Multimodal Large Language Models (MLLMs) transforming Computer Vision

This article introduces what is a Multimodal Large Language Model (MLLM), their applications using challenging prompts, and the top models reshaping Computer Vision.

1. What is a Multimodal Large Language Model (MLLM)?

An MLLM merges the reasoning capabilities of Large Language Models (LLMs) with the ability to receive, reason, and output with multimodal information. This includes text, images, audio, and video.

1.1 The rise of multimodality in Artificial Intelligence

The adoption of the Transformer architecture, invented by Google in 2017, has significantly impacted Computer Vision, leading to models like Vision Transformers (ViT). With the rise of LLMs, MLLMs naturally emerged.

1.2 MLLMs vs VLMs vs Foundation Models

- **MLLMs**: Can work with more modalities (text, images, audio, video) and have stronger reasoning skills.
- **VLMs (Vision Language Models)**: A specialized category that integrates text and image inputs and generates text outputs. They are less performant in reasoning skills compared to MLLMs.
- **Foundation Models**: Some consider MLLMs to be Foundation Models, such as Google’s Vertex AI which lists Claude 3, PaliGemma, or Gemini 1.5 as Foundation Models.

1.2 Architecture

The architecture of an MLLM is divided into three parts:

- **Modality encoder**: Condenses raw data formats (visuals, sound) into a streamlined representation. Often utilizes a pre-trained encoder (e.g., CLIP).
- **LLM backbone**: A language model that acts as the “brain” of the MLLM, processing features generated by the encoder and outputting responses in text.
- **Modality interface (connector)**: An intermediary link between the encoder and the LLM, crucial for connecting text with other modalities effectively since LLMs only interpret text.

2. Applications and use cases of Multimodal Models in Computer Vision

The article tests three top MLLMs (GPT-4o, LLaVA 7b, and Apple Ferret 7b) using challenging queries:

2.1 Counting objects in presence of occlusion

- **GPT-4o**: Yielded incorrect coordinates for bounding boxes.
- **LLaVA**: Incapable of detecting all objects and provided wrong bounding box locations.
- **Apple Ferret**: Successfully detected all objects, even occluded ones, demonstrating strong spatial understanding.

2.2 Autonomous driving: understanding and planning for risk

- **LLaVA**: Performed poorly, hallucinating and not identifying objects correctly.
- **GPT-4o**: Shines in returning reasoned detailed responses in text but performs poorly in clearly detecting bounding boxes.
- **Apple Ferret**: Detected the majority of objects with accurate bounding box coordinates, performing better than GPT-4o in this specific task.

2.3 Sports analytics: detecting objects and scene understanding

(Content truncated in the original extract, but the general idea is to test MLLMs in complex scenarios like sports analytics where unimodal fine-tuned architectures like YOLO typically excel.)

Key Takeaway: Apple's Ferret model showed strong performance in tasks requiring precise object detection and spatial understanding, even with occlusions, outperforming GPT-4o and LLaVA in these specific computer vision challenges.

Introduction to Vision Language Models (VLMs)

VLMs are AI systems that seamlessly combine image understanding with natural language processing. They connect what they see with the words that describe it, allowing machines to “see” and “read” at the same time.

How are VLMs different from traditional language models?

Feature	Normal Large Language Model (LLM)	Vision Language Model (VLM)
Primary Input	Text only	Images and Text
Knowledge Base	Derived from text corpora	Grounded in both text and visual data
Core Ability	Text-based reasoning and generation	Multimodal reasoning across vision and text
Example Task	“Write a poem about the ocean.”	“Describe what is happening in this picture of a beach.”

VLM Architecture

VLMs combine image and text processing into a unified framework, integrating modules that extract and align visual and textual features. Key components include:

- **Image Encoder**: Extracts features from images using a Vision Transformer (ViT).
- **Vision–Language Projector**: Aligns image embeddings with text embeddings.
- **Tokenizer + Embedding Layer**: Converts input text into token IDs and maps them to dense vectors.
- **Positional Encoding**: Adds spatial or sequential information to embeddings.
- **Shared Embedding Space**: Combines projected image tokens with text embeddings into a unified sequence.
- **Decoder-Only Language Model**: Generates output text autoregressively.

Datasets for Vision Language Models

High-quality training data for VLMs requires aligned multimodal data (images paired with text). Some key datasets include:

- **LAION-5B**: Massive dataset with over 5 billion image-text pairs.
- **PMD (Public Model Dataset)**: Contains 70 billion image-text pairs.
- **VQA (Visual Question Answering)**: Used to fine-tune VLMs for visual reasoning tasks.
- **ImageNet**: Primarily used for image classification and object recognition.

Evolution of VLMs

- **Pioneering Models (e.g., CLIP)**: OpenAI’s CLIP (Contrastive Language-Image Pre-training) learned to determine how well a text description matched an image, creating a powerful shared embedding space.
- **Generative Models (e.g., Flamingo, LLaVA)**: Models like Google’s Flamingo and LLaVA (Large Language and Vision Assistant) built upon CLIP, enabling complex, conversational dialogues about images.

Applications of VLMs

- **Image Captioning**: Automatically generating descriptive captions for images.
- **Visual Question Answering (VQA)**: Answering specific questions about an image.
- **Image-Text Retrieval**: Revolutionizing search by allowing image and text queries.

Challenges & Limitations of Vision-Language Models (VLMs)

- **Model Complexity**: Harder to train and deploy.
- **Bias**: Can inherit biases from training data.
- **Limited Understanding**: Rely on pattern recognition rather than reasoning.
- **Hallucinations**: May generate incorrect answers confidently.
- **Generalization**: Struggle with unseen data.
- **High Computational Cost**: Require substantial resources.
- **Ethical Concerns**: Data collection without consent raises issues.

Multimodal AI: A Guide to Open-Source Vision Language Models

This article focuses on open-source Vision Language Models (VLMs), which are designed to understand and process both visual and text information.

Notable Open-Source VLMs:

Gemma 3

- Developed by Google, built on the same research behind Gemini 2.0.
- Supports advanced text, image, and short video understanding.
- Available in 1B, 4B, 12B, and 27B sizes, offering flexibility.

- Key features: Multilingual support (140+ languages), portable and efficient, supports agentic workflows.
- Cautions: Limited video comprehension (especially long-form or audio-visual), image input only normalized (may limit fine-grained understanding).

Llama 3.2 Vision

- Developed by Meta, designed to process both text and images.
- Available in 11B and 90B parameter sizes, outperforming many open-source and proprietary models in image-text tasks.
- Integrates a pre-trained image encoder into the language model using adapters.
- Key features: Multimodal capabilities (captioning, image-based Q&A, visual reasoning), strong performance, customizability.
- Cautions: Room for improvement in math-heavy tasks, English only for image+text applications.

NVLM 1.0

- Family of multimodal LLMs developed by NVIDIA, achieving state-of-the-art results.
- Rivals top proprietary models like GPT-4o and open-access models like Llama 3-V 405B.
- Consists of three distinct architectures:
    - **NVLM-D**: Decoder-only, unified multimodal reasoning, better at OCR.
    - **NVLM-X**: Cross-attention-based, computationally efficient for high-resolution images.
    - **NVLM-H**: Hybrid, superior performance in multimodal reasoning and image processing.
- Key features: Powerful image reasoning, improved text-only performance.

- Cautions: Non-commercial use only (research and hobbyist), limited variant release (only NVLM-1.0-D-72B publicly available).

Molmo

- Family of open-source VLMs developed by the Allen Institute for AI.
- Available in 1B, 7B, and 72B parameters, performing on par with proprietary models like GPT-4V, Gemini 1.5 Pro, and Claude 3.5 Sonnet.
- Key to performance: Unique training data, **PixMo**, consisting of 1 million image-text pairs (dense captioning and supervised fine-tuning data).
- Innovative data collection: Annotators provided spoken descriptions of images, capturing detailed spatial positioning and relationships among objects efficiently.