] used IG to obtain feature attributions for a BERT-based cross-encoder model. In their experiments, they used an empty query baseline and an empty document baseline, which are padding tokens.

3 METHODOLOGY

Dense Retriever In this paper, we use pre-indexing of the documents. The pre-indexed document embeddings are used to retrieve top-K passages using dot product similarity between document and query embeddings. The query and document embeddings are obtained using a DistilBERT model, which has already been fine-tuned on MSMARCO to retrieve relevant documents to a query [2].

We use the dense retriever models that are tuned to work with the dot product as their similarity measure. We utilize open sourced GPL models and use SentenceTransformer [11] framework. For all the GPL models, maximum sequence length is set to 350. We use mean pooling over output token embeddings, disregarding the special tokens such as [CLS] and [SEP]. For TREC-COVID, and FIQA

(a) Instance Based

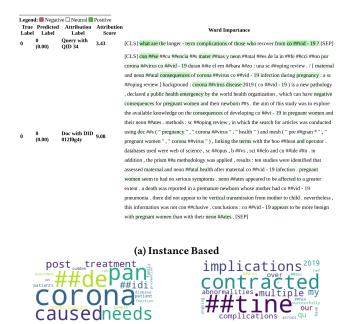**(b) Ranking Based Positive Attribution**

(c) Ranking Based Negative Attribution

Figure 1: Attribution analysis for random query and relevant document for TREC-COVID. The model used was GPL/msmarco-distilbert-margin-mse. The word sizes are determined by the summed attribution over top ranked 25 documents

dataset we evaluate provided the models, and asses the performance improvement on the corresponding test sets.

Baseline As stated by Sundararajan et al. [14], a good baseline should give a score of zero and should convey an empty signal. Inspired by the recent work in cross-encoder explanations, we use the [PAD] token to create our baseline. However, we deviate from them by introducing a new method to calculate the query and document attributions. To calculate the query token attributions, we replace the query tokens with the [PAD] tokens and leave the document tokens untouched. Then, we replace the document tokens with [PAD] tokens and leave the query tokens untouched. We run the Integrated Gradient analysis for both the query and the document baselines. This method calculates the input attributions for both the query and the document tokens.

Ranking Analysis The aforementioned method only works for instance-based explanations. However, in the information retrieval setting, we are also interested in the explainability of the document ranking. For this task, we select the top 25 documents retrieved by the model for a specific query. We aggregate the token attributions over the top retrieved documents by summing them up to generate the overall attributions of a token. This way, we get the most important tokens for the ranking. The tokens that appear often in the top rankings and contribute positively get a higher attribution score for the ranking overall. We use word cloud visualization, the word sizes are determined by the summed attribution over 25 documents.

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	Query with QID 34	3.43	[CLS] what are the longer - term complications of those who recover from co ##vid - 19 ? [SEP]
0	0 (0.00)	Doc with DID 012f0g4y	9.08	[CLS] con ##se ##cu ##encia ##s mater ##nas y neon ##atal ##es de la in ##fe ##cci ##on por corona ##virus co ##vid - 19 duran ##te el em ##bara ##zo : una sc ##oping review . / [maternal and neon ##atal consequences of corona ##virus co ##vid - 19 infection during pregnancy : a sc ##oping review] background : corona ##virus disease 2019 (co ##vid - 19) is a new pathology , declared a public health emergency by the world health organization , which can have negative consequences for pregnant women and their newborn ##s . the aim of this study was to explore the available knowledge on the consequences of developing co ##vi - 19 in pregnant women and their neon ##ates . methods : sc ##oping review , in which the search for articles was conducted using dec ##s (" pregnancy " , " corona ##virus " , " health ") and mesh (" pre ##gnan * " , " pregnant women " , " corona ##virus ") , linking the terms with the boo ##lean and operator . databases used were web of science , sc ##opus , b ##vs , sci ##elo and cu ##ide ##n . in addition , the prism ##a methodology was applied . results : ten studies were identified that assessed maternal and neon ##atal health after maternal co ##vid - 19 infection . pregnant women seem to had no serious symptoms . neon ##ates appeared to be affected to a greater extent . a death was reported in a premature newborn whose mother had co ##vid - 19 pneumonia . there did not appear to be vertical transmission from mother to child . nevertheless , this information was not con ##clusive . conclusions : co ##vid - 19 appears to be more benign with pregnant women than with their neon ##ates . [SEP]

(a) Instance Based

(b) Ranking Based Positive Attribution   **(c) Ranking Negative Attribution**

Figure 2: Attribution analysis for random query and relevant document for TREC-COVID. The model used was GPL/trec-covid-msmarco-distilbert-gpl. The word sizes are determined by the summed attribution over top ranked 25 documents

Title Attribution Analysis For TREC-COVID data, title is appended in the beginning of the document, and then the domain adaptation is performed. FIQA does not have a title attribute. Also for MSMARCO models, this is not the case as they do not include a title.

Hence for TREC-COVID, we can test for the difference in total attribution for the title to understand the domain adaptation affect. For each query, we randomly select a relevant document, and compute the document token attributions for both base and domain adapted model. Later, we sum the attribution for title tokens.
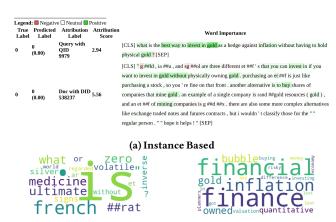
4 RESULTS

Our experiments, shown in Table 1 before, confirm the effectiveness of domain adaption observed in earlier research, on TREC-COVID NDCG@10 improves from 65.1 to 71.6 (+6.5 abs., +10%) and on FIQA from 26.7 to 36.8 (+10.1 abs., +38%). With the proposed method, we now analyze if we can interpret how the domain adaptation is affecting the model.

4.1 Input Attribution

Figures 1 and 3 show the attribution analysis for the baseline model.

Figures 1a and 3a display the positive and negative attributions for both query and document pairs in models trained on MSMARCO using DistilBERT. Positive attributions enhance the similarity between the query and the document, while negative attributions lowers it.

In Figure 1a, the query tokens ["complications," "co," "##vid"] contribute positively. Additionally, the document tokens ["co," "##vid," "disease"] also contribute positively, whereas ["neon"] contributes negatively.

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	Query with QID 9979	2.94	[CLS] what is the best way to invest in gold as a hedge against inflation without having to hold physical gold ? [SEP]
0	0 (0.00)	Doc with DID 538237	5.56	[CLS] " g ##ld , ia ##u , and sg ##ol are three different et ##f ' s that you can invest in if you want to invest in gold without physically owning gold . purchasing an et ##f is just like purchasing a stock , so you ' re fine on that front . another alternative is to buy shares of companies that mine gold . an example of a single company is rand ##gold resources (gold) , and an et ##f of mining companies is g ##d ##x . there are also some more complex alternatives like exchange traded notes and futures contracts , but i wouldn ' t classify those for the " " regular person . " " hope it helps ! " [SEP]

(a) Instance Based

(b) Ranking Based Positive Attribution   **(c) Ranking Based Negative Attribution**

Figure 3: Attribution analysis for random query and relevant document for FIQA. The model used was GPL/msmarco-distilbert-margin-mse. The word sizes are determined by the summed attribution over top ranked 25 documents

In Figure 3a, the query tokens ["invest," "gold," "best"] contribute positively, while ["against," "?"] contribute negatively. Furthermore, the document tokens ["gold," "without," "buy"] contribute positively, whereas ["g," "sg"] contribute negatively.

For both queries, the DistilBERT model effectively matches query terms with document terms. Tokens that appear in both the query and the document receive high positive attribution scores. As expected, we observe no attributions for the [CLS] and [SEP] tokens.

Ranking Based Attribution Figures 1b, 1c, 3b, and 3c depict the word clouds of positive and negative attributions revealed by the proposed method in a ranking scenario.
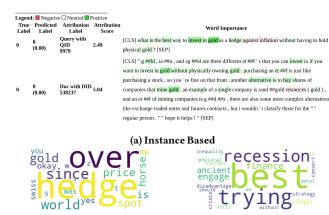
Despite the query not explicitly mentioning ["corona," "disease"], Figure 1b illustrates that these are important words identified by the model. Another notable finding is the appearance of "complications" in the negative attribution word cloud shown in Figure 1c, which is a term present in the query text.

For the FIQA dataset, Figures 3b and 3c show that the model assigns positive attributions to document tokens ["what," "is," "french"] and negative attributions to ["financial," "finance"]. Additionally, "gold" appears among the negatively contributing terms.

4.2 Domain Adaptation

Figures 2 and 4 show the attribution analysis after domain adaptation.

In Figure 2a, the query tokens ["what," "co," "##vid"] contribute positively. Additionally, the document tokens ["pregnancy," "consequences," "corona"] contribute positively. The query contribution of "complications" has decreased compared to the non-domain-adapted model, while the contribution of "what" has increased. Furthermore, the domain-adapted model assigns more importance to the first sentence, which is the title of the paper. The attribution for document tokens ["corona," "consequences," "pregnant"] has also increased.

In Figure 4a, the query tokens ["invest," "gold," "hedge"] contribute positively, whereas ["against," "inflation"] contribute negatively. Additionally, the document tokens ["gold," "invest," "buy"] contribute positively, while ["resources," "is"] contribute negatively. The query

Legend: ■ Negative □ Neutral ■ Positive				
True Label	**Predicted Label**	**Attribution Label**	**Attribution Score**	**Word Importance**
0	0 (0.00)	Query with QID 9979	2.49	[CLS] what is the best way to invest in gold as a hedge against inflation without having to hold physical gold ? [SEP]
0	0 (0.00)	Doc with DID 538237	5.04	[CLS] " g ##ld , ia ##u , and sg ##ol are three different et ##f ' s that you can invest in if you want to invest in gold without physically owning gold . purchasing an et ##f is just like purchasing a stock , so you ' re fine on that front . another alternative is to buy shares of companies that mine gold . an example of a single company is rand ##gold resources (gold) , and an et ##f of mining companies is g ##d ##x . there are also some more complex alternatives like exchange traded notes and futures contracts , but i wouldn ' t classify those for the " " regular person . " " hope it helps ! " [SEP]

(a) Instance Based

(b) Ranking Based Positive Attribution

(c) Ranking Based Negative Attribution

Figure 4: Attribution analysis for random query and relevant document for FIQA. The model used was GPL/fiqa-msmarco-distilbert-gpl. The word sizes are determined by the summed attribution over top ranked 25 documents

contribution of "inflation" has decreased after domain adaptation, whereas "best" has increased. Moreover, the attribution for document tokens ["gold," "alternative," "g"] has increased, whereas ["resources," "without"] has decreased.

Both models display positive attributions for the input tokens in the sentence, "you can invest in if you want to invest in gold without physically owning gold."

Ranking For TREC-COVID, Figures 1b and 2b illustrate that both the non-domain-adapted model and the domain-adapted model place emphasis on "corona" and "pan." A notable finding is the increase in attribution to "treatment" and the appearance of "post" in the domain-adapted model.

Regarding negative attributions, Figures 1c and 2c show that the non-domain-adapted model focuses more on "establishment," "affect," and other non-related query words. In contrast, the domain-adapted model focuses on "contracted" and "implications." In both cases, they do not assign negative attributions to query-related terms. Additionally, the domain-adapted model identifies "2019" as a negatively contributing token to the ranking.

For FIQA, Figures 3b and 4b reveal that the domain-adapted model places greater emphasis on "hedge," "gold," and "over." Conversely, the non-domain-adapted model places more emphasis on "is" and "french." Furthermore, Figures 3c and 4c show that the non-domain-adapted model assigns high negative attribution to "finance" and "financial," whereas the domain-adapted model assigns high negative attribution to "best" and "trying." In this scenario, the domain-adapted model focuses more on query-related terms such as "hedge" and "gold" compared to the non-domain-adapted model.

Title attributions Figure 5 depicts that domain adapted model puts positive attribution to the title compared to base model. The base model puts significantly negative attribute to the title. This quantitative finding is consistent with the qualitative observations above.
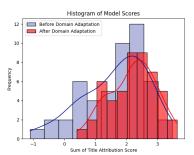

Figure 5: Sum of title attribution scores for TREC-COVID

5 CONCLUSIONS AND DISCUSSION

This paper proposes to use Integrated Gradients for instance-based and ranking-based explanations of dense retrievers. We find that Integrated Gradients is a feasible choice for the interpretability of dense retriever models. Our initial findings show that dense retriever models are capable of both soft matching and term matching as expected. For example, even though the query may not contain the word "volatile," documents containing "volatile" are ranked higher, with a positive attribution score assigned to the term. Furthermore, we observe that negative attributions are also applicable in the dense retriever setting, with certain tokens contributing negatively to the similarity score. As illustrated in Figure 1, these are typically non-relevant tokens for the query.

We find that domain-adapted models tend to place more positive attributions on domain-specific vocabulary, such as "corona" and "hedge," compared to the baseline model. Additionally, the domain-adapted model better captures the dependencies of document terms like "treatment" and "rehabilitation," which are relevant to the query but not explicitly stated in the query terms, compared to the non-domain-adapted model. Moreover, the positive attribution given to the title by the domain-adapted model in the TREC-COVID domain suggests that the model relies more on the document title to generate similarity scores compared to the non-domain-adapted model. This behavior may be influenced by the domain adaptation process in GPL models, where the title is concatenated at the beginning of the document [17]. Thus, the model may learn to focus more on the beginning of the document, recognizing the title as an important element for retrieving research papers related to the query.

A LIMITATIONS

Interpretable and explainable neural ranking models is an important, but also very hard to study problem in IR. Our initial experiments in this paper merely aim to raise interest in, and show the viability of, interpretability analysis of dense retrievers. As there is no standard quantitative way of evaluating attributions produced by IG, we opted for a deep qualitative analysis of a small sample of queries. We plan to considerably expand this in future work. We also plan to compare the IG method to other interpretability methods, such as [8] or [12], in future research.

To gain a comprehensive understanding of the white/black box model, we also want to expand the analysis to the global attributions of models. However, IG is not capable of producing such global explanations. We aim to extend this our research beyond the top ranked

documents, as understanding the attributions for lower-ranked documents is crucial for more comprehensible and reliable IR research. Moreover, it is possible to analyze the most influential document tokens for a query similarity by computing document token attributions using all the documents in the corpus. We plan to conduct such very computationally demanding analysis in future research.

REFERENCES

[1] Avishek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. 2023. Explainable Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3448–3451. https://doi.org/10.1145/3539618.3594249

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[3] Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One?. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 250–262. https://doi.org/10.18653/v1/2022.findings-emnlp.19

[4] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[5] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/3404835.3462891

[6] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6086–6096. https://doi.org/10.18653/v1/P19-1612

[7] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1442–1459. https://doi.org/10.18653/v1/2021.emnlp-main.109

[8] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[9] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 1075–1088. https://doi.org/10.18653/v1/2021.eacl-main.92

[10] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith B. Hall, and Ryan T. McDonald. 2020. Zero-shot Neural Retrieval via Domain-targeted Synthetic Query Generation. *ArXiv* abs/2004.14503 (2020). https://api.semanticscholar.org/CorpusID:216867871

[11] 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*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[12] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, John DeNero, Mark Finlayson, and Sravana Reddy (Eds.). Association for Computational Linguistics, San Diego, California, 97–101. https://doi.org/10.18653/v1/N16-3020

[13] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html

[14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs.LG]

[15] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 296–310. https://doi.org/10.18653/v1/2021.naacl-main.28

[16] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. (4 2021). https://arxiv.org/abs/2104.08663v4

[17] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *arXiv preprint arXiv:2112.07577* (4 2021). https://arxiv.org/abs/2112.07577

[18] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 9 (2019), 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768

[19] Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-Shot Dense Retrieval with Momentum Adversarial Domain Invariant Representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4008–4020. https://doi.org/10.18653/v1/2022.findings-acl.316

[20] 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1941–1944. https://doi.org/10.1145/3397271.3401325

[21] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2023. Dense Text Retrieval based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* (dec 2023). https://doi.org/10.1145/3637870 Just Accepted.