Multimodal RAG

Up until now, most AI models have been limited to a single modality (a single type of data like text or images or video). Recently, there has been significant progress in AI models being able to handle multiple modalities (majorly text and images). With the emergence of these Large Multimodal Models (LMMs) a multimodal RAG system becomes possible.

"Generate any type of output from any type of input providing any type of context"

The high-level features of multimodal RAG are -

1. Ability to **query/prompt in one or more modalities** like sending both text and image as input.
2. Ability to **search and retrieve not only text** but also images, tables, audio files related to the query
3. Ability to **generate text, image, video etc.** irrespective of the mode(s) in which the input is provided.

Approaches

Using MultiModal Embeddings

Using LMMs Only

Large MultiModel Models

Google DeepMind Flamingo

salesforce BLIP

Microsoft KOSMOS-1

Tencent Macaw-LLM

OpenAI GPT4V

Google Gemini

LlaVA

LAVIN

LLaMA - Adapter

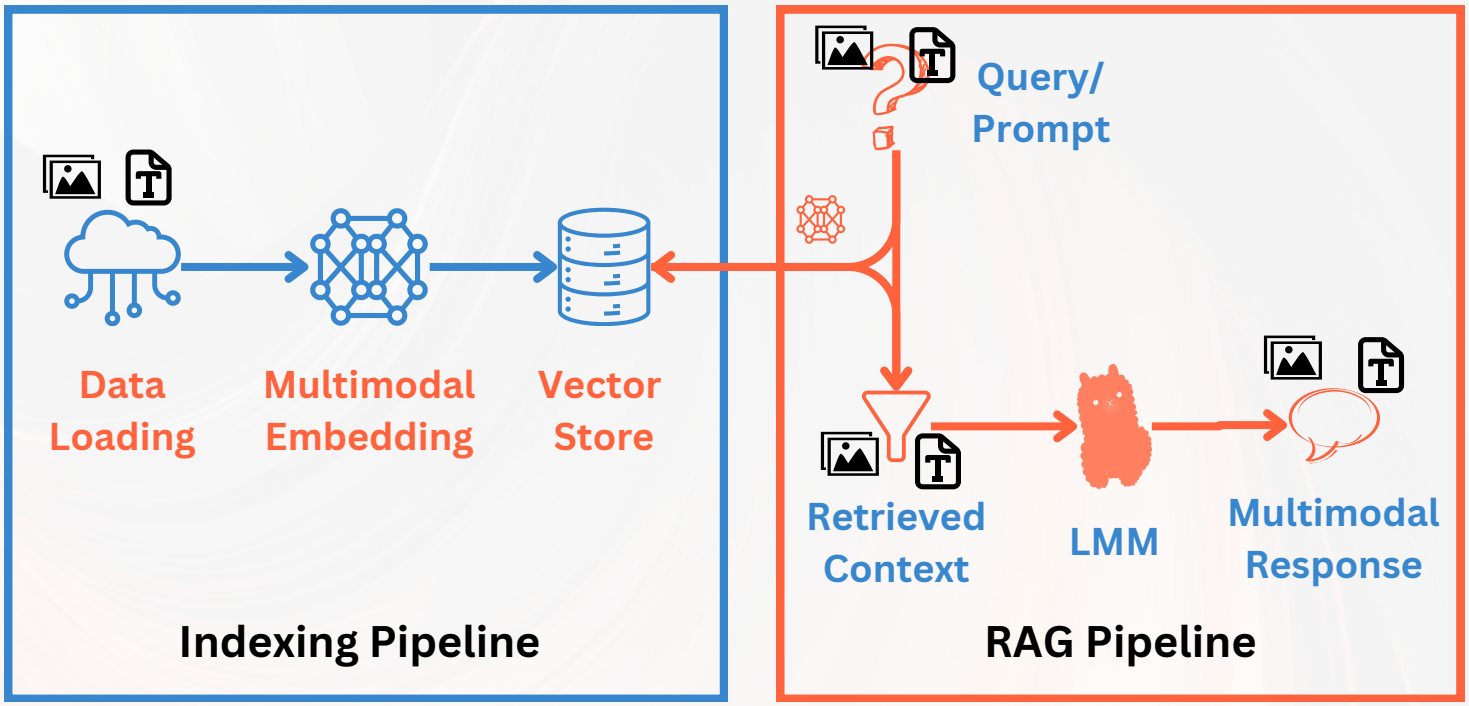
FUYU

Abhinav Kimothi

Multimodal RAG Approaches

Using MultiModal Embeddings

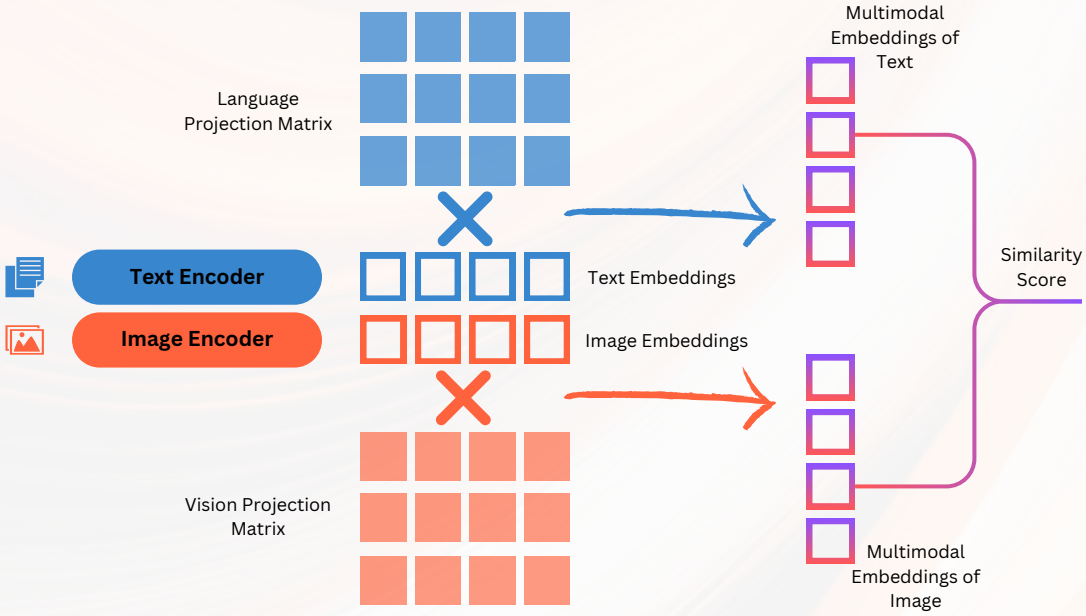
- Multimodal embeddings (like **CLIP**) are used to embed images and text
- User Query is used to retrieve context which can be image and/or text
- The image and/or text context is passed to an LMM with the prompt.
- The LMM generates the final response based on the prompt



Multimodal RAG using Multimodal Embeddings

CLIP : Contrastive Language-Image Pre-training

Mapping data of different modalities into a shared embedding space



CLIP is an example of training multimodal embeddings

OpenAI's CLIP (Contrastive Language-Image Pre-training), maps both images and text into the same semantic embedding space. This allows CLIP to "understand" the relationship between texts and images for powerful applications

Abhinav Kimothi

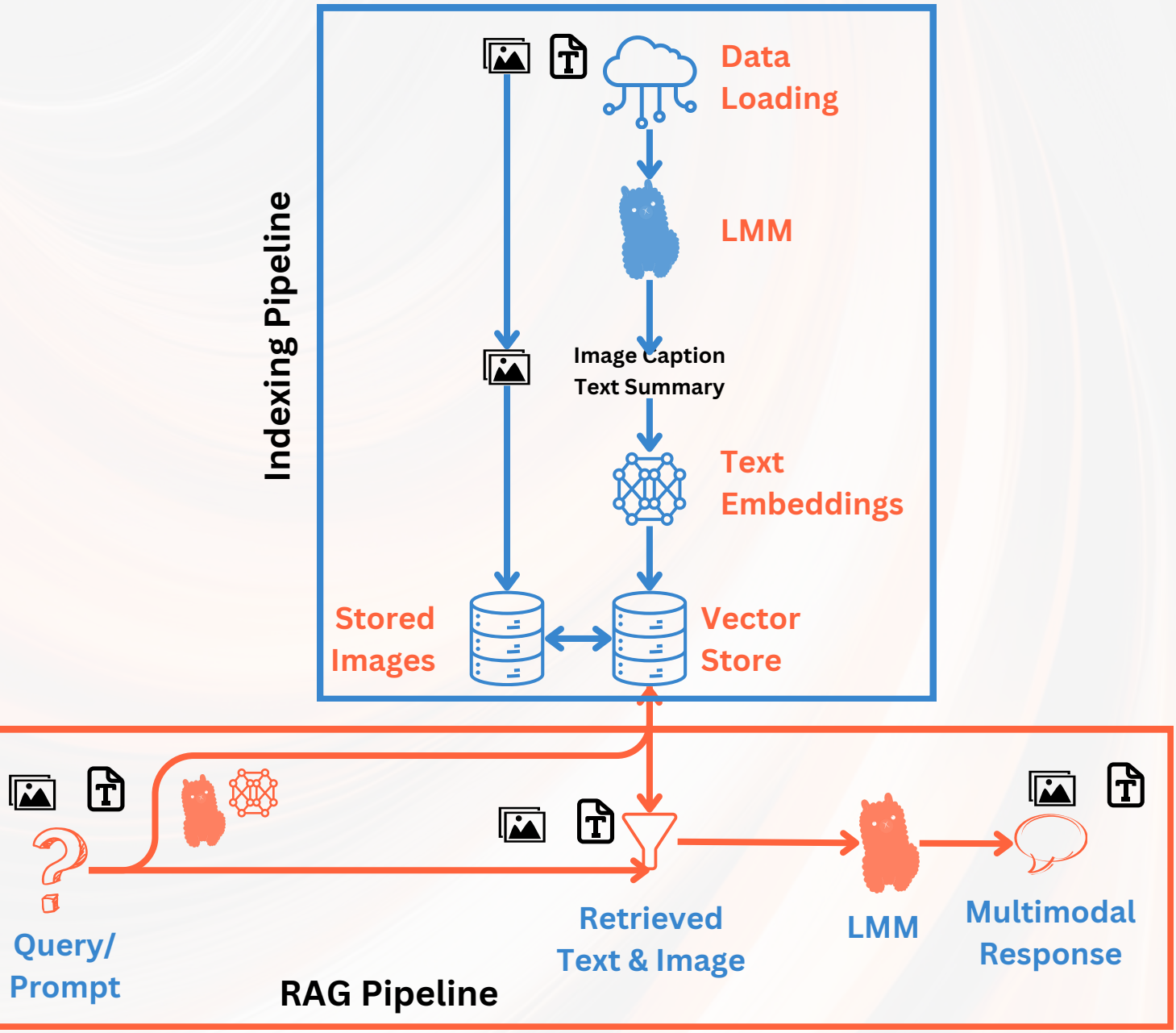
Using LMMs to produce text summaries from images

Indexing

- An LLM is used to generate captions for images in the data
- The image captions and text summaries are stored as text embeddings in a vector database
- A mapping is maintained from the image captions to the image files

Generation

- User enters a query (with text and image)
- Image captions are generated using an LLM and embeddings are generated
- Text summaries and image captions are searched. Images are retrieved based on the relevant image captions.
- Retrieved text summaries, captions and images are passed to the LMM with the prompt. The LMM generates a multimodal response



Resources

LangChain

Cookbook

Multi-Vector Retriever for RAG on tables, text, and images

LlamaIndex

Tutorial

Multi-modal Retrieval Augmented Generation with LlamaIndex

Other Blogs on RAG

Getting the Most from LLMs: Building a Knowledge Brain for Retrieval Augmented Generation

The advancements in the LLM space have been mind-boggling. However, when it comes to using LLMs in real scenarios, we...

medium.com

Evaluation of RAG Pipelines for more reliable LLM applications

Building a PoC RAG pipeline is not overtly complex. LangChain and LlamaIndex have made it quite simple. Developing...

medium.com

LLM Notes

Generative AI with Large Language Models (Coursera Course Notes)

Generative AI with LLMs The world of Language Model-based AI has captured the imagination of all tech enthusiasts and...

abhinavkimothi.gumroad.com