

Multimodal Large Language Model (MLLM)

Nasy

Jul 16, 2023

Outline

1 Introduction

2 Multimodal Instruction Tuning (M-IT)

3 MiniGPT-4

4 Reference

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 Models¹

Perception (Coarse-Grained Tasks)		Perception (Fine-Grained Tasks)	
Existence 🌐	[Y] Is there a elephant in this image? [N] Is there a hair drier in this image?	[Y] Is there a refrigerator in this image? [N] Is there a donut in this image?	[Y] Is this movie directed by francis ford coppola ? [N] Is this movie directed by franklin j. schaffner ? 
Count 📊	[Y] Is there a total of two person appear in the image? [N] Is there only one person appear in the image?	[Y] Are there two pieces of pizza in this image? [N] Is there only one piece of pizza in this image?	[Y] Is this movie titled twilight (2008) ? [N] Is this movie titled the horse whisperer (1998) ? 
Position 💎	[Y] Is the motorcycle on the right side of the bus? [N] Is the motorcycle on the left side of the bus.	[Y] Is the baby on the right of the dog in the image? [N] Is the baby on the left of the dog in the image?	[Y] Is the actor inside the red box called Audrey Hepburn ? [N] Is the actor inside the red box called Chris April ? 
Color 🎨	[Y] Is there a red coat in the image? [N] Is there a yellow coat in the image?	[Y] Is there a red couch in the image? [N] Is there a black couch in the image?	[Y] Is the actor inside the red box named Jim Carrey ? [N] Is the actor inside the red box named Jari Kinnunen ? 
Perception (OCR Task)		Scene 🌱	
OCR 📄	[Y] Is the phone number in the picture " 0131 555 6363 "? [N] Is the phone number in the picture " 0137 556 6363 "?	[Y] Is the word in the logo " high time coffee shop "? [N] Is the word in the logo " high tide coffee shop "?	[Y] Does this image describe a place of moat water ? [N] Does this image describe a place of marsh ? 
			[Y] Is this picture captured in a place of galleys ? [N] Is this picture captured in a place of physics laboratory ? 
Landmark 🏛		Artwork 🎨	
		Landmark 🏛	[Y] Is this an image of Beijing Guozijian ? [N] Is this an image of Klinikkirche (Pfaffendorf) ? 
			[Y] Is this artwork displayed in the type of still-life ? [N] Is this artwork displayed in the type of mythological ? 
Cognition (Reasoning Tasks)			
Commonsense Reasoning 🧠			
Text Translation 🇮🇪	Code Reasoning 🛡		
老味道	[Y] Appropriate to translate into English classic taste ? [N] Appropriate to translate into English strawberry flavor ? 	共同努力	[Y] Appropriate to translate into English work hard together ? [N] Appropriate to translate into English be filled with intrigue ? 
Numerical Calculation 🗣	[Y] Should the value of "a" in the picture equal 3 ? [N] Should the value of "a" in the picture equal 2 ? 		[Y] Python code. Is the output of the code " Hello "? [N] Python code. Is the output of the code " World "? 

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-4²
- Visual Instruction Tuning (LLaVA)³
- mPLUG-Owl⁴

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

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

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

MiniGPT-4

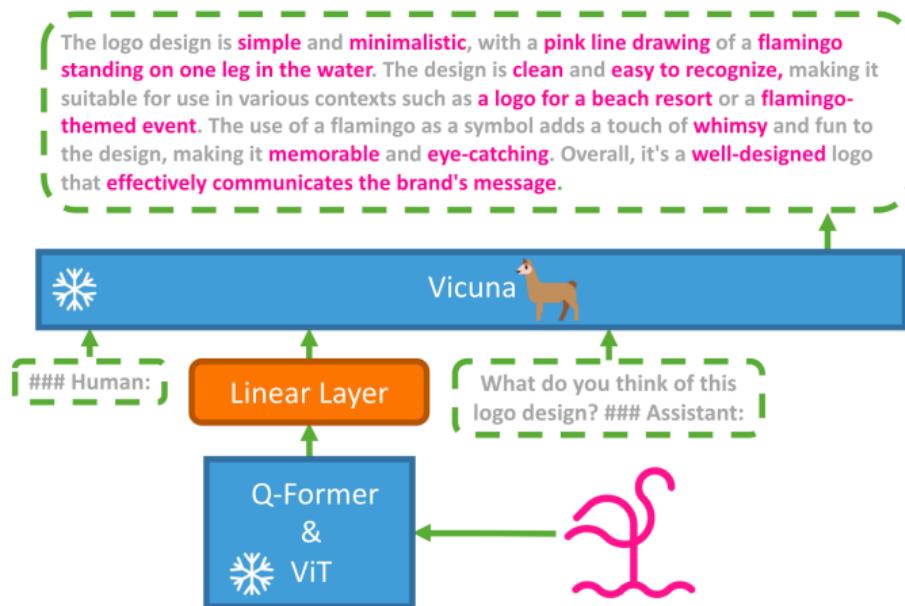


Figure 1: **The architecture of MiniGPT-4.** It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. MiniGPT-4 only requires training the linear projection layer to align the visual features with the Vicuna.

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: <ImageFeature> 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: <ImageFeature> <Instruction> ###Assistant:**

LLaVA

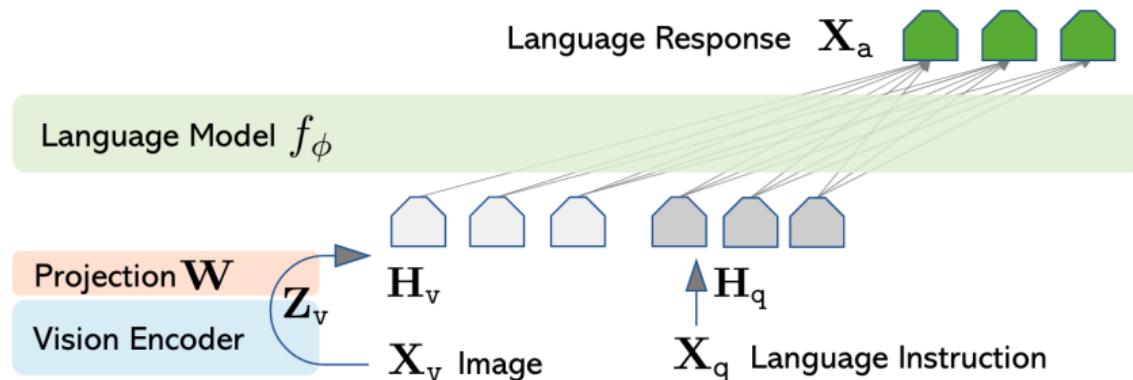


Figure 1: LLaVA network architecture.

Reference I

- [1] Chaoyou Fu et al. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. July 1, 2023. arXiv: [2306.13394 \[cs\]](https://arxiv.org/abs/2306.13394). 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](https://doi.org/10.48550/arXiv.2304.08485). arXiv: [2304.08485 \[cs\]](https://arxiv.org/abs/2304.08485). 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](https://doi.org/10.48550/arXiv.2304.14178). arXiv: [2304.14178 \[cs\]](https://arxiv.org/abs/2304.14178). 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](https://doi.org/10.48550/arXiv.2304.10592). arXiv: [2304.10592 \[cs\]](https://arxiv.org/abs/2304.10592). URL: <http://arxiv.org/abs/2304.10592> (visited on 07/16/2023). preprint.