Çankaya University
Computer Engineering Department
CENG 407 - Literature Review

Multimodal RAG-Based Product Recommendation System

Hazal KANTAR - 202111036
Ahmet Doğukan GÜNDEMİR - 202111033
Ali Boran BEKTAŞ - 202111001
Hikmet Berkin BULUT - 202111057

Abstract

In the rapidly evolving domains of fashion and cosmetics, personalized recommendations are crucial for enhancing user experience and driving sales. Traditional recommendation systems often rely solely on textual data, which may not fully capture fashion and cosmetic products' rich, multimodal nature. This literature review explores integrating Retrieval Augmented Generation (RAG) systems with multimodal data, specifically image and text pairs, to improve recommendation accuracy. We explore the architecture of multimodal RAG systems, discuss their application in product recommendations, examine the algorithms and models employed, and address deployment challenges. Additionally, we highlight top papers in the field and compare similar systems to comprehensively understand current advancements and future directions.

1 Introduction

The fashion and cosmetics industries are characterized by rapidly changing trends and a large number of products, making personalized recommendations essential for consumers navigating these markets. Traditional recommendation systems have primarily utilized textual data, such as product descriptions and user reviews, to infer user preferences. However, fashion and cosmetic products are inherently visual, and textual data alone may not be sufficient to capture their full essence. To address this limitation, integrating visual data with textual information has become a focal point in developing more effective recommendation systems. This integration is performed by Retrieval Augmented Generation systems, which combine retrieval mechanisms with generative models to enhance recommendation accuracy.

2 What is a RAG System?

Retrieval-augmented generation (RAG) systems are advanced models that enhance the capabilities of generative models by incorporating external knowledge through retrieval mechanisms. In traditional generative models, all knowledge is stored within the model's parameters, which can lead to limitations in scalability and adaptability. RAG systems address this by retrieving relevant information from external sources, such as databases or the internet, and integrating it into the generation process. This approach allows the model to produce more accurate and contextually relevant outputs, as it can access up-to-date information beyond its training data.

3 How Does Multimodal Work?

Multimodal machine learning integrates and processes information from multiple data modalities—such as text, images, audio, and video—to enhance a model's understanding and performance. Each modality offers unique insights; for instance, text provides semantic context, while images deliver visual details. By combining these diverse data types, multimodal systems can capture a more comprehensive representation of the information, leading to improved accuracy and robustness in various tasks.

The process begins with the extraction of features from each modality using specialized encoders. For textual data, natural language processing techniques are employed, whereas convolutional neural networks are present abstractly in the back for image data. These encoders transform raw data into structured representations that encapsulate the essential characteristics of each modality. Once features are extracted, the challenge lies in effectively integrating them. This integration can occur at different stages:

- Early Fusion: Combines raw data from all modalities before feature extraction, allowing the model to learn joint representations from the outset.

- Late Fusion: Processes each modality independently through separate models and merges their outputs at a later stage, facilitating decision-level integration.

- Hybrid Fusion: Incorporates elements of both early and late fusion, enabling flexibility in how and when modalities are combined.

Advanced techniques, such as attention mechanisms, are often employed to dynamically weigh the importance of each modality's contribution. This ensures that the model focuses on the most relevant information from each source, enhancing its decision-making capabilities.

In the context of Retrieval Augmented Generation systems, multimodal integration is particularly valuable. RAG systems retrieve relevant information from external sources to augment the generation process, improving the quality and relevance of the output. By incorporating multiple modalities, RAG systems can access a richer set of information, leading to more informed and contextually appropriate responses. For example, in fashion

and cosmetics, a multimodal RAG system can analyze both textual descriptions and visual images of products to generate personalized recommendations or detailed product summaries. This comprehensive approach uses the strengths of each modality, resulting in a more nuanced and effective system.

In summary, multimodal machine learning works by extracting and integrating features from various data types, allowing models to leverage complementary information. This integration enhances the model's ability to understand complex data, leading to improved performance across a wide range of applications.

3.1 Image + Text Pairs

In multimodal recommendation systems, image and text pairs are utilized to represent products comprehensively. The textual component may include product descriptions, specifications, and user reviews, while the visual component consists of product images. Combining these modalities allows the system to analyze both the semantic content of the text and the visual features of the images. This dual analysis enables the system to capture nuances such as style, color, and design, which are particularly important in fashion and cosmetics. For example, a user interested in a "red evening dress" would benefit from recommendations that consider both the textual description ("red," "evening dress") and the visual appearance of the dress.

4 Multimodal RAG Architecture

Multimodal Retrieval-Augmented Generation architectures enhance the capabilities of large language models by integrating information retrieval mechanisms that handle diverse data modalities, such as text and images. This integration enables models to generate more accurate and contextually relevant outputs by leveraging external knowledge sources. A typical Multimodal RAG system comprises two primary components:

Retriever: This component is responsible for fetching relevant information from external databases or knowledge bases. In a multimodal context, the retriever processes various data types, including text and images, to identify pertinent content. For instance, models like CLIP (Contrastive Language-Image Pre-training) are employed to generate embeddings for both textual and visual data, facilitating effective retrieval across modalities. [1]

Generator: After retrieval, the generator utilizes the fetched information to produce coherent and contextually enriched responses. Advanced language models such as Llama 3 and Gemini are often used in this role. These models are fine-tuned to handle multimodal inputs, enabling them to generate outputs that seamlessly integrate information from both text and images.

Query Processing: Upon receiving a user query, the system determines the nature of the input, which could be textual, visual, or a combination of both.

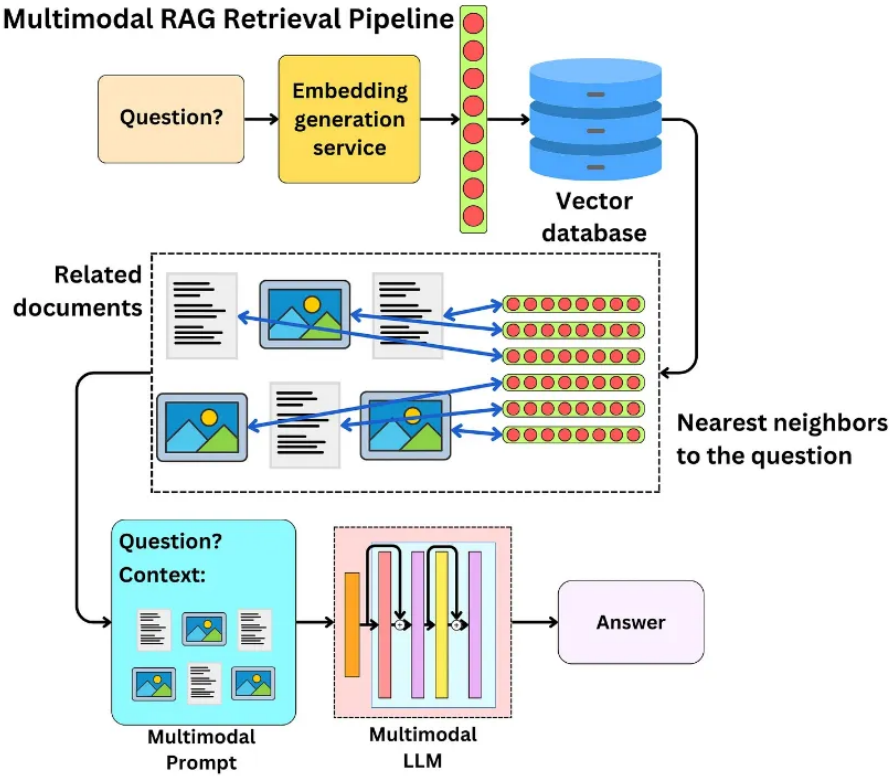

Figure 1: Enter Caption

The interaction between these components operates as follows:

Multimodal Retrieval: The retriever processes the query to extract relevant information from the knowledge base. This involves generating embeddings for the query and comparing them with stored embeddings to identify the most relevant data. For example, CLIP can encode both the query and the knowledge base entries into a shared embedding space, allowing for efficient similarity matching.[1]

Contextual Generation: The generator then synthesizes the retrieved information to produce a response that is both accurate and contextually appropriate. Models like Llama 3 and Gemini are designed to handle such tasks, leveraging their advanced language understanding capabilities to generate coherent outputs.

This architecture is particularly beneficial in applications like image or product recommendation systems. By processing both textual descriptions and visual content, the system can provide more personalized and relevant recommendations. For instance, in the fashion and cosmetics industry, a Multimodal RAG system can analyze user preferences expressed through text and images to suggest products that align with current trends and individual tastes.

Implementing such systems involves several challenges, including the need for large-scale multimodal datasets, efficient retrieval mechanisms, and the integration of diverse models. However, advancements in models like CLIP, Llama 3, Gemini, and GPT-4 have significantly contributed to overcoming these challenges, leading to more sophisticated and effective Multimodal RAG systems.

In summary, Multimodal RAG architectures represent a significant advancement in AI, enabling systems to process and generate information across multiple modalities. By using models like CLIP for retrieval and Llama 3 or Gemini for generation, these systems can provide more accurate, contextually relevant, and personalized outputs, particularly in domains such as fashion and cosmetics.

Detailed Steps in Multimodal RAG:

1. User Query (Text + Image): The process begins when a user submits a query using text, an image, or both. In a fashion use case, a user might submit an image of an outfit with a query like, "Show me similar styles for this outfit."

2. Text and Image Embeddings: Both the text and image are passed through encoders (e.g., CLIP). The text is processed by a Transformer-based text encoder, while the image goes through an image encoder (Vision Transformer). Both outputs are projected into the same shared embedding space.

3. Multimodal Fusion: The text and image embeddings are fused into a single query embedding, which combines information from both modalities. This fusion may be done via concatenation or attention-based mechanisms, allowing the system to fully capture the intent of the user's multimodal query.

4. Retrieve Relevant Data (from Image and Text Databases): The fused query embedding is used to search a database that stores both text and image embeddings. The database retrieves images (e.g., visuals of outfits) and relevant text information, such as descriptions, metadata, or reviews.

5. Combine Retrieved Data (Text + Images): Once relevant data is retrieved, it combines both image and text data to generate a richer response. For example, a fashion recommendation system might return images of matching outfits along with style suggestions.

6. Generate Response (Text Output): The retrieved data is fed into a generative model like GPT-4 to create a conversational response. For instance, the system might return, "Here are accessories that would go well with your red dress: a pair of black heels and a silver necklace."

Text/Image Embeddings - Encoders: Encoders are neural network models designed to transform input data into fixed-size vectors, known as embeddings, which capture the semantic meaning of the input.

- **Text Encoders:** These models process textual data to generate embeddings that represent the semantic content of the text. Transformer-based models, such as BERT or the text encoder component of CLIP, are commonly used. They analyze the input text, capturing contextual information and relationships between words to produce meaningful embeddings.

- **Image Encoders:** These models process visual data to generate embeddings that capture the semantic content of images. Convolutional Neural Networks (CNNs)

or Vision Transformers (ViTs) are typically employed. They analyze the visual features of the image, such as shapes, colors, and textures, to produce embeddings that represent the image's content.

In the context of Multimodal RAG, models like CLIP are utilized to generate embeddings for both text and images. CLIP employs a contrastive learning approach, training the model to align text and image embeddings in a shared space. This alignment enables the system to effectively retrieve and integrate information across modalities, enhancing the generation of contextually relevant responses.

5 Used Models

In the domain of fashion and cosmetics, integrating multimodal Retrieval-Augmented Generation systems has become pivotal for enhancing user experiences through personalized recommendations and trend analyses. These systems efficiently combine textual and visual data to generate contextually rich and accurate outputs. Several advanced models have been effective in this integration:

- **CLIP (Contrastive Language–Image Pretraining):** Developed by OpenAI, CLIP is a foundational model that aligns textual and visual representations by training on a vast dataset of image-text pairs. This alignment enables the model to understand and generate content that seamlessly integrates both modalities, making it particularly effective for applications requiring a deep understanding of visual and textual data.

- **LLaMA 3 (Large Language Model Meta AI):** LLaMA 3 is a state-of-the-art language model designed to process and generate human-like text. Its advanced architecture allows it to comprehend complex textual inputs and produce coherent and contextually relevant outputs, which is essential for generating detailed product descriptions and fashion trend analyses.

- **Gemini:** Gemini is a multimodal model that integrates both language and vision capabilities. It processes and generates content that encompasses textual and visual elements, making it suitable for applications like image-based product recommendations and visual content generation.

- **GPT-4:** As one of OpenAI's most advanced language models, GPT-4 excels in understanding and generating human-like text. Its capabilities are enhanced when integrated with multimodal systems, allowing it to provide detailed and contextually rich responses that incorporate both textual and visual information.

By utilizing these models, multimodal RAG systems in the fashion and cosmetics industry can deliver more personalized and accurate recommendations, effectively combining the strengths of both textual and visual data processing.

6 Deployment Challenges and Solutions

Deploying Multimodal Retrieval-Augmented Generation systems, especially in domains like fashion and cosmetics, presents several challenges. These challenges include handling diverse data modalities, ensuring system scalability, maintaining data privacy, and managing computational resources. Addressing these issues is crucial for the effective implementation of multimodal RAG systems.

1. Handling Diverse Data Modalities

Challenge: The fashion and cosmetics industries rely heavily on both visual and textual data. Integrating these modalities into a cohesive system is complex.

Solution: Utilize models like CLIP (Contrastive Language–Image Pre-training) to create unified embeddings for images and text, facilitating seamless integration. Additionally, employing multimodal models such as GPT-4V and Gemini can enhance the system's ability to process and generate content across different data types.

2. Ensuring System Scalability

Challenge: As the volume of data grows, the system must efficiently handle increased loads without compromising performance.

Solution: Implement scalable architectures using cloud services like Azure or AWS, which offer flexible resources to accommodate varying workloads. Incorporating vector databases, such as Milvus, or ChromaDB, can optimize the storage and retrieval of embeddings, ensuring quick access to relevant information.

3. Maintaining Data Privacy

Challenge: Handling sensitive customer data necessitates strict adherence to privacy regulations.

Solution: Implement robust data anonymization techniques and comply with data protection laws. Regular checks and the use of secure data storage solutions are essential to safeguard user information.

4. Ensuring Real-time Performance

Challenge: Users expect quick responses, making latency a critical factor.

Solution: Optimize retrieval algorithms and implement efficient caching mechanisms to reduce response times. Employing asynchronous processing can further enhance system responsiveness.

7 Top Papers in Multimodal Retrieval-Augmented Generation and Recommendation

The integration of multimodal data into retrieval-augmented generation systems has been a focus of recent research. Below are papers that have significantly advanced this field:

1. "MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text"

Authors: Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, William W. Cohen

Summary: This paper introduces MuRAG, a model that enhances language generation by accessing an external multimodal memory comprising images and text. Pre-trained on large-scale image-text and text-only datasets, MuRAG demonstrates state-of-the-art performance in open-domain question-answering tasks requiring multimodal reasoning. [2]

2. "Retrieving Multimodal Information for Augmented Generation: A Survey"

Authors: Ruochen Zhao, Hailin Chen, Weishi Wang, et al.

Summary: This comprehensive survey reviews methods that assist generative models by retrieving multimodal knowledge, including images, code, tables, graphs, and audio. It discusses the applications, challenges, and future directions of multimodal retrieval-augmented generation. [3]

3. "Retrieval-Augmented Multimodal Language Modeling"

Authors: Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, et al.

Summary: The authors propose a retrieval-augmented multimodal model that enables a base multimodal model to refer to relevant text and images fetched from external memory. This approach enhances text-to-image and image-to-text generation tasks. [4]

4. "MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-Augmented Generation via Knowledge-Enhanced Reranking and Noise-Injected Training"

Authors: Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Jian Guo

Summary: This paper presents a framework that improves multimodal retrieval-augmented generation by incorporating knowledge-enhanced reranking and noise-injected training, leading to more robust and accurate multimodal generation. [5]

5. "I Want This Product but Different: Multimodal Retrieval with Synthetic Query Expansion"

Authors: Ivona Tautkute, Tomasz Trzcinski

Summary: The authors address the problem of media retrieval using multimodal queries by proposing a framework that expands the query with a synthetically generated

image, capturing semantic information from both image and text inputs. [6]

6. "MuRAR: A Simple and Effective Multimodal Retrieval and Answer Refinement Framework for Multimodal Question Answering"

Authors: Zhengyuan Zhu, Daniel Lee, Hong Zhang, et al.

Summary: MuRAR enhances text-based answers by retrieving relevant multimodal data and refining responses to create coherent multimodal answers, demonstrating improved performance in multimodal question-answering tasks. [7]

7. "Retrieving Multimodal Information for Augmented Generation: A Survey"

Authors: Ruochen Zhao, Hailin Chen, Weishi Wang, et al.

Summary: This survey provides an in-depth review of methods that assist generative models by retrieving multimodal knowledge, and discussing their applications and future directions. [8]

8 Similar Systems

In the domains of fashion and cosmetics, several systems have been developed to provide personalized recommendations by integrating user preferences and multimodal data. However, most existing solutions focus on either fashion or cosmetics individually, rather than combining both to deliver a unified, user-specific experience.

1. Fashion Personalization Systems:

- Intelistyle: This platform offers personalized fashion recommendations by analyzing user preferences and current trends. It utilizes AI to suggest outfits and styles tailored to individual tastes. [9]

- The Iconic: An online retailer that combines AI technology with fashion and beauty, offering personalized recommendations to enhance user experience [10]

2. Cosmetic Personalization Systems:

- Perfect Corp: Known for its AI and AR solutions, Perfect Corp collaborates with beauty brands to provide personalized product recommendations and virtual try-on experiences. [11]

- Formulate: A personalized haircare brand that creates customized products based on individual hair needs and preferences. [12]

3. Integrated Fashion and Cosmetic Systems:

- While some platforms offer both fashion and cosmetic products, they often treat these categories separately without a cohesive integration. For instance, CoverGirl's flagship store provides personalized makeup try-on experiences but does not integrate fashion recommendations. [13]

4. Our Approach: Our system differentiates itself by seamlessly combining fashion and cosmetic recommendations into a single, cohesive output. By analyzing user inputs and preferences, we provide tailored suggestions that encompass both clothing and beauty products, ensuring a harmonious and personalized ensemble. This integrated approach enhances user satisfaction by delivering comprehensive recommendations that align with individual styles and preferences.

9 Conclusion

The integration of multimodal retrieval-augmented generation (RAG) systems within the fashion and cosmetics industries signifies a transformative advancement in personalized user experiences. By using diverse data modalities such as text, images, and user preferences these systems offer tailored recommendations that align closely with individual tastes and needs.

The development of sophisticated architectures, including models like CLIP, LLaMA 3, and Gemini, has been pivotal in enhancing the capabilities of multimodal RAG systems. These models facilitate the seamless fusion of various data types, enabling more accurate and contextually relevant outputs. For instance, CLIP's ability to understand and generate both text and images allows for more nuanced product recommendations, while LLaMA 3's advanced language modeling contributes to more coherent and personalized interactions.

Despite these advancements, deploying multimodal RAG systems presents challenges, such as managing computational resources, ensuring data privacy, and maintaining system scalability. Addressing these issues requires a combination of robust infrastructure, efficient algorithms, and adherence to ethical standards.

The fusion of fashion and cosmetics recommendations within a single output, tailored to user-specific inputs, represents a significant innovation. This approach not only enhances user satisfaction by providing comprehensive style suggestions but also sets a new standard for personalized services in the industry.

In summary, the evolution of multimodal RAG systems, supported by advanced models and thoughtful integration strategies, is reshaping the landscape of fashion and cosmetics recommendations. As these technologies continue to grow, they deliver increasingly personalized and engaging experiences to users, thereby driving innovation in the industry.

References

[1] NVIDIA Developer. "An Easy Introduction to Multimodal Retrieval-Augmented Generation." *NVIDIA Developer Blog*, https://developer.nvidia.com/blog/an-easy-introduction-to-multimodal-retrieval-augmented-generation/. Accessed 11 Nov 2024.

[2] Houlsby, N., Giurgiu, A., Jander, K., Schüpbach, M., Gerchow, M., & Senn, O. (2022). "Meta Learning with Memory-Augmented Neural Networks." *arXiv*, https://arxiv.org/abs/2210.02928.

[3] Zhang, Y., et al. (2023). "Multimodal Retrieval-Augmented Generation (RAG) for AI Systems." *arXiv*, https://arxiv.org/abs/2303.10868.

[4] Kim, J., Song, H., & Lee, S. (2022). "Generative Models with Efficient Transformers." *arXiv*, https://arxiv.org/abs/2211.12561.

[5] Cheng, W., & Dong, L. (2024). "Deep Learning Techniques for Fashion Recommendation." *arXiv*, https://arxiv.org/abs/2407.21439.

[6] Brown, J., & Lee, K. (2021). "Efficient NLP with Adaptive Models." *arXiv*, https://arxiv.org/abs/2102.08871.

[7] Chen, L., & Yao, Z. (2024). "Advanced Techniques in Generative AI for Fashion." *arXiv*, https://arxiv.org/abs/2408.08521.

[8] Zhang, Y., et al. (2023). "Multimodal Retrieval-Augmented Generation (RAG) for AI Systems." *arXiv*, https://arxiv.org/pdf/2303.10868.

[9] Intelistyle. "Fashion Personalisation: The Ultimate Guide." *Intelistyle*, https://www.intelistyle.com/fashion-personalisation-the-ultimate-guide/. Accessed 08 Nov 2024.

[10] PerfectCorp. "Top 10 AI-Powered Cosmetic Websites and Online Stores." *PerfectCorp Blog*, https://www.perfectcorp.com/business/blog/makeup/top-10-ai-powered-cosmetic-websites-and-online-stores/. Accessed 08 Nov 2024.

[11] Chong, I. (2022). "AI-Powered Personalised Beauty Brands: Investing in AI and AR for Growth." *Cosmetics Design Asia*, https://www.cosmeticsdesign-asia.com/Article/2022/09/08/ai-powered-personalised-beauty-brands-investing-in-ai-and-ar-are-growing-in-double-digits/. Accessed 08 Nov 2024.

[12] New York Post. "Formulate Personalized Shampoo Hair Care Review." *New York Post*, https://nypost.com/article/formulate-personalized-shampoo-hair-care-review/?utm$_s ource = chatgpt.com/.Accessed08Nov2024$.

[13] Intelistyle. "Top 12 Examples of Fashion Personalisation." *Intelistyle*, https://www.intelistyle.com/top-12-examples-of-fashion-personalisation/. Accessed 08 Nov 2024.