Multimodal Large Language Model (MLLM)

Nasy

Jul 16, 2023

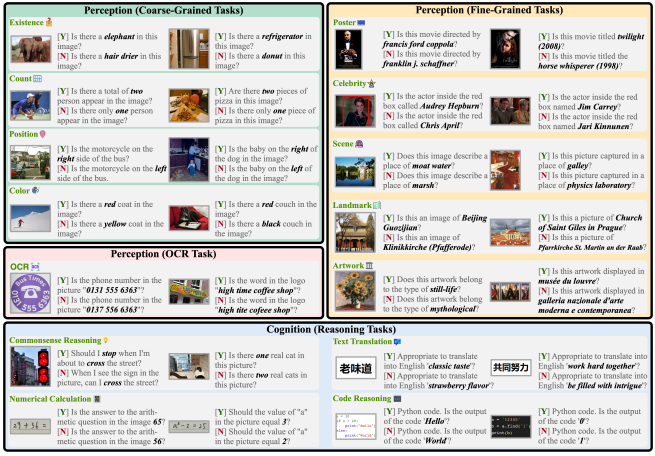
Outline

Introduction

- What is Multimodal Large Language Model (MLLM)?
  - LLM-based model with the ability to receive and reason with multimodal information.
- Future?
  - MLLM is more in line with the way humans perceive the world.
  - MLLM offers a more user-friendly interface.
- Examples
  - GPT4
  - LLaVA
  - MiniGPT-4
  - ...

Benchmark

MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models[1]



[1] Chaoyou Fu et al. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. July 1, 2023. arXiv: 2306.13394 [cs]. URL: http://arxiv.org/abs/2306.13394 (visited on 07/16/2023). preprint.

Overview

- Multimodal Instruction Tuning (M-IT)
- Multimodal In-Context Learning (M-ICL),
- Multimodal Chain-of-Thought (M-CoT),
- LLM-Aided Visual Reasoning (LAVR)

Multimodal Instruction Tuning (M-IT)

- Dataset:
  - Existing benchmark datasets
  - Self-instruction
- Model:
  - Align foreign embeddings to the LLMs
  - Resort to expert models to translate foreign modalities into natural languages that LLMs can ingest
- Fine-tuning:
  - LoRA

Multimodal Instruction Tuning (M-IT)

- Works:
  - MiniGPT-4[2]
  - Visual Instruction Tuning (LLaVA)[3]
  - mPLUG-Owl[4]

[2] Deyao Zhu et al. *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models.* Apr. 20, 2023. DOI: 10.48550/arXiv.2304.10592. arXiv: 2304.10592 [cs]. URL: http://arxiv.org/abs/2304.10592 (visited on 07/16/2023). preprint.

[3] Haotian Liu et al. *Visual Instruction Tuning.* Apr. 17, 2023. DOI: 10.48550/arXiv.2304.08485. arXiv: 2304.08485 [cs]. URL: http://arxiv.org/abs/2304.08485 (visited on 07/16/2023). preprint.

[4] Qinghao Ye et al. *mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality.* Apr. 27, 2023. DOI: 10.48550/arXiv.2304.14178. arXiv: 2304.14178 [cs]. URL: http://arxiv.org/abs/2304.14178 (visited on 07/16/2023). preprint.

MiniGPT-4

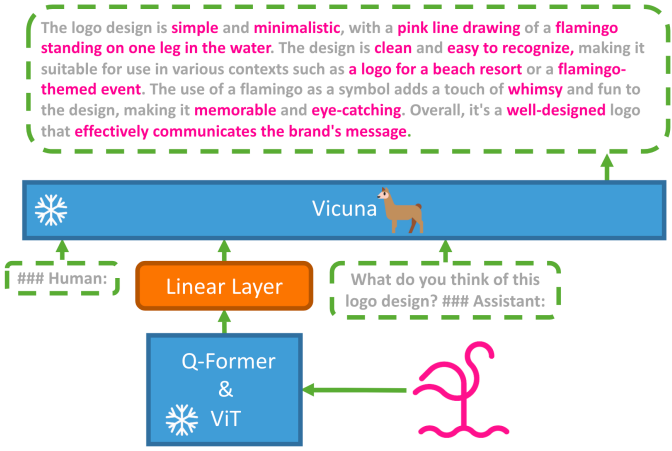


Figure 1: **The architecture of MiniGPT-4.** It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. MiniGPT-4 only requires training the linear projection layer to align the visual features with the Vicuna.

MiniGPT-4 methods

- Vicuna (LLaMA)
- BLIP-2 (ViT backbone with pre-trained Q-Former)
- Linear projection to bridge the gap
- Two stages
  - Pretraining the model on a large collection of aligned image-text pairs to acquire visionlanguage knowledge.
  - Fine-tune the pretrained model with a smaller but high-quality image-text dataset with a designed conversational template to enhance the model's generation reliability and usability.

MiniGPT-4 First stage

- Train
  - Conceptual Caption, SBU, and LAION
  - 20,000 training steps with a batch size of 256, covering approximately 5 million image-text pairs.
- Issues
  - Generating repetitive words or sentences, fragmented sentences, or irrelevant content.

MiniGPT-4 First stage alignment

- Two step
  - Align the image-text pairs generation (Template)
    - `###Human: <Img><ImageFeature></Img> Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:`
  - Data post processing (ChatGPT, fix the generated text)

MiniGPT-4 Final fine-tuning

- Finetune the pretrained model with the curated high-quality image-text pairs.
- `###Human: <Img><ImageFeature></Img> <Instruction> ###Assistant:`

LLaVA

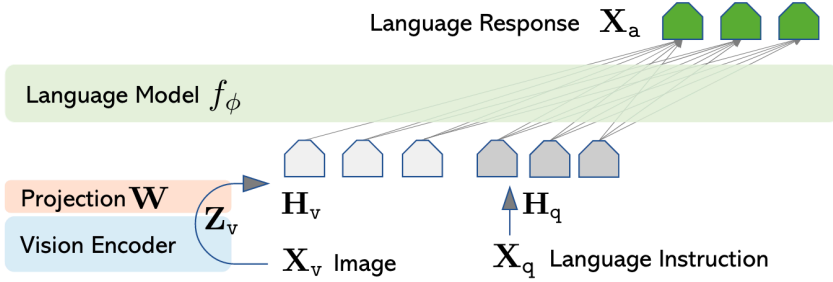


Figure 1: LLaVA network architecture.

Reference I

[1] Chaoyou Fu et al. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. July 1, 2023. arXiv: 2306.13394 [cs]. URL: http://arxiv.org/abs/2306.13394 (visited on 07/16/2023). preprint.

[2] Haotian Liu et al. *Visual Instruction Tuning*. Apr. 17, 2023. DOI: 10.48550/arXiv.2304.08485. arXiv: 2304.08485 [cs]. URL: http://arxiv.org/abs/2304.08485 (visited on 07/16/2023). preprint.

[3] Qinghao Ye et al. *mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality*. Apr. 27, 2023. DOI: 10.48550/arXiv.2304.14178. arXiv: 2304.14178 [cs]. URL: http://arxiv.org/abs/2304.14178 (visited on 07/16/2023). preprint.

[4] Deyao Zhu et al. *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models*. Apr. 20, 2023. DOI: 10.48550/arXiv.2304.10592. arXiv: 2304.10592 [cs]. URL: http://arxiv.org/abs/2304.10592 (visited on 07/16/2023). preprint.