Related Work

Hao Xiong
Soochow University
libraxionghao@gmail.com

1 Related Work

Open-domain Question Answering (QA for short) requires a system to answer questions based on evidence which are retrieved from a large corpus such as Wikipedia. Current approaches consist of retriever, reranker and reader networks, where the retriever retrieves a small number of documents, and the reranker reranks the retrieved documents for the reader to answer the questions.

Retriever Text retrieval aims to find related documents from a large corpus based on a query. Earlier work (Chen et al., 2017) relied on bag-of-words-based sparse retrievers such as TF-IDF (Chen et al., 2017) and BM25 (Robertson and Zaragoza, 2009). Recently, some work improve the traditional sparse retriever with neural networks, Dai and Callan (2019) use BERT (Devlin et al., 2019) to dynamically generate term weights, and Mao et al. (2021a) utilize text generation method to expand queries or documents to make better use of sparse retriever.

More recent work showed that neural retrievers can generate effective dense representations for retrieval when trained on open-domain QA datasets (Karpukhin et al., 2020). Qu et al. (2021) improve the approach further with the hard negative sampling by iterative training. Izacard and Grave (2022) distill knowledge from reader to retriever. There also exists some researchers focus on the pre-training of dense retrieval (Gao and Callan, 2021).

Reranker Previous work showed that the pre-trained language models demonstrated an outstanding capability in enhancing the performance of both sparse and dense retrievers. Nogueira and Cho (2019) present a supervised reranker for retrieval tasks based on BERT to rerank the retrieved documents. Mao et al. (2021b) propose to rerank by the reader predictions without training a new model. Sachan et al. (2022) use a large language model as the reranker directly, compared to the previous method. Nonetheless, it requires large amounts of computation and time at training and inference stage and performs not well as fine-tuned reranker. Chuang et al. (2023) propose a query reranker to select the best query expand for improving document retrieval.

Reader Reader models for open-domain QA are required to read multiple documents which are more than 100 documents to avoid missing the target document from the large-scale knowledge base. The reader models was divided into two primary categories, the extractive readers (Karpukhin et al., 2020), which encode each document separately and marginalize the predicted answer probabilities and extract the answer spans from the probabilities.

While the generative readers generating the answer in a sequence-to-sequence manner,including the Fusion-in-Decoder (FiD) model (Izacard and Grave, 2020) and the Retrieval-Augmented Generation model (Lewis et al., 2020). The FiD model concatenate the encoded representations of documents, which can then be decoded by the decoder and generate the answer for the query.

2 Method

OpenQA aims to answer factoid questions without pre-specified domains. We assume that a large collection of documents C (i.e., Wikipedia) are given as the resource to answer the questions and a retriever-reader architecture is used to tackle the task, where the retriever retrieves a small subset of the documents $D \subset C$ and the reader reads the documents D to extract (or generate) an answer. Our goal is to improve the effectiveness and efficiency of the retriever and consequently improve the performance of the reader.

In this section, we propose an optimized training approach to dense passage retrieval for open-domain QA, namely RocketQA. We first introduce the background of the dual-encoder architecture,

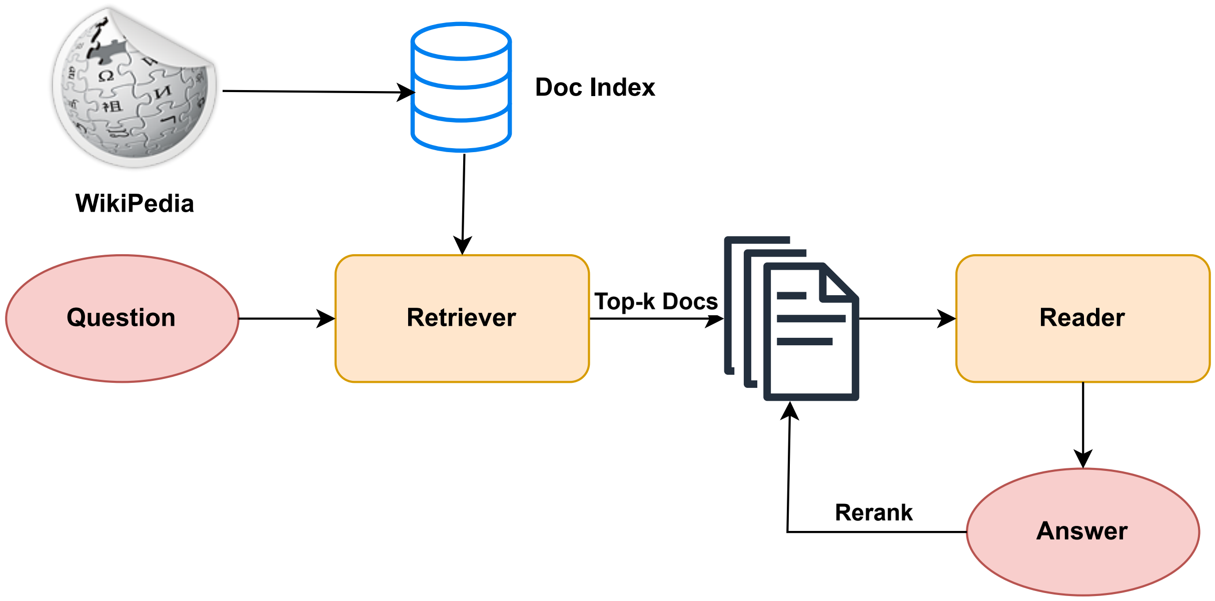

Figure 1: High-level SEAL architecture, composed of an autoregressive LM paired with an FM-Index, for which we show the first (F) and last (L) columns of the underlying matrix (more details in Sec). The FM-index constraints the autoregressive generation (*e.g.*, after *carbon* the model is contrained to generate either *tax*, *dioxide* or *atom* in the example) and provides the documents matching (*i.e.*, containing) the generated ngram (at each decoding step).

and then describe the three novel training strategies in RocketQA. Lastly, we present the whole training procedure of RocketQA.

2.1 Retriever

2.2 Reader

2.3 Reranker

Given an initially retrieved passage list R and topN predictions of the reader $A^{[:N]}$,RIDER forms a reranked passage list R' as follows. RIDER scans R from the beginning of the list and appends to R' every passage $p \in R$ if p contains any reader prediction $a \in A^{[:N]}$ after string normalization (removing articles and punctuation) and tokenization.

Then, the remaining passages are appended to R' according to their original order. Intuitively, if the reader prediction is perfect, the retrieval accuracy after reranking is guaranteed to be optimal. Specifically, if the reader prediction is correct, it is guaranteed that the retrieval accuracy after reranking is better, since RIDER moves all passages containing the correct answer to the top (or at least the same if those passages are all at the top before reranking). If the reader prediction is wrong, RIDER could still be better if the predicted answer co-occurs with the correct answer, the same, or worse if the predicted answer is misleading. In practice, if the reader performs reasonably well, RIDER is also likely to rerank passages well. Overall, we observe quantitatively that RIDER leads to consistent gains in terms of both retrieval accuracy and QA performance without refining the retriever (reader) or even any training itself despite the noise in reader predictions.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. *arXiv preprint arXiv:2305.17080*.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard and Edouard Grave. 2022. Distilling knowledge from reader to retriever for question answering.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021a. Generation-augmented retrieval for open-domain question answering.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021b. Rider: Reader-guided passage reranking for open-domain question answering. *arXiv preprint arXiv:2101.00294*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering.

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

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. *arXiv preprint arXiv:2204.07496*.