Retrieval-Augmented Generation: Is Dense Passage Retrieval Retrieving?

Benjamin Reichman
Georgia Institute of Technology
bzr@gatech.edu

Larry Heck
Georgia Institute of Technology
larryheck@gatech.edu

Abstract

Dense passage retrieval (DPR) is the first step in the retrieval augmented generation (RAG) paradigm for improving large language models (LLM) performance. DPR fine-tunes pre-trained networks to enhance the alignment of the embeddings between queries and relevant textual data. A deeper understanding of DPR fine-tuning will be required to fundamentally unlock the full potential of this approach. In this work, we explore DPR-trained models mechanistically by using a combination of probing, layer activation analysis, and model editing. Our experiments show that DPR training decentralizes how knowledge is stored in the network, creating multiple access pathways to the same information. We also uncover a limitation in this training style: the internal knowledge of the pre-trained model bounds what the retrieval model can retrieve.

1 Introduction

Large Language Models (LLMs) in the past few years went from being a research topic in natural language processing to being a tool utilized daily by hundreds of millions of people and integrated into a wide variety of businesses. With this meteoric rise, these models have been critiqued for frequently hallucinating, confidently outputting incorrect information (Bang et al., 2023). Such inaccuracies can not only mislead people but also erode trust in LLMs. Trust in these systems to give accurate information is crucial to their ability to help people and continue their adoption.

The retrieval augmented generation (RAG) paradigm is one proposed way to fix this hallucination problem (Lewis et al., 2020). Unlike traditional LLM interactions, where a query directly prompts an output from the model, RAG introduces an intermediary step. Initially, a 'retrieval' model processes the query to gather additional information from a knowledge base, such as Wikipedia or the broader internet. This additional information alongside the original query is fed to the LLM, increasing the accuracy of the answers that the LLM generates.

For RAG to be effective, the underlying retrieval model has to excel at finding accurate and relevant information. Typically, the performance of these models is evaluated based on metrics that consider the top-5, top-20, top-50, and top-100 retrieved passages. However, recent studies indicate that LLMs predominantly use information from the top-1 to top-5 passages, underscoring the importance of not only accuracy in retrieval but also precision in ranking (Liu et al., 2023; Xu et al., 2024). Retrieval model performance improves greatly as one goes from top-1 to top-100, highlighting a central issue in the current RAG pipeline. One solution involves integrating a 'reranking' model, which adjusts the order of retrieved passages to improve the relevance of the top-ranked passages (Nogueira et al., 2019, 2020). However, this approach adds the computational and maintenance cost of an additional model in the pipeline and can also introduce errors. The alternative option is to improve retrieval models so that they can jointly retrieve and rank passages well.

Retrieval methods can be broadly categorized into two types: sparse and dense (Zhao et al., 2023). Sparse methods encode queries and passages into sparse vectors, usually based on terms that appear in said queries and passages (Robertson and Zaragoza, 2009; Sparck Jones, 1972). Dense methods employ language models to encode the semantic information in queries and passages into dense vectors (Karpukhin et al., 2020). The type of dense methods that we will be exploring in this paper share two commonalities: (a) the joint training of two or more encoding models – one for embedding a query and the other for embedding a knowledge base, (b) contrastive training. These commonalities were introduced in the dense passage retrieval method which a lot of subsequent methods are inspired from. In this paper, we analyze the original dense passage retrieval method using the BERT backbone. We analyze DPR from multiple perspectives to understand what is changing in the backbone model during the training process.

2 Results and Conclusion

This paper set out to discover the purpose DPR-style fine-tuning served and discover insights into how DPR-trained BERT operates. We found that the middle segment of the model, where the model is processing a mix of syntactic and semantic features according to Geva et al.'s (2021), impacts performance the most. Through linear probing, alongside experiments where we added and removed knowledge from pre-trained BERT, we determined that BERT does not appear to acquire new information through DPR fine-tuning. Instead, we observed that the efficacy of retrieval hinges on the activation of shared facts/memories between the BERT models used to encode the query and the context passages. This mechanism implies that incorrect retrieval could occur if a query or context passage inadvertently activates irrelevant or incorrect memories. Moreover, the absence of necessary facts or webs of knowledge within the model hampers its ability to retrieve information.

However, the crucial insight came from analyzing the changes in BERT's activations before and after DPR-style training. We found that DPR-style training alters the model's internal representation of facts, transitioning from a highly centralized to a decentralized representation of facts. Pre-trained BERT's representations are very centralized with a select few neurons being activated across a wide array of facts and only a few neurons being strongly activated for each fact, suggesting a limited number of pathways for fact or memory activation. The representations in DPR-trained BERT, on the other hand, are a lot less centralized. DPR-trained BERT engages more neurons, more robustly for each fact, and diminishes the uniform reliance on specific neurons across different facts. This decentralization makes it so that each fact/memory has a lot more pathways to get triggered, which in turn allows for more potential inputs to trigger the same set of memories. Such a shift not only underscores the primary objective of DPR training—to diversify the model's retrieval capabilities across an expanded set of queries and passages—but also delineates a crucial mechanism by which these models improve their retrieval performance.

Our findings suggest several areas of focus for future work: (1) Accelerate knowledge representation decentralization with new unsupervised training methods. Current methods for DPR rely on labeled queries and passage pairs. However, only relying on this labeled data limits how much decentralization can occur. (2) Optimize retrieval methods that operate with uncertainty. More detailed model analysis is required to determine how the model processes a query when it is missing key knowledge for the retrieval. The analysis should reveal methods to more robustly and gracefully degrade with increased levels of uncertainty. (3) Directly map a model's internal knowledge to the set of best documents to retrieve. These approaches should better leverage the model's knowledge as shown in (Tay et al., 2022; Pradeep et al., 2023; Wang et al., 2022; Bevilacqua et al., 2022; Ziems et al., 2023).

In the most fundamental sense, Dense Passage Retrieval (DPR) achieves its namesake function—it retrieves, locating and returning relevant context to the user given a query. Yet, as our evidence suggests, DPR models appear constrained to retrieving information based on the knowledge that preexists within their parameters, either innately or through augmentation. This operational boundary delineates a significant caveat: facts must already be encoded within the model for useful context to be accessible for retrieval. Absent these facts or their associative networks, retrieval seems to falter. Thus, if retrieval is understood as the capacity to recall or recognize knowledge already familiar to the model, then indeed, DPR models fulfill this criterion. However, if we extend our definition of retrieval to also encompass the ability to navigate and elucidate concepts previously unknown or unencountered by the model—a capacity akin to how humans research and retrieve information—our findings imply that DPR models fall short of this mark.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Dai, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *IJCNLP-AACL*, pages 675–718. ACL.

Michele Bevilacqua, Giuseppe Ottaviano, et al. 2022.

Autoregressive search engines: Generating substrings as document identifiers. *NeurIPs*, 35:31668–31683.

Mor Geva, Roei Schuster, et al. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*, pages 5484–5495, Online and Punta Cana, Dominican Republic. ACL.

Vladimir Karpukhin, Barlas Oguz, et al. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, Online. ACL.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, et al. 2023. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172.

Rodrigo Nogueira, Zhiying Jiang, et al. 2020. Document ranking with a pretrained sequence-to-sequence model. In *ACL Findings: EMNLP 2020*, pages 708–718, Online. ACL.

Rodrigo Nogueira, Wei Yang, et al. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Ronak Pradeep, Kai Hui, et al. 2023. How does generative retrieval scale to millions of passages? In *EMNLP*, pages 1305–1321, Singapore. ACL.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yi Tay, Vinh Tran, et al. 2022. Transformer memory as a differentiable search index. *NeurIPs*, 35:21831–21843.

Yujing Wang, Yingyan Hou, et al. 2022. A neural corpus indexer for document retrieval. *NeurIPs*, 35:25600–25614.

Peng Xu, Wei Ping, Xianchao Wu, et al. 2024. Retrieval meets long context large language models. In *ICLR*.

Wayne Xin Zhao, Jing Liu, et al. 2023. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.* Just Accepted.

Noah Ziems, Wenhao Yu, et al. 2023. Large language models are built-in autoregressive search engines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada. ACL.