Module 10 of 10

Practice 6: Multimodal RAG with ColPali-like approach

Grid University

Phase 6 - Multimodal RAG with ColPali-like approach

Objective: Explore and implement a more end-to-end multimodal RAG system, leveraging models like ColPali (or more recent equivalents like LLaVA, GPT-4V for querying, or specialized multimodal embedding models) to handle the direct processing of combined image and text data.

Tasks:

1 **Visual Document Ingestion and Preprocessing:**

- Develop capabilities to ingest documents where layout and visual elements are key (e.g., PDFs).
- Implement a process for converting document pages into images and segmenting these images into patches for detailed analysis.

2 **Multimodal Embedding Generation:**

- Select and integrate a Vision Language Model (VLM) or a similar model architecture (e.g., PaliGemma, or custom combination) capable of generating contextualized embeddings from visual document patches, capturing both text and visual features.
- Establish a pipeline for generating these multimodal embeddings and storing them in a vector database, including relevant metadata linking back to source document page and patch location.

3 **Multimodal Retrieval System:**

- Adapt or implement a query embedding process suitable for matching against visual patch embeddings.
- Develop and integrate a retrieval mechanism, such as a "late interaction" or "MaxSim" approach (similar to ColBERT/ColPali), to identify and rank the most relevant document page patches based on the user's query.

4 **Contextual Generation with Visual Context:**

- Ensure the retrieved visual patches (or entire pages) are effectively provided as context to a multimodal LLM.
- The generative LLM must be capable of interpreting and synthesizing information from both the textual query and the provided visual context to formulate answers.

5 **Enhanced Source Attribution:**

- Implement features to visually indicate or reference the specific regions or patches within the source document that contributed to the generated answer, improving transparency.

6 **Compare the results with previous pipeline.**

Good job!