Relevance-Guided Retrieval under Continually Expanding Knowledge Bases

Hyundong Jin, Heayoun Choi, and Eunwoo Kim

School of Computer Science and Engineering
Chung-Ang University, South Korea
{jude0316, heayounchoi, eunwoo}@cau.ac.kr

***Abstract*—Retrieval-Augmented Generation (RAG) enhances large language models by incorporating external knowledge to support responses in information-intensive tasks. As knowledge bases expand to accommodate emerging concepts and domains, existing approaches that apply flat retrieval over the entire corpus encounter substantial limitations in efficiency and relevance. To address these challenges, we propose a retrieval method that structures the knowledge base into semantically coherent subsets and confines retrieval to the most relevant subset for each query. This hierarchical procedure enables scalable retrieval under continual expansion of the knowledge base by reducing search space and mitigating irrelevant matches. Experimental results on knowledge-intensive benchmarks show that the proposed method consistently improves both retrieval accuracy and computational efficiency, providing a practical approach to retrieval-augmented generation in expanding knowledge bases.**

***Index Terms*—Retrieval-augmented generation, Continual learning**

I. Introduction

Recently, large language models (LLMs) [1], [2] have demonstrated outstanding performance across diverse tasks [3]. However, their reliance on fixed, model-internal knowledge limits their capacity to answer queries that demand current or specialized information beyond the pretraining corpus. To address this limitation, Retrieval-Augmented Generation (RAG) has been introduced to augment LLMs with non-parametric memory obtained through retrieval of relevant external content. By incorporating retrieved material into the generation stage, LLMs can produce more accurate and contextually enriched responses, particularly for knowledge-intensive tasks.

With the increasing adoption of LLMs in real-world applications that require access to diverse and continuously expanding information, the incorporation of external knowledge sources becomes essential, as shown in Figure 1. Existing RAG methods [4], [5] typically rely on flat retrieval that ignores the evolving structure of knowledge, leading to significant performance drops as the corpus expands incrementally.

Acknowledgment. This research was supported in part through BK21 FOUR (Fostering Outstanding Universities for Research) Program funded by Ministry of Education of Korea (No.I22SS7609062) and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University))

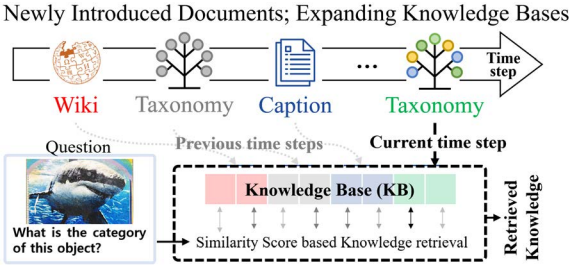

Fig. 1. An illustration of the proposed retrieval setting with an expanding knowledge base. At each time step, external knowledge comprising images and corresponding texts is introduced, and retrieval is performed to obtain relevant information for answering the given query.

To this end, we propose a retrieval approach that structures the knowledge base into semantically coherent subsets, making the retrieval stage scalable. Instead of querying the entire corpus, the method selects the most pertinent subset for each query, enabling efficient and accurate retrieval. We empirically evaluate the effectiveness of our approach on multiple knowledge-intensive benchmarks. Experimental results demonstrate consistent improvements in both retrieval efficiency and response accuracy, highlighting the proposed method as a robust and scalable solution for retrieval-augmented generation under growing knowledge demands.

II. Method

We propose a retrieval framework for scalable integration of a continually expanding knowledge base, where each time step introduces a new collection of image–text documents. The proposed method aims to reduce the retrieval search space by organizing the knowledge base into semantically coherent subsets and restricting retrieval to the most relevant portions. The framework comprises two stages: knowledge base construction and subset-aware retrieval. The overall framework is illustrated in Figure 2.

At each time step, a new collection of image–text documents arrives. The images are encoded using a vision encoder [6], and the resulting embeddings are averaged to produce a representative vector for the document set. This vector is compared with the representative embeddings of existing groups in the knowledge base. If a semantically similar group is identified, the new set is merged into that group.

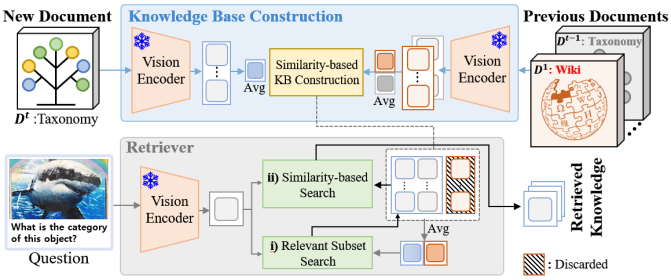

Fig. 2. An overview of the proposed retrieval framework with an expanding knowledge base. New documents are incrementally incorporated by similarity-based grouping, and retrieval proceeds by selecting a relevant subset followed by fine-grained search within it.

For every visual query, retrieval is performed in two stages First, the query embedding, obtained from the same vision encoder, is matched against group-level representatives to select the most relevant subset. Second, within the chosen subset, the query embedding is compared with individual entry embeddings to identify the top-matching image. The text corresponding to the retrieved image is returned as external knowledge for the language model. This hierarchical procedure reduces computational cost and prevents degradation in retrieval accuracy by filtering out irrelevant subsets of the knowledge base and performing retrieval only within the selected subset.

III. Experiments

A. Implementation details

We evaluated the proposed retrieval method using an image classification dataset with a detailed category structure. The dataset was divided into ten subsets, each comprising 20 distinct classes. We assumed an initial knowledge base composed of 100K image–text pairs, provided by [7], and incrementally augmented it by incorporating the training set of one additional subset at each time step. We employed the pre-trained retriever from [5] to extract image embeddings and construct the evolving knowledge base. For evaluation, we used the test set corresponding to the union of all subsets seen up to the current time step. We compared the proposed method with two baselines: (1) retrieval from the entire accumulated knowledge base and (2) retrieval from a single subset selected using the group-matching strategy of the proposed method.

B. Results

Table I reports retrieval performance in terms of Recall@1 and Recall@5 as the number of time steps increases (i.e., after 2, 4, 6, 8, and 10 knowledge base updates). The proposed method consistently maintains high retrieval performance, with Recall@1 of 0.8515 and Recall@5 of 0.9142 even at the final time step, demonstrating consistent retrieval accuracy despite the growth of the knowledge base. When retrieval is performed over the entire accumulated knowledge base, the performance remains lower than that of the proposed method at all time steps. The reduced performance is attributed to the inclusion of numerous irrelevant entries, which hinders precise matching. The method of retrieving from a single subset selected using the group-matching strategy of the proposed method shows a clear decline in performance, with Recall@1 dropping from 0.8460 at step 2 to 0.6471 at step 10, as the limited candidate pool fails to cover relevant entries in an semantically diverse knowledge base. In contrast, the proposed method combines group-level filtering with instance-level retrieval, preserving accuracy while enabling scalability.

TABLE I
Results of retrieval performance as knowledge base expands over multiple time steps.

Method		2	4	6	8	10
Full	R@1	0.8569	0.8359	0.8301	0.8059	0.8026
	R@5	0.9589	0.8334	0.9313	0.9159	0.9117
Single	R@1	0.8460	0.7672	0.7059	0.6642	0.6471
	R@5	0.8781	0.8220	0.7508	0.7072	0.6833
Ours	R@1	0.9275	0.8756	0.8704	0.8517	0.8515
	R@5	0.9753	0.9409	0.9344	0.9171	0.9142

IV. Conclusion

In this paper, we addressed the challenge of scalable retrieval in the context of continually expanding knowledge bases, which hampers both accuracy and efficiency in retrieval-augmented generation. We organised the knowledge base into semantically coherent subsets and adopted a two-stage retrieval procedure that first selects a relevant subset with group-level representatives, and then performs fine-grained search within that subset. This design preserves prior retrieval fidelity while it limits computational cost during expansion. Experiments on knowledge-intensive benchmarks achieved higher retrieval accuracy along with reduced inference time compared to other retrieval baselines. The proposed framework presents a promising solution for efficient retrieval in environments of continually expanding knowledge bases.

References

[1] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[4] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1818–1826.

[5] Y. Yan and W. Xie, "Echosight: Advancing visual-language models with wiki knowledge," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 1538–1551.

[6] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[7] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can pre-trained vision and language models answer visual information-seeking questions?" *arXiv preprint arXiv:2302.11713*, 2023.