MMKG-RAG: Retrieval-Augmented Generation with Multi-Modal Knowledge Graph

Shuaitao Zhao, Shijie Luo, Xinyuan Lu, and Weixiong Rao(✉)

School of Computer Science and Technology, Tongji University, Shanghai, China
{shuaitaozhao,sjlaw,2333089,wxrao}@tongji.edu.cn

Abstract. Multi-modal content nowadays has become increasingly prevalent in modern applications. However, most existing RAG-based LLMs still focus on textual data, and fail to understand rich semantic information in multi-modal content. To overcome the issue above, we demonstrate a novel Multi-Modal Knowledge Graph (MMKG)-based RAG system, namely MMKG-RAG. After constructing comprehensive multi-modal knowledge graphs, MMKG-RAG provides the foundation for sophisticated multi-modal content retrieval and question answering capabilities. Compared to the multi-modal LLMs with no MMKG, the developed MMKG-RAG offers more powerful understanding and retrieval, and meanwhile improves its reasoning capability across modalities.

Keywords: Multi-Modal Knowledge Graph · Large language Models · Retrieval-Augmented Generation

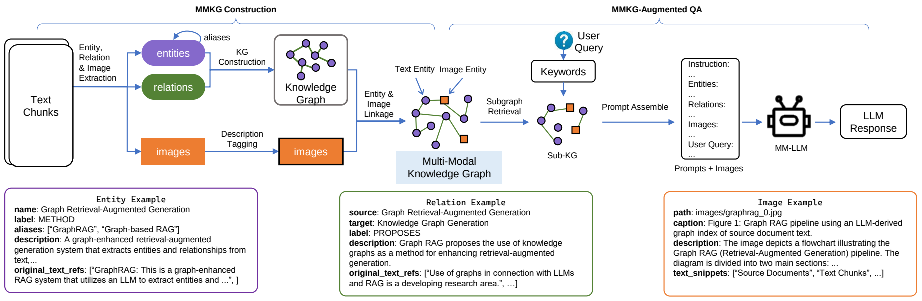
1 Introduction

Multi-modal content nowadays has become increasingly prevalent. Multi-modal Large Language Models (MM-LLM) have stirred up much interest among researchers and practitioners in their impressive skills in many multi-modal content processing tasks [6]. However, MM-LLM still do not work well to understand domain-specific content beyond their training data, and suffer from such issues as hallucination [4]. With the help of external knowledge, RAG techniques have emerged as a viable solution, yet with limited capacity of performing complex reasoning tasks [2]. Recent progress in knowledge graph-based RAG systems, GraphRAG [1] and LightRAG [3], has provided global reasoning capabilities. Nevertheless, these RAG-based previous works mainly focus on textual data, and fail to understand rich semantic information in multi-modal content.

To overcome the issue above, we demonstrate a novel Multi-Modal Knowledge Graph (MMKG)-based RAG system, namely MMKG-RAG, to integrate both textual and visual information. MMKG-RAG leverages MM-LLMs to construct comprehensive multi-modal knowledge graphs, which next provide the foundation for sophisticated multi-modal content retrieval and question-answering capabilities. Compared to the multi-modal LLMs with no MMKG, the developed MMKG-RAG offers more powerful understanding and retrieval capacities, and meanwhile significantly improves its reasoning ability across modalities.

2 System Overview

As shown in Fig. 1, MMKG-RAG consists of two key components. The *MMKG construction module* first builds an MMKG from multi-modality content sources, such as text and images. Then, the *MMKG-augmented QA module* retrieves the MMKG to generate multi-modal response for input questions.



Entity Example
name: Graph Retrieval-Augmented Generation
label: METHOD
aliases: [“GraphRAG”, “Graph-based RAG”]
description: A graph-enhanced retrieval-augmented generation system that extracts entities and relationships from text,...
original_text_refs: [“GraphRAG: This is a graph-enhanced RAG system that utilizes an LLM to extract entities and ...”,]

Relation Example
source: Graph Retrieval-Augmented Generation
target: Knowledge Graph Generation
label: PROPOSES
description: Graph RAG proposes the use of knowledge graphs as a method for enhancing retrieval-augmented generation.
original_text_refs: [“Use of graphs in connection with LLMs and RAG is a developing research area.”, …]

Image Example
path: images/graphrag_0.jpg
caption: Figure 1: Graph RAG pipeline using an LLM-derived graph index of source document text.
description: The image depicts a flowchart illustrating the Graph RAG (Retrieval-Augmented Generation) pipeline. The diagram is divided into two main sections: ...
text_snippets: [“Source Documents”, “Text Chunks”, ...]

Fig. 1: System architecture

2.1 MMKG Construction Module

MMKG-RAG employs a two-stage pipeline to construct the multi-modal knowledge graph (MMKG), involving textual and image entities and relations (illustrated by Fig. 1). In the first stage, MMKG-RAG processes textual data by segmenting input documents into manageable chunks to create a textual KG. It exploits the available MM-LLM (e.g., GPT-4o) to create a collection of $\langle \textit{entity}, \textit{relation} \rangle$ pairs. The created entities are typically textual ones. In addition, MMKG-RAG identifies entity aliases to mitigate entity redundancy within the textual KG.

In the second stage, MMKG-RAG extends the textual KG above by rich modalities such as images. That is, MMKG-RAG first extracts images and identifies their contextual text from the input data source. With the help of an available MM-LLM GPT-4o, MMKG-RAG extracts those images within the input data source, together with their contextual information including figure captions, embedded text, and semantic descriptions. Then, MMKG-RAG creates image entities with the extracted images, and adds them into the textual KG above. Subsequently, MMKG-RAG establishes relation links between the created image entities and available textual entities. Such textual entities are retrieved by searching the extracted contextual information within the textual KG above. Here, MMKG-RAG further exploits the capacity of semantic analysis provided by the available MM-LLM GPT-4o to refine the most appropriate linkages between image entities and textual ones. In this way, MMKG-RAG generates a multi-modal knowledge graph that maintains both structural and semantic relationships across different modalities.

2.2 MMKG-Augmented QA Module

This module involves three steps: query analysis, subgraph retrieval, and answer generation. For a given query, MMKG-RAG first exploits the available MM-LLM GPT-4o to determine whether or not external KG retrieval is required. For those simple queries that can be directly answered by GPT-4o, MMKG-RAG then bypasses the retrieval process, thereby enhancing efficiency [5].

Otherwise, MMKG-RAG extracts keywords and performs query expansion to broaden the search scope of KG retrieval. For each keyword of the input query, MMKG-RAG retrieves a relevant subgraph from the constructed MMKG, involving both textual and image entities and relations. That is, MMKG-RAG first organizes the retrieved textual entities and relations into structured prompts. Such textual prompts, together with the retrieved image entities, then enable the available MM-LLM GPT-4o to generate comprehensive and contextually appropriate responses that effectively leverage both textual and image entities. Compared to the answers directly on GPT-4o, the developed MMKG-RAG generates more reasonable response.

3 Demonstration

Fig. 2 illustrates the screenshots of the two modules above. MMKG-RAG provides a user interface (UI) developed by Gradio [1]. End users first upload the available PDF document sources which may contain both texts and images. Given the uploaded documents, MMKG-RAG converts them into markdown format via an open source tool Marker [2]. MMKG-RAG can optionally generate multiple MMKGs and next merge them together into a single one. Or, MMKG-RAG allows incremental updates of the input PDFs into an existing MMKG. After that, on the QA UI, users can choose an available MMKG, and MMKG-RAG then generates the answers of input queries with help of the chosen MMKG. Such generated answer can effectively incorporate both textual and image information. We demonstrate the MMKG-RAG demo system by implementing a multi-modal knowledge graph. Detailed implementation information and a demonstration video can be accessed in our GitHub repository[3]. In this demo, we provide two following scenarios.

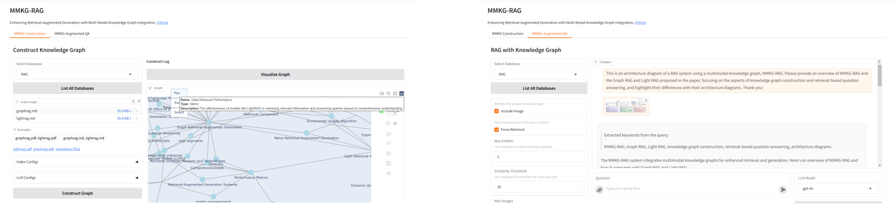
Visually Enhanced Knowledge QA demonstrates complex multi-modal answer generation. In this demo, users may inquire " *the components of GraphRAG and their functions*", MMKG-RAG generates comprehensive responses by combining both textual answer and retrieved images. By such tools as Mermaid[4], MMKG-RAG generates intuitive architectural diagrams that facilitate an intuitive and clear understanding of complex concepts.

[1] https://www.gradio.app

[2] https://github.com/VikParuchuri/marker

[3] https://github.com/wenzhaoabc/mmkg-rag

[4] https://mermaid.js.org/



MMKG-RAG

Enhancing Retrieval-Augmented Generation with Multi-Modal Knowledge Graph Integration. GitHub

Construct Knowledge Graph

Select Database: RAG

Index Graph: graphrag.md 35.0 KB; lightrag.md 39.9 KB

Examples: graphrag.pdf, lightrag.pdf | graphrag.md, lightrag.md

lightrag.pdf graphrag.pdf markdown files

Name: Data Retrieval Performance
Type: Metric
Description: The effectiveness of models like LightRAG in retrieving relevant information and answering queries based on comprehensive understanding.

MMKG-RAG

Enhancing Retrieval-Augmented Generation with Multi-Modal Knowledge Graph Integration. GitHub

RAG with Knowledge Graph

Select Database: RAG

Whether the answer contains images: ☑ Include Image

Force retrieval even if the query is direct: ☑ Force Retrieval

Max Entities (The maximum number of entities selected): 5

Similarity Threshold (The similarity threshold for the retrieval,0-100): 20

This is an architecture diagram of a RAG system using a multimodal knowledge graph, MMKG-RAG. Please provide an overview of MMKG-RAG and the Graph RAG and Light RAG proposed in the paper, focusing on the aspects of knowledge graph construction and retrieval-based question answering, and highlight their differences with their architecture diagrams. Thank you!

Extracted keywords from the query:

MMKG-RAG, Graph RAG, Light RAG, knowledge graph construction, retrieval-based question answering, architecture diagrams

The MMKG-RAG system integrates multimodal knowledge graphs for enhanced retrieval and generation. Here's an overview of MMKG-RAG and how it compares with Graph RAG and Light RAG:

LLM Model: gpt-4o

(a) MMKG construction

(b) MMKG-augmented QA

Fig. 2: User interfaces of MMKG-RAG

Cross-modal Context Understanding can process input multi-modal documents to generate meaningful responses. For example, when users submit an input image regarding *neural networks* together with a textual query, MMKG-RAG then understands the submitted images, and generates responses that merge textual descriptions with visualized aids, thereby effectively highlighting important components and their relations within the neural network.

4 Conclusion and Future Work

The developed MMKG-RAG integrates multi-modal knowledge graph into retrieval-augmented generation. By effectively combining textual and image information, we demonstrate two running examples to generate multi-modal response. As future work, we continue the optimization of MMKG construction, cross-modal retrieval, and the extension of additional modalities such as audio and video.

References

1. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J.: From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024)
2. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
3. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: Lightrag: Simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779 (2024)
4. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems (2023)
5. Jeong, S., Baek, J., Cho, S., Hwang, S.J., Park, J.C.: Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. arXiv preprint arXiv:2403.14403 (2024)
6. Wang, Y., Wang, L., Zhou, Q., Wang, Z., Li, H., Hua, G., Tang, W.: Multimodal llm enhanced cross-lingual cross-modal retrieval. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8296–8305 (2024)