Notes video:

Embeddings should be Modality independent I,e video, audio, image, text

Contrastive representaion learning: task of combining multiple different embedding model into one multimodal embdding model to generate one vector space .i.e. can either train one model to to be multimodel or unify different models

How it works: provide negative and positive examples of similar concepts.  
ignore this concept for now. Required prerequisite: CNN

Multimodal embeddings: image of lion will be placed close to the video of the lion and similarly to the text of the lion. It produces joint vector space for different modalities

My insights:

There is a difference between images understood by embedding model and vision Model.

image embedding model convert image into single vector while vision model converts it into different patches.

*! will search on that*

How llm understand patches of an image of document: feature extraction through patch+ OCR extraction + spatial ques

Spatial ques aware models : docLLm

L3 notebook: markdown function seems interesting

Llm can even see hidden msg inside the picture..how? in 4th video, needs more research at the end!

L4, so the whole pipline of MRAG would be  
store the images as embeddings, store text data as embedding, then query the text and query the image, get the image, give it to llm for interpretation…..seems stupid

Three major ways to conduct multimodel rag  
Option 1:

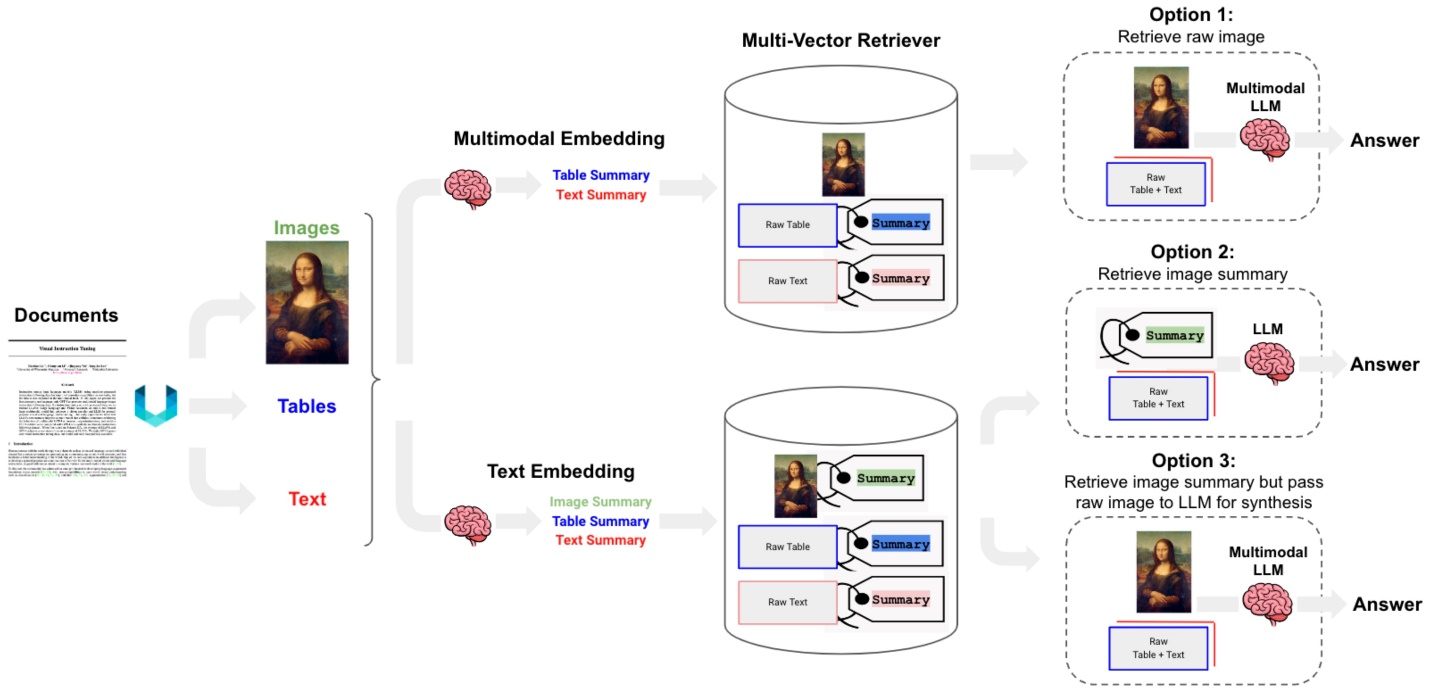
* Use multimodal embeddings to embed images and text
* Retrieve both using similarity search
* Pass raw images and text chunks to a multimodal LLM for answer synthesis

Option 2:

* Use a multimodal LLM to produce text summaries from images
* Embed and retrieve text
* Pass text chunks to an LLM for answer synthesis

Option 3

* Use a multimodal LLM to produce text (feature extraction) from images
* Embed and retrieve image summaries (feature extraction) with a reference to the raw image
* Pass raw images and text chunks to a multimodal LLM for answer synthesis



Option 3 is the best choice.

Started using unstructured library to extract tables, images and text from pdf….awesoomely worked. Will understand everything about it

Ok so progress so far in option 3, we split text, tables and images from pdf. Have a raw form of all of theses and summary (feature extraction) form of all of these (yes including text)  
then create a doc store for raw forms and and vectore store for sumaaries of all these….vectors  
why doc store? How they have been linked \*\* very important.

One pros of 3 could be, we don’t need multi m embedding model.

Text vs texts\_4k tokens. These two variables has an issue when embeddings summaries

Could there be a way I have texts in embeddings, no summaries for text. But rest stay same, if I ca find out linking mechanism….found it..complex…, I could do better. (metadata)

Unstructured library : need more research, can we give it any pointers to get a picture right.

The library supports a wide range of document types, including:

* Text files (.txt)
* Word documents (.docx, .doc, .odt)
* Presentations (.pptx, .ppt)
* Spreadsheets (.xlsx, .csv, .tsv)
* Emails (.eml, .msg)
* Rich Text Format (.rtf)
* E-books (.epub)
* Web pages (.html, .xml)
* Images (.png, .jpg, .heic)
* **Partitioning Strategies**: The function offers several strategies for processing PDFs:
  + "auto": Automatically chooses the strategy based on document characteristics and function kwargs.
  + "hi\_res": Identifies the layout of the document using a high-resolution strategy, necessary for extracting tables.
  + "ocr\_only": Applies OCR to extract text from images within the PDF.
  + "fast": Uses a faster, less accurate method for text extraction.

Conclusion:

* **Multimodal embeddings create a joint vector space for different modalities.** This means you can search and compare data across images, text, audio, and video, leading to more powerful and comprehensive applications.
* **LLMs understand images through patches, OCR, and spatial cues.** This allows LLMs to process visual information in a way that's similar to how humans understand images, enabling them to analyze documents and images with greater depth.
* **Option 3 (**(feature extraction) from **images with a multimodal LLM, embedding those extracted feature text, and then passing raw data to a multimodal LLM) is the most effective approach for Multimodal RAG.** This combines the efficiency of text-based retrieval with the richness of multimodal understanding.
* **Combining document stores and vector stores is crucial for Multimodal RAG.** Document stores hold the raw data, while vector stores enable efficient similarity search. Linking these two effectively is key to a robust system.

References:

Paper to improve vision model performance through Mrag <https://cs.stanford.edu/~myasu/blog/racm3/>

Read docllm paper for creating a solution to better parse pdf data:

<https://arxiv.org/pdf/2401.00908>

Future Tasks:

KAI

Multimodality test on video, with different llms. With respect to every aspect of a video, e.g. audio

Testing timeframe refrences….

Experiment: give a video and its bad transcript and tell it to correct transcript (assumption: llm can understand timestamps from video and correct the timestamps text in bad transcript )

Me

Create your own BEST Mrag pipeline. Without abstractions of langchain.