Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) are advanced AI systems that can understand and generate information across **multiple types of data**, not just text.

In simple words:
LLMs = models that understand text
MLLMs = models that understand text + images + audio + video + other modalities

What "Multimodal" Means……

Modality = type of data.
MLLMs can process more than one modality, such as:

Text (questions, paragraphs, code)

Images (photos, diagrams, charts)

Audio (speech, music)

Video (clips, scenes)

Sensor data (e.g., from robots)

Applications of Multimodal LLMs…….

1. **Image Understanding**
   – Captioning, object detection, diagram solving
2. **Education**
   – Solve math problems from photos
   – Check handwritten answers
3. **Healthcare**
   – Analyse X-rays with explanations
4. **Finance**
   – Analyse charts + news + text
5. **Robotics**
   – Navigate environments using sensor + image I/O
6. **Creative tasks**
   – Generate images, logos, storyboards

Examples of Popular MLLMs……

GPT and GPT-5 family

Google Gemini

Meta LLaVA / LLaMA Vision

OpenAI CLIP + LLM hybrids

DeepMind Flamingo

PaLM-E (robotics multimodal model)

How MLLMs Work……..

Multimodal LLMs contain:

1. **Encoders**
   Convert each data type into embeddings (numerical representations).
   - Image encoder
   - Audio encoder
   - Video encoder
   - Text encoder (transformer)
2. **Fusion Module**
   Combines information across modalities.
3. **LLM Core**
   A transformer-based model that performs reasoning and generation.
4. **Decoders**
   Convert embeddings back into text, images, or audio.

Key Features of Multimodal LLMs……..

1. Understand multiple data types

Example:
Upload an image → ask questions about it → get answers in text.

2. Combine information across modalities

Example:
Provide a chart + a question → model interprets the chart & provides insights.

3. Generate multimodal outputs

- Generate images from text prompts
- Generate captions for images
- Convert speech to text or text to speech

4. Better reasoning

Because they see multiple forms of information, they can perform:

- Visual reasoning
- Mathematical reasoning with diagrams
- Chart reading
- Code understanding with screenshots
- Spatial or geometric reasoning

When you think about large language models (LLMs), multimodality might not be the first thing that comes to mind. After all, they are *language* models! But we can quickly see that models can be much more useful if they're able to handle types of data other than text. It's very useful, for example, if a language model is able to glance at a picture and answer questions about it. A model that is able to handle text and images (each of which is called a *modality*) is said to be *multimodal*, as we can see in Figure 9-1.

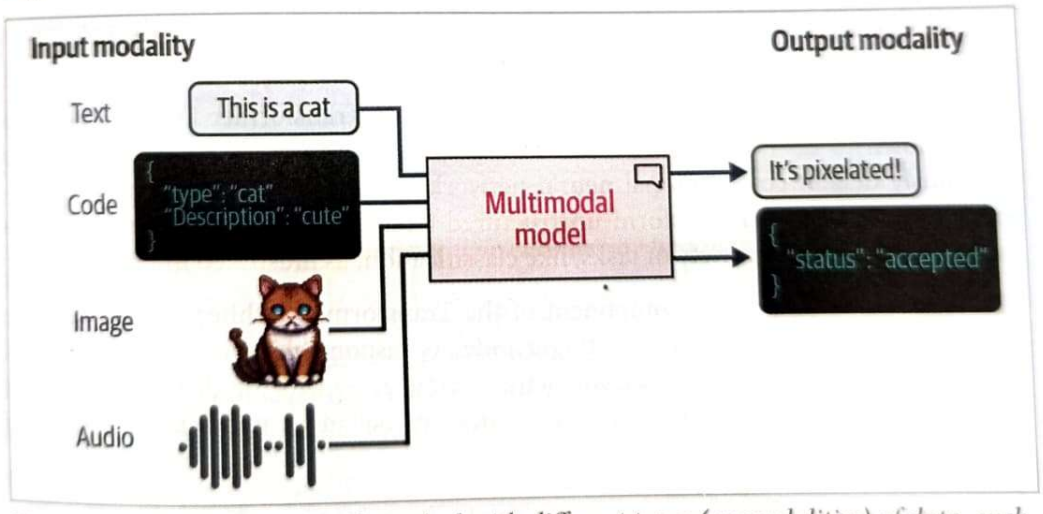

Figure 9-1. Models that are able to deal with different types (or modalities) of data, such as images, audio, video, or sensors, are said to be multimodal. It's possible for a model to accept a modality as input yet not be able to generate in that modality.

Transformers for vision……

Transformers for **vision** in LLMs are models that apply the **Transformer architecture** originally developed for text to process **images**. This capability is a key reason modern **Multimodal Large Language Models (MLLMs)** (like GPT, Gemini etc.) can understand pictures, charts, diagrams, and videos.

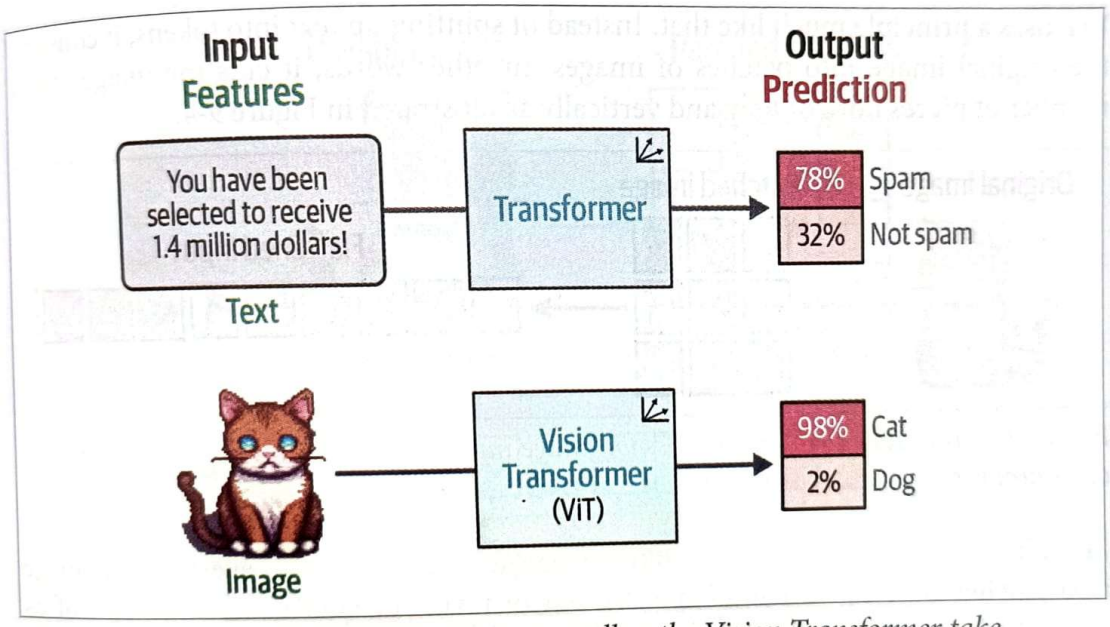

Figure 9-2. Both the original Transformer as well as the Vision Transformer take unstructured data, convert it to numerical representations, and finally use that for tasks like classification.

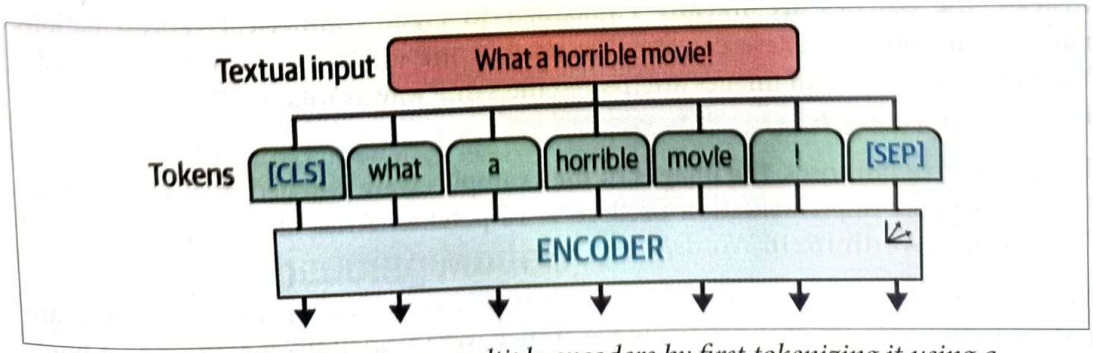

Figure 9-3. Text is passed to one or multiple encoders by first tokenizing it using a tokenizer.

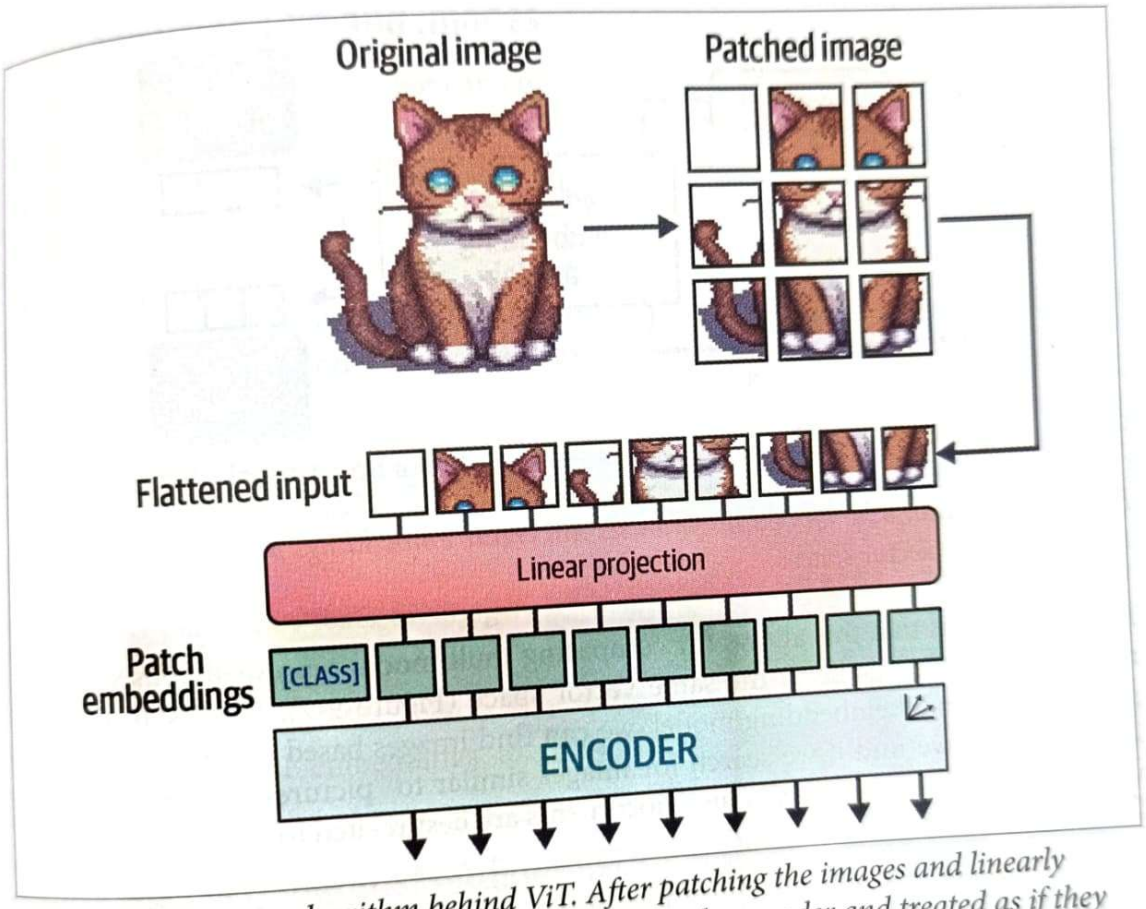

Figure 9-5. The main algorithm behind ViT. After patching the images and linearly projecting them, the patch embeddings are passed to the encoder and treated as if they were textual tokens.

Contrastive Language Image Pre-training……

CLIP: Connecting Text and Images

CLIP is an embedding model that can compute embeddings of both images and texts. The resulting embeddings lie in the same vector space, which means that the embeddings of images can be compared with the embeddings of text.[3] This comparison capability makes CLIP, and similar models, usable for tasks such as:

Zero-shot classification
: We can compare the embedding of an image with that of the description of its possible classes to find which class is most similar.

Clustering
: Cluster both images and a collection of keywords to find which keywords belong to which sets of images.

Search
: Across billions of texts or images, we can quickly find what relates to an input text or image.

Generation
: Use multimodal embeddings to drive the generation of images (e.g., stable diffusion[4]).

How Can CLIP Generate Multimodal Embeddings?

The procedure of CLIP is actually quite straightforward. Imagine that you have a dataset with millions of images alongside captions as we illustrate in Figure 9-8.

Figure 9-8. The type of data that is needed to train a multimodal embedding model.

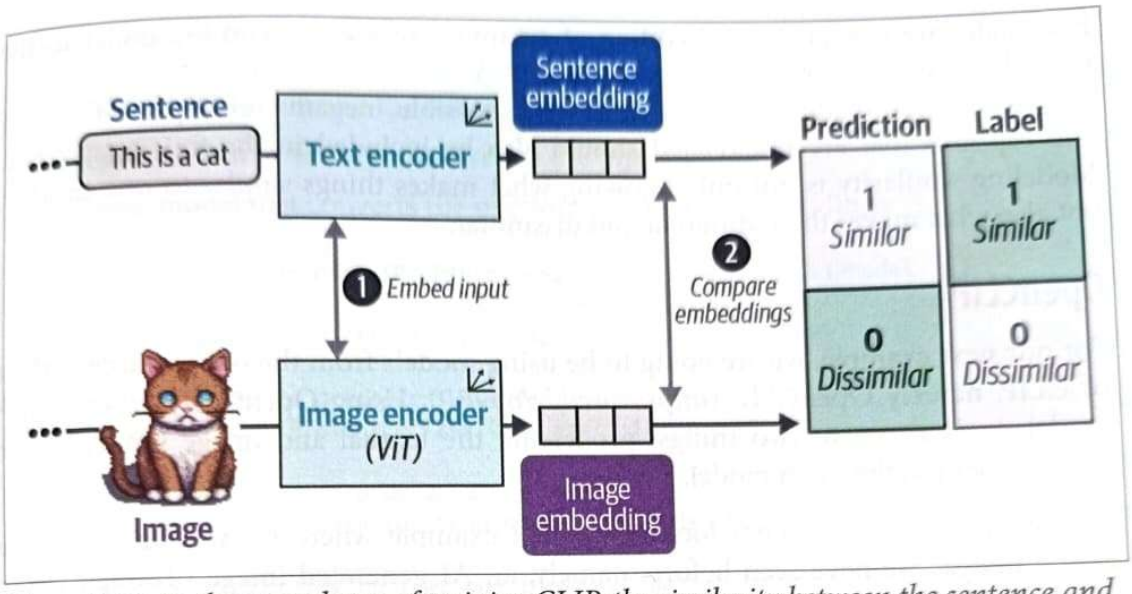

Figure 9-10. In the second step of training CLIP, the similarity between the sentence and image embedding is calculated using cosine similarity.

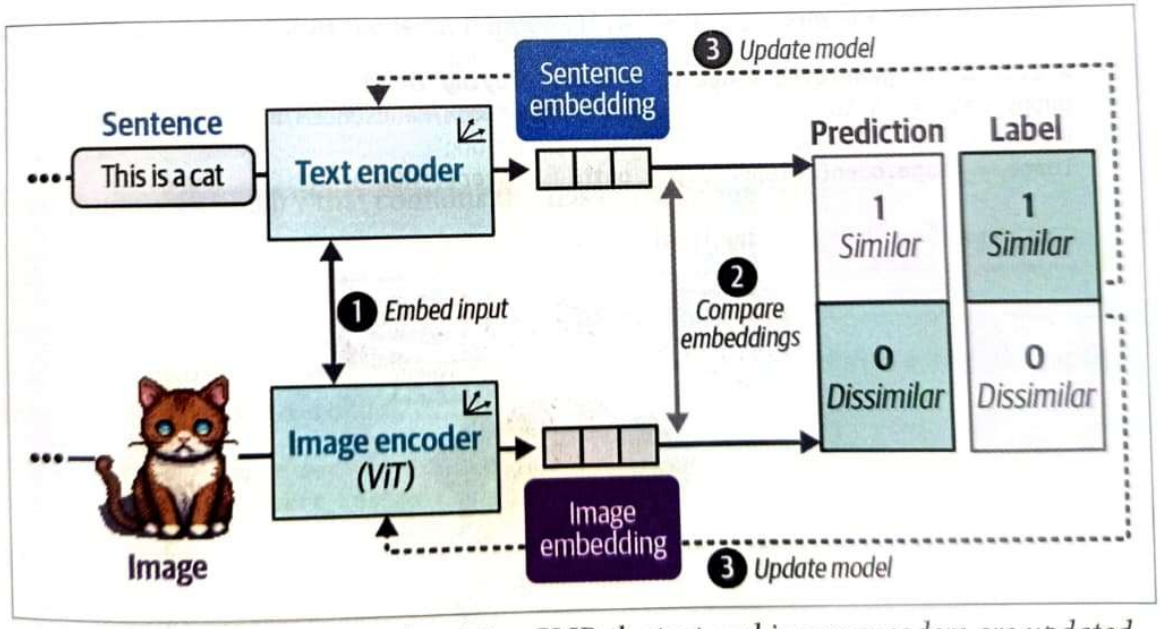

Figure 9-11. In the third step of training CLIP, the text and image encoders are updated to match what the intended similarity should be. This updates the embeddings such that they are closer in vector space if the inputs are similar.

Making Text Generation Models Multimodal……

Making Text Generation Models Multimodal taking a text-only LLM (like GPT, BERT, LLaMA, etc.) and adding the ability to understand and generate other types of data, such as:

- **Images**
- **Audio**
- **Video**
- **Charts**
- **Sensor data**
- **Embedding streams**

This process transforms an LLM into a **Multimodal Large Language Model (MLLM)**.

Why Do We Need Multimodality?

Because **real-world information is not only text**.

Examples:

- A student uploads a math diagram → LLM should solve it.
- A doctor uploads an X-ray → LLM should diagnose it.
- A trader uploads a candlestick chart → LLM should interpret it.
- A user uploads an audio clip → LLM should transcribe and explain it.

A text-only LLM cannot do these tasks.
A multimodal one can.

In the case of text generation models, we would like it to reason about certain input images. For example, we could give it an image of a pizza and ask it what ingredients it contains. You could show it a picture of the Eiffel Tower and ask when it was built or where it is located. This conversational ability is further illustrated in Figure 9-15.

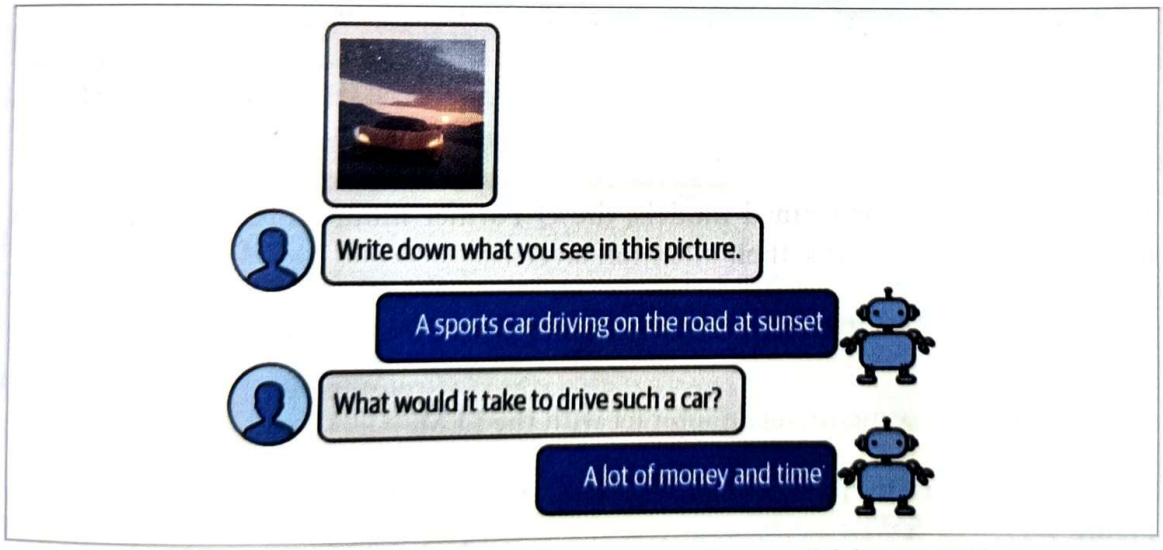

Figure 9-15. An example of a multimodal text generation model (BLIP-2) that can reason about input images.

To bridge the gap between these two domains, attempts have been made to introduce a form of multimodality to existing models. One such method is called *BLIP-2: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation 2*. BLIP-2 is an easy-to-use and modular technique that allows for introducing vision capabilities to existing language models.

Bootstrapping Language-Image Pre-training

It is a **vision-language model (VLM)** designed to connect:

- **Images**
- **Text**

BLIP is mainly used as the **image–text encoder** in many **multimodal LLM systems**.

How BLIP Works .....

BLIP consists of three components:

1. Vision Encoder (ViT)

Converts images into embeddings (like “image tokens”).

2. Text Encoder / Decoder

Reads or generates text.

3. Multimodal Fusion Module

Aligns image and text representations for joint understanding.

BLIP is trained on massive image–text datasets, allowing it to learn:

- visual content
- captions
- relationships between text and images

Benefits of BLIP

1. Strong Image–Text Alignment

BLIP aligns the meaning of text and images very accurately, improving:

Captioning
Question answering
Image-based reasoning

This alignment is crucial for multimodal LLMs.

2. High-Quality Image Captioning

3. BLIP can be used in different modes:

Image-to-text
Text-to-image
Vision–text matching

4. Reduces Noisy Data During Training

BLIP has a **caption-filtering mechanism**, meaning:

It removes low-quality captions

Keeps clean and relevant pairs

This leads to **better training** and more robust multimodal understanding.

BLIP is a vision-language model that helps LLMs understand images, align them with text, and perform tasks like captioning and visual Q&A, making multimodal AI more accurate and more powerful.