Large Multimodal Model and Its Potential Impact to Document Understanding

overview

1. What is LMM?

2. Why LMM?

3. LMM Architeture

4. LMM Training Strategy and Data

5. LMM for Document Understanding

What is LMM?

Large Multimodal Model <-> **Multimodal Large Language Model** <-> **Large Vision-Language Models**

- LLM-based model with the ability to **receive**, **reason**, and **output** with multimodal information.
- LMM manifests two representative traits:
  - LMM is based on LLM with billion scale parameters, which is not available in previous multimodal models.
  - LMM uses new training paradigms to unleash its full potential, such as multimodal instruction tuning to encourage the model to follow new instructions.

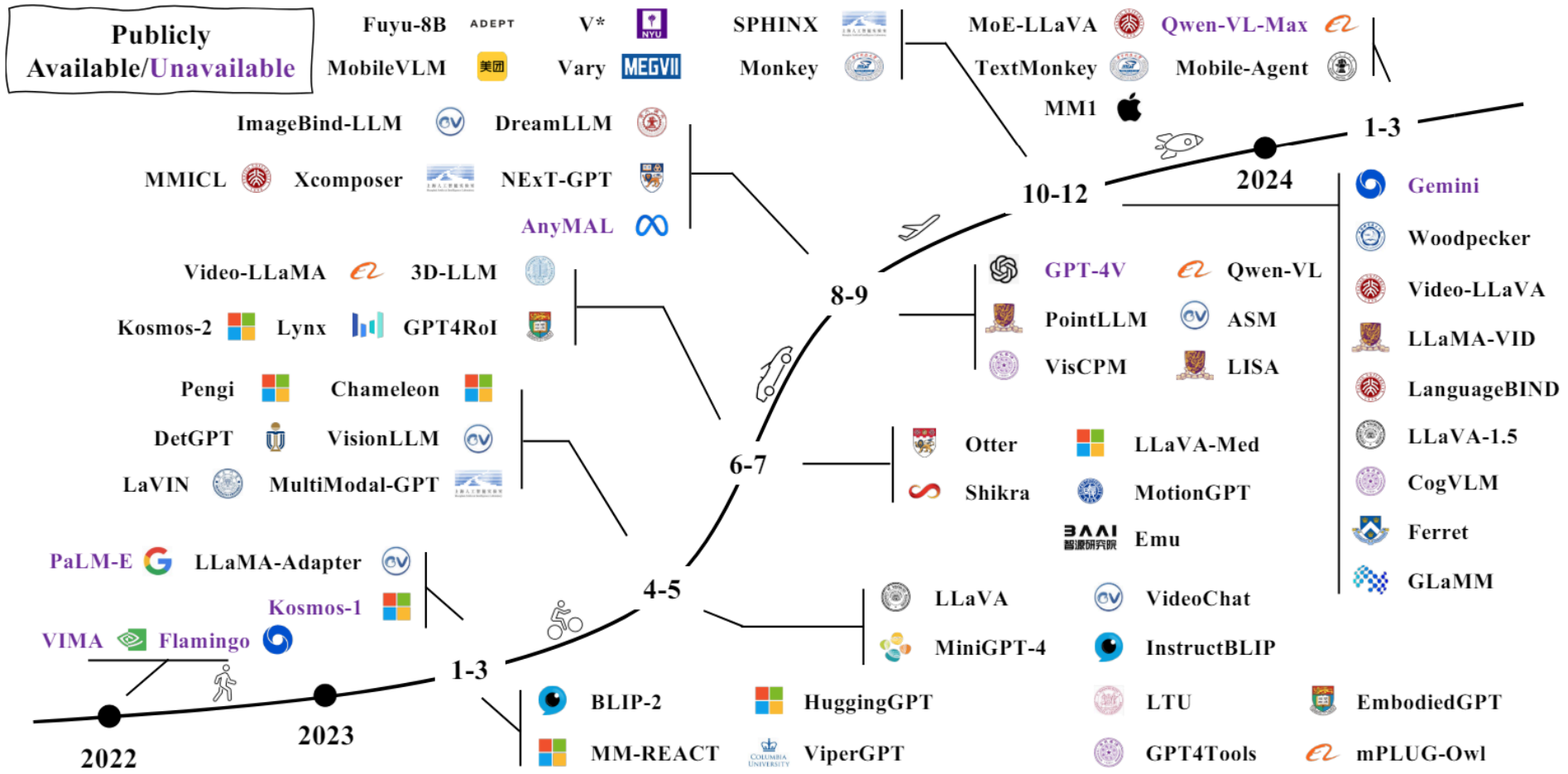
The motivation

This research is to see if we can leverage the recent development of LMM to benefit the Document Understanding solution in ATO.

1. improving the performance of models in the DU pipeline;
2. improving the throughput of the DU pipeline;
3. simplifying the DU pipeline by replacing multiple models with one or two models;
4. simplifying the training or inference of the models in the DU pipeline.

Recent LMM Development

On the other hand, there has been a boom in research and development of LMM in recent years, particularly since last year.



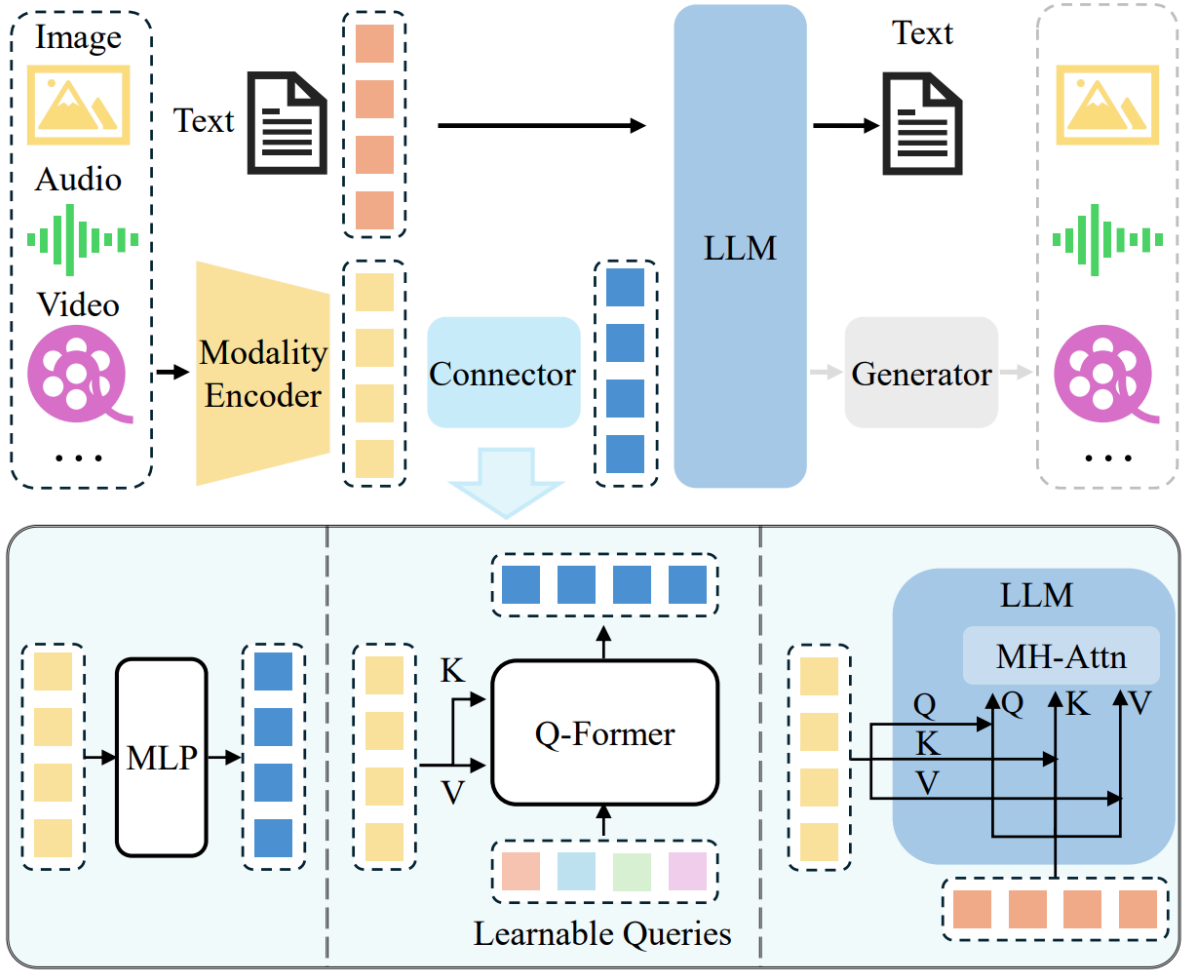
LMM Architecture (1)

A typical LMM can be abstracted into three modules, i.e.,

1. a pre-trained modality encoder,
2. a pre-trained LLM, and
3. a modality interface to connect them.

LMM Architecture (2)

The typical LMM architechture is as follows



LMM Training Strategy and Data (1)

Pre-training: entails large-scale text-paired data

(1) aligning different modalities and
(2) providing world knowledge.

A common approach for pre-training is to keep pre-trained modules (e.g. visual encoders and LLMs) frozen and train a learnable interface.

LMM Training Strategy and Data (2)

Instruction-Tuning

Intuitively, instruction tuning aims to teach models to better understand the instructions from users and fulfill the demanded tasks. Instruction tuning learns how to generalize to unseen tasks rather than fitting specific tasks like the two counterparts.

A multimodal instruction sample often includes an optional instruction and an input-output pair. The instruction is typically a natural language sentence describing the task, such as, “Describe the image in detail.” The input can be an image-text pair like the VQA task or only an image like the image caption task. The output is the answer to the instruction conditioned on the input.

Some research in LMM

1. MM1
2. LLaVA-1.5
3. InternVL

MM1:

Key contribution is study the importance of various architecture components and data choices.

Through careful and comprehensive ablations of the image encoder, the vision language connector, and various pre-training data choices.

Findings:

1. large-scale multimodal pre-training using a careful mix of image-caption, interleaved image-text, and text-only data is crucial.
2. the image encoder together with image resolution and the image token count has substantial impact, while the vision-language connector design is of comparatively negligible importance.

LLaVA-1.5

(1) Scaling to high-resolution image inputs.

(2) Compositional capabilities.

(3) Data efficiency: randomly downsampling

(4) Data scaling.

InternVL

(1) Strong Vision Encoder: we explored a continuous learning strategy for the large-scale vision foundation model—InternViT-6B, boosting its visual understanding capabilities, and making it can be transferred and reused in different LLMs.

(2) Dynamic HighResolution: we divide images into tiles ranging from 1 to 40 of 448×448 pixels according to the aspect ratio and resolution of the input images, which supports up to 4K resolution input.

(3) High-Quality Bilingual Dataset: we carefully collected a high-quality bilingual dataset that covers common scenes, document images, and annotated them with English and Chinese question-answer pairs, significantly enhancing performance in OCR- and Chinese-related tasks.

Recommendation

When building Vector Databases, shall we also consider the vectors for multimodal data such as images particularly (scanned or photo) documents?

Further resources

Youtube Video:

- Mastering Multimodal Models: Exploring Idefics2
- Multimodal Generative AI Demystified

Website:

- InternVL
- LLaVA
- TextMonkey
- Awesome-Multimodal-Large-Language-Models
- Awesome-Document-Understanding

Further materiel for this presentation

You might have watched a nice video a few weeks ago called "Multimodal Generative AI Demystified". The Multimodal models mentioned in that video includes many types of input and many types of output.

Note: With Document Understanding in consideration, this presentation will be focusing on multimodel on Image and Text. And it will also not involve any image generation.