SEMINAR
Vision-Language Models Survey

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology
Vietnam National University Ho Chi Minh City

Table of Contents

Table of Contents

Background

With the advent of deep learning, visual recognition research has achieved great success by leveraging end-to-end trainable deep neural networks (DNNs). However, the shift from traditional machine learning to deep learning comes with two new grand challenges:

- The slow convergence of DNN training under the classical setup of Deep Learning from scratch
- The laborious collection of large-scale, task specific, and crowd-labelled data in DNN training.

Background

- Recently, a new learning paradigm Pre-training, Finetuning and Prediction has demonstrated great effectiveness in a wide range of visual recognition tasks.
- Inspired by the advances in natural language processing, a new deep learning paradigm named Vision-Language Model Pre-training and Zero-shot Prediction has appeared.

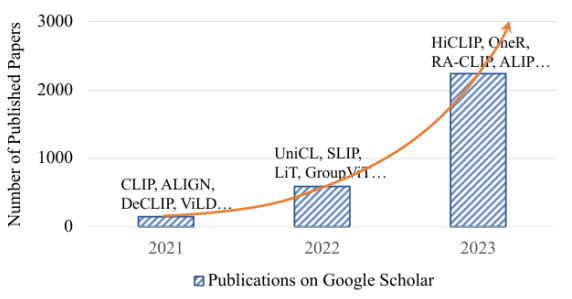


Figure 1: Number of publications on visual recognition VLMs (from Google Scholar)

Table of Contents

Introduction

Definition

VLMs learn to map the relationships between text data and visual data such as images or videos, allowing these models to generate text from visual input or understand natural language commands in the context of visual information[1].

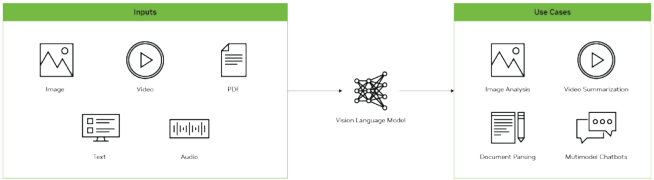


Figure 2: Visualization of a Vision-Language Model

[1] https://www.ibm.com/think/topics/vision-language-models

Introduction

Core Capabilities

- Unlike traditional computer vision models, VLMs are not bound by a fixed set of classes or a specific task such as classification or detection.
- Retrained on a vast corpus of text and image / video caption pairs, VLMs can be trained in natural language and used to handle many classic vision tasks plus new AI-powered generative tasks such as summarization and visual question answering.

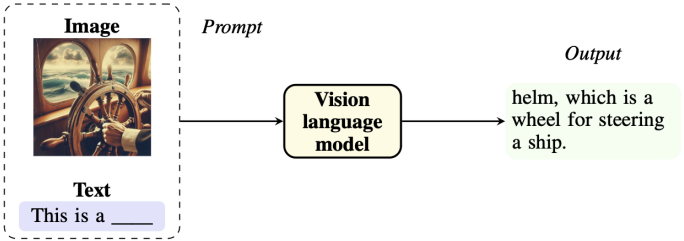


Figure 3: Illustrations of Visual Question Answering

Introduction

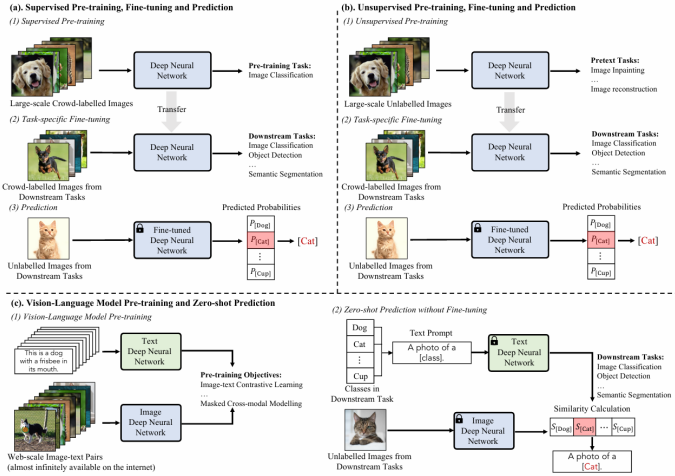


Figure 4: Three DNN training paradigms in visual recognition.

Introduction

1. Traditional Machine Learning and Prediction
Hand-crafted features, poor scalability.

↓

2. Deep Learning from Scratch and Prediction
Complicated feature engineering, slow convergence of DNN, laborious collection of large-scale, task-specific, and crowd-labelled data.

↓

3. Supervised Pre-training, Fine-tuning and Prediction
Accelerate network convergence, well-performing models with limited task-specific training data, requires large-scale labelled data in pre-training.

↓

4. Unsupervised Pre-training, Fine-tuning & Prediction
Self-supervised learning to learn useful and transferable representations from unlabelled data, e.g., masked image modelling, contrastive learning; still requires a fine-tuning stage.

↓

5. VLM Pre-training and Zero-shot Prediction
Motivated by great success in NLP; matches image-text embeddings; leverages large-scale image-text pairs available online.

Table of Contents

Vision-language tasks and variants

Early VLMs focus on image-level visual recognition tasks, whereas recent VLMs are more general-purpose, which can also work for dense prediction tasks that are complex and require localization related knowledge.

Table 1: Core Vision Tasks

Task Type	Description	Strong VLM Models
Image Classification	Predicts a label for the whole image (e.g., bear, dog, car...)	**CLIP**, **BLIP**, **ALIGN**, **Florence**
Object Detection	Detects multiple objects with bounding boxes	**Grounding DINO**, **OWL-ViT**, **GLIP**
Semantic Segmentation	Labels each pixel by class (not separating instances)	**CLIPSeg**, **LSeg**
Instance Segmentation	Labels and distinguishes each object instance individually	**SAM**, **SEEM**, **GRIT**

Table 2: Advanced Multimodal Tasks

Task Type	Description	Strong VLM Models
Image Captioning	Generates a text description for the image	**BLIP**, **BLIP-2**, **GIT**, **MiniGPT-4**
Visual Question Answering (VQA)	Answers questions based on the image content	**BLIP-2**, **Flamingo**, **MiniGPT-4**, **LLaVA**
Visual Grounding	Locates regions in the image based on a textual description	**Grounding DINO**, **OWL-ViT**, **GLIP**
Text-to-Image Retrieval	Finds matching images from a text query (or vice versa)	**CLIP**, **ALIGN**, **Florence**
Multimodal Reasoning	Performs reasoning using both image and language modalities	**GPT-4V**, **Kosmos-2**, **MiniGPT-4**, **LLaVA**

VLMs on vision tasks

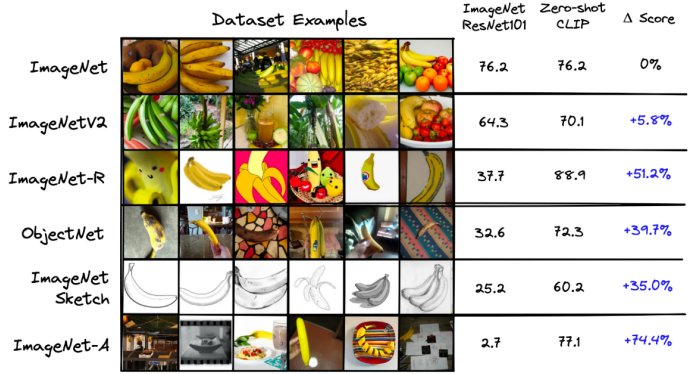


Figure 5: ResNet-101 fine-tuned on ImageNet vs. zero-shot CLIP

VLMs on vision tasks

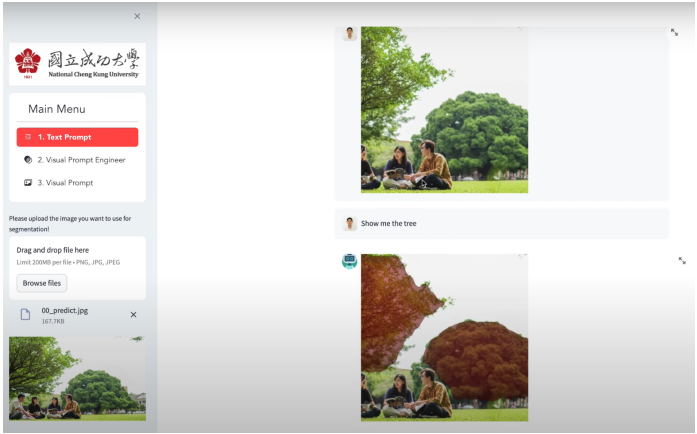


Figure 6: Basic processing systems apply CLIPSeg for Object Segmentation tasks

VLMs on vision tasks

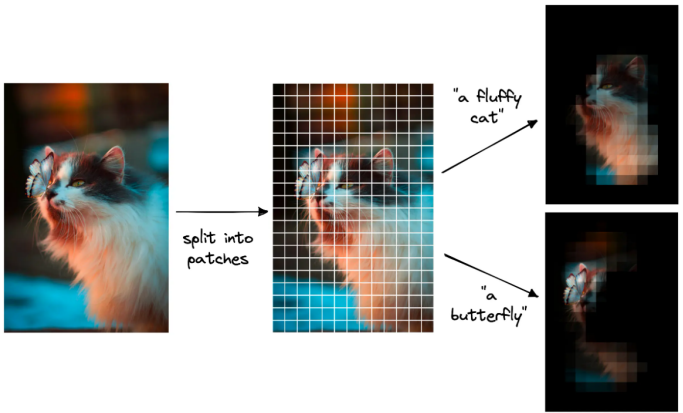


Figure 7: CLIP in object detection and localization

Table of Contents

VLM Pre-training Frameworks

Three curent approached frameworks:

- Two-Tower VLM
- Two-Leg VLM
- One-Tower VLM

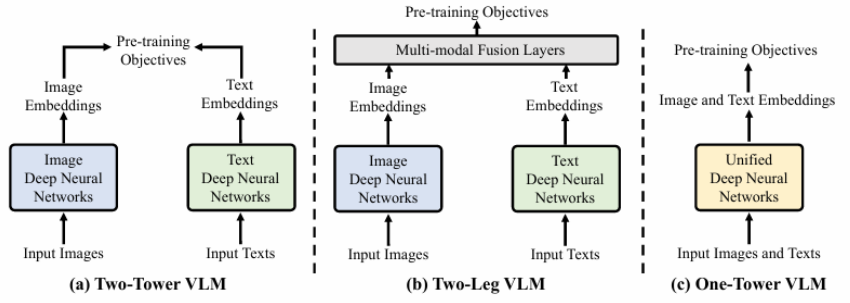


Figure 8: three current architectures of VLM

CNN-based Architectures for Image Features

CNN-based Architectures:

- Different ConvNets such as VGG, ResNet, and EfficientNet have been widely used for learning image features.
- These models rely on convolutional layers to capture spatial hierarchies in images.

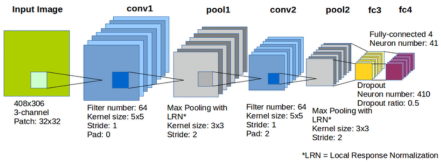


Figure 9: Convolutional architecture

Transformer-based Architectures for Image Features

Vision Transformer (ViT):

- ViT applies Transformer blocks to image patches instead of word tokens.
- Each block includes multi-head self-attention and a feed-forward network.
- ViT has shown competitive or superior performance compared to CNNs on many visual tasks.

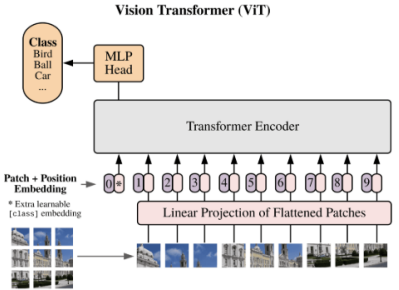
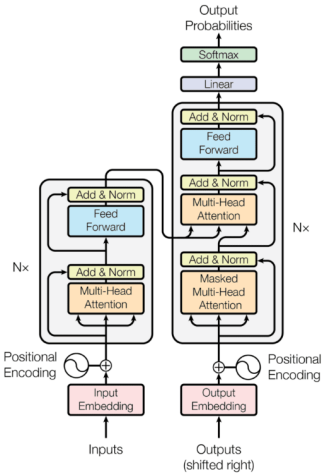


Figure 10: Vision transformer architecture

Learning Language Features

Transformer-based Architectures for Language:

- Transformer and its variants are widely used to learn text representations.



VLM Pre-training Objectives

VLM Pre-training Objectives As the core of VLM, various vision language pre-training objectives have been designed to learn a rich vision-language correlation.

Three types of contrastive learning:

- Contrastive objectives
- Generative objectives
- Alignment objectives

Contrastive Objectives

Contrastive Objectives: Learn discriminative representations by pulling paired samples close and pushing others away.

Three types of contrastive learning:

- Image Contrastive Learning
- Image-Text Contrastive Learning
- Image-Text-Label Contrastive Learning

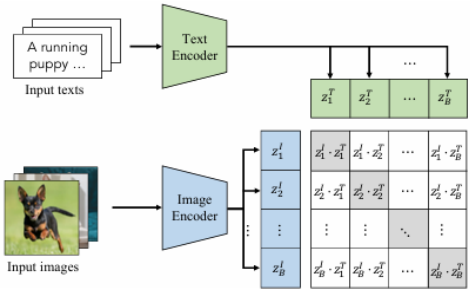
Image Contrastive Learning

Given a batch of B images, image contrastive learning (e.g., InfoNCE) aims to bring positive pairs (z_i^I, z_i^{I+}) close and push negative keys $z_j^I, j \neq i$ away.

Loss Formulation

$$\mathcal{L}_I^{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(z_i^I \cdot z_{i+}^I/\tau)}{\sum_{j=1,j\neq i}^{B+1} \exp(z_i^I \cdot z_j^I/\tau)}$$

Image-text Contrastive Learnings



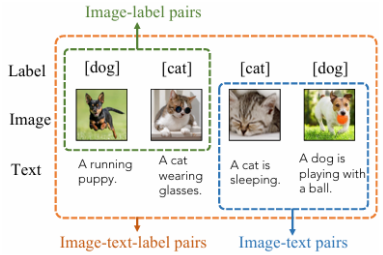
pulling the embeddings of paired images and texts close while pushing others

Loss Formulation

$$\mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_i^T/\tau\right)}{\sum_{j=1}^{B}\exp\left(z_i^I \cdot z_j^T/\tau\right)}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^T \cdot z_i^I/\tau\right)}{\sum_{j=1}^{B}\exp\left(z_i^T \cdot z_j^I/\tau\right)}$$

Image-Text-Label Contrastive Learning



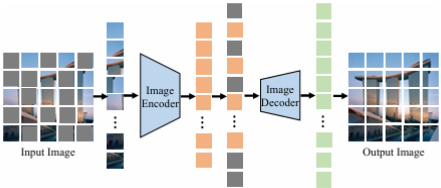
Introduces supervised contrastive learning into image-text contrastive learning by grouping samples with the same label as positives.

Loss Formulation

$$\mathcal{L}_{I\to T}^{\text{ITL}} = -\sum_{i=1}^{B} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp(z_i^I \cdot z_k^T/\tau)}{\sum_{j=1}^{B} \exp(z_i^I \cdot z_j^T/\tau)}$$

$$\mathcal{L}_{T\to I}^{\text{ITL}} = -\sum_{i=1}^{B} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp(z_i^T \cdot z_k^I/\tau)}{\sum_{j=1}^{B} \exp(z_i^T \cdot z_j^I/\tau)}$$

Masked Image Modeling - Generative



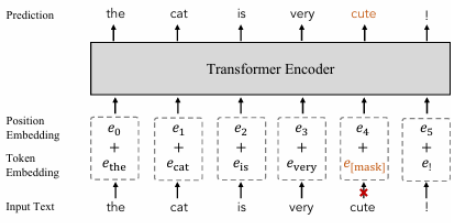
Randomly masks a subset of image patches and trains the model to reconstruct them from the unmasked patches.

Loss Formulation

$$\mathcal{L}_{\text{MIM}} = -\frac{1}{B}\sum_{i=1}^{B} \log f_\theta(x_i^I \mid \hat{x}_i^I)$$

where x_i^I is the masked patch set and $\hat{x}_i^I$ is the unmasked patch set of image x_i^I.

Masked Language Modeling - Generative



Randomly masks a subset of text tokens and reconstructs them based on the unmasked tokens.

Loss Formulation

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{B}\sum_{i=1}^{B} \log f_\phi(x_i^T \mid \hat{x}_i^T)$$

where x_i^T is the masked token set and $\hat{x}_i^T$ is the unmasked set.

Masked Cross-Modal Modeling - Generative

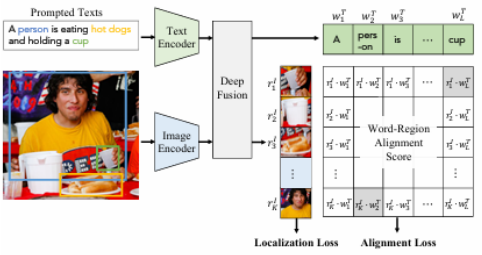
Combines MIM and MLM by jointly masking both image patches and text tokens, and reconstructing them conditioned on the unmasked parts.

Loss Formulation

$$\mathcal{L}_{\text{MCM}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log f_\theta(x_i^I \mid \hat{x}_i^I, \hat{x}_i^T) + \log f_\phi(x_i^T \mid \hat{x}_i^I, \hat{x}_i^T)\right]$$

where x_i^I, $\hat{x}_i^I$ are masked/unmasked image patches and x_i^T, $\hat{x}_i^T$ are masked/unmasked tokens.

Image-Text Matching - Alignment



Models global correlation between images and texts using a binary classification loss over the alignment score.

Loss Formulation

$$\mathcal{L}_{IT} = p \log S(z^I, z^T) + (1-p) \log(1 - S(z^I, z^T))$$

Region-Word Matching - Alignment

Models local cross-modal correlation between image regions and words for dense visual recognition tasks such as object detection.

Loss Formulation

$$\mathcal{L}_{RW} = p \log S_r(r^I, w^T) + (1-p) \log(1 - S_r(r^I, w^T))$$

Transfer Learning

Motivation:

- **Distribution gap:** Downstream tasks may differ in image styles and text formats.
- **Objective gap:** VLMs are trained with general objectives, while downstream tasks require task-specific objectives (e.g., classification, detection).

Transfer Techniques:

- **Prompt Tuning:** Modifies input text/image with learnable prompts. Includes:
  - Text Prompt Tuning (e.g., CoOp, CoCoOp, DualCoOp, PLOT)
  - Visual Prompt Tuning (e.g., VP, RePrompt)
  - Text-Visual Prompt Tuning (e.g., UPT, MAPLE)
- **Feature Adapter:** Adds lightweight trainable layers after VLM encoders (e.g., CLIP-Adapter, Tip-Adapter, SVL-Adapter)
- **Other Methods:**
  - Direct Fine-tuning (e.g., Wise-FT)
  - Architecture Modification (e.g., MaskCLIP)
  - Cross-modal Attention (e.g., VT-CLIP, CALIP)

Knowledge Distillation

Motivation:

- **Architecture Flexibility:** Distill VLM knowledge into task-specific models without retaining the VLM structure.
- **Representation Gap:** VLMs offer image-level features, while downstream tasks need region/pixel-level understanding.

For Object Detection:

- **Embedding Alignment:** Align detector and VLM features (e.g., ViLD, HierKD, RKD)
- **Prompt-based Distillation:** Learn detection-specific prompts (e.g., DetPro, PromptDet)
- **Pseudo-label Supervision:** Use VLM-generated pseudo boxes/masks (e.g., PB-OVD, XPM, P3OVD)
- **Region Bag Distillation:** Aggregate multiple region embeddings (e.g., BARON, RO-ViT)

Knowledge Distillation

For Semantic Segmentation:

- **Two-stage Pipeline:** Segment-then-classify approach (e.g., ZegFormer, ZSSeg)
- **Direct Pixel-level Distillation:** Match VLM with pixel-wise features (e.g., CLIPSeg, LSeg, MaskCLIP+)
- **Prompt/Descriptor Learning:** Enhance generalization beyond base classes (e.g., ZegCLIP, OVSeg)
- **Weak Supervision with VLM:** Refine CAMs using CLIP (e.g., CLIP-ES, CLIMS)

Goal: Transfer general VLM knowledge to dense prediction tasks while enabling open-vocabulary capability.

Table of Contents

What is SOTA?

SOTA = State of the Art

- Refers to the best-performing model/method for a specific task.
- Achieves top results on benchmarks (e.g., VQA, captioning).
- Represents cutting-edge R&D in AI.

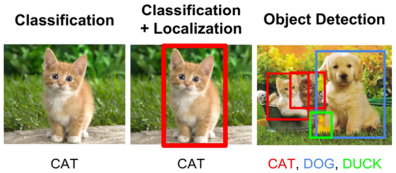
SOTA Models by Objectives

VLMs' Tasks from Basic to Advanced

- Core Vision Tasks (Basic Visual Recognition)
- Vision-Language Alignment (Zero-shot Learning)
- Captioning & OCR (Image Description and Reading)
- Visual Question Answering (Vision-Language Reasoning)
- Instruction Following & Visual Dialogue (Advanced Multimodal Interaction)

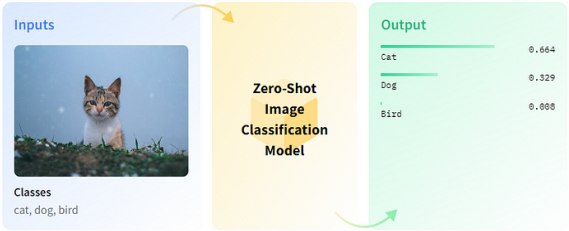
SOTA Models for Core Vision Tasks

Task	Description	SOTA Models
Image Classification	Assign a label to an image	ViT(Google, 2020)
Object Detection	Detect and classify objects in an image	YOLOv8, DETR
Image Segmentation	Segment specific objects or regions	SAM (Meta, 2023)



SOTA Models for Vision-Language Alignment

Task	Description	SOTA Models
Zero-Shot Classification	Classify images without fine-tuning	CLIP (OpenAI, 2021)
Multimodal Retrieval	Search for images based on text or vice versa	CLIP, BLIP-2

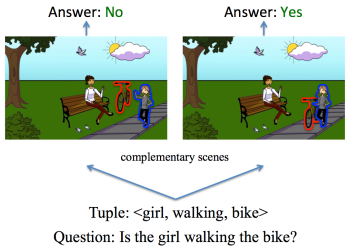


SOTA Models for Captioning & OCR

Task	Description	SOTA Models
Image Captioning	Generate textual descriptions for images	BLIP-2, Flamingo
OCR + Grounding	Recognize text and ground it to visual objects	Kosmos-2, GPT-4o

SOTA Models for Visual Question Answering

Task	Description	SOTA Models
VQA (Visual QA)	Answer natural language questions about images	Flamingo, BLIP-2
Visual Reasoning (Few-shot)	Logical reasoning on image-based questions	MiniGPT-4, GPT-4o



SOTA Models for Instruction Following & Visual Dialogue

Task	Description	SOTA Models
Visual Instruction Following	Perform tasks based on visual input + commands	GPT-4o, MiniGPT-4
Multimodal Chat/Agents	Real-time interaction with images + text + audio	GPT-4o, Gemini 1.5

SOTA Models for Instruction Following & Visual Dialogue

Figure 12: Multimodal Chat/Agents

Table of Contents

Real-World Products Using VLMs

Product	Use Case	Backed by
ChatGPT Vision	VQA, OCR, diagram understanding	OpenAI
Gemini Pro	Video understanding + logic	Google DeepMind
Claude 3	Charts, tables, multimodal dialogue	Anthropic
Perplexity AI	Visual search and info retrieval	VLM Hybrid (CLIP + BLIP)
Adobe Firefly	Prompt-based image generation	Adobe
Google Lens	Product & object recognition	ViT + CLIP
MidJourney	AI art and concept generation	Diffusion + vision guidance

CORE VLM COMPONENTS

Core VLM Components

OCR

- Optical Character Recognition
- A technology that enables machines to read and extract text from images or scanned documents.
- Use in VLMs: Allows models to understand and process text inside images, such as signs, forms, or screenshots.
- Used in Kosmos-2 and ChatGPT Vision to read charts, menus, or handwritten notes.

Figure 13: Optical Character Recognition

Core VLM Components

ViT

- ViT stands for Vision Transformer — a model architecture introduced by Google Research in 2020 that applies the transformer architecture (originally designed for NLP) directly to image data.
- Outperforms CNNs on many vision tasks and is used in many modern VLMs as the image encoder.
- Forms the backbone of models like CLIP, DINOv2, and Google Lens.

Core VLM Components

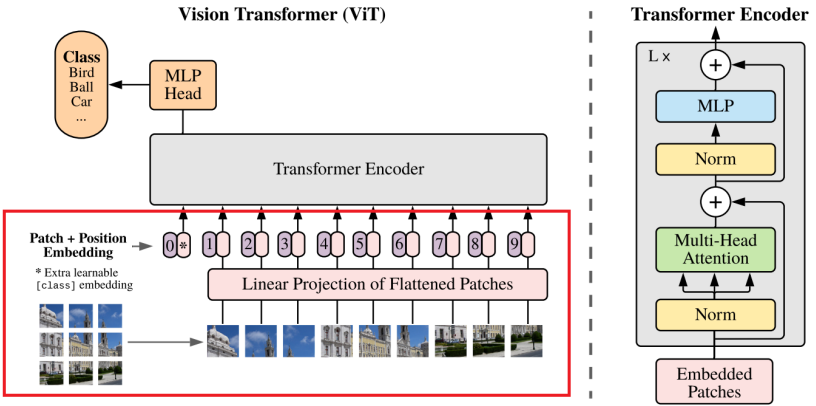


Figure 14: Vision Transformer (ViT) Architecture

Core VLM Components

Diffusion Models

- A class of generative models that create images from noise by learning how to reverse a "blurring" process step-by-step.
- Use in VLMs: Powers text-to-image generation tools by combining language understanding with image synthesis
- Used in DALL-E 2, Adobe Firefly, and MidJourney for prompt-based image creation.

Core VLM Components

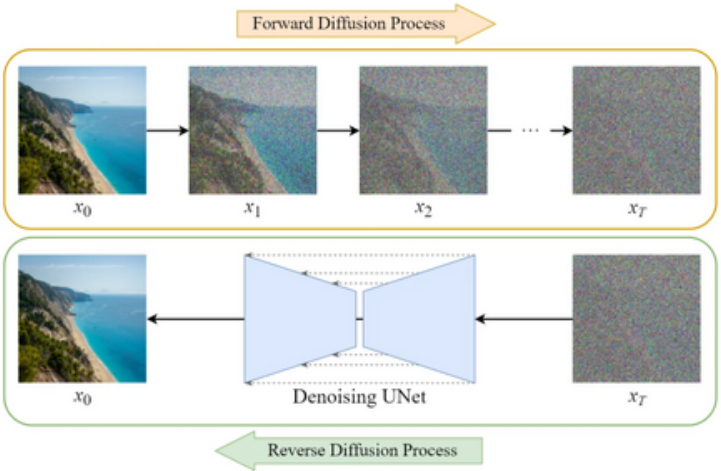


Figure 15: Stable Diffusion Progress

Table of Contents

Future Research Directions

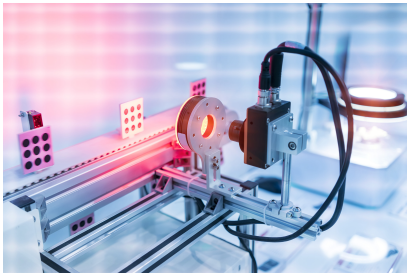
Multilingual VLMs

- Most current VLMs are trained in English, limiting access for non-English speakers.
- **Challenges:**
  - Lack of high-quality multilingual vision-language datasets
  - Tokenization issues (e.g., BPE, WordPiece) optimized for English
  - Fragmentation of common words in non-English scripts
  - Loss of semantics and struggles with languages like Thai or Japanese
- **Real-World Need:**Image-based learning, customer service, and healthcare in multilingual regions.

Future Research Directions

Robotics + Instruction Following

- Robots need to understand natural language commands and interact in visual environments (e.g., “pick up the red cup”).
- **Real-World Need:**
  - Autonomous systems that see, understand, and respond to humans naturally
  - Ability to process independently in real situation using visual input from environment



Future Research Directions

Efficient Inference on Edge/Mobile Devices

- **Why it matters:**Many VLMs are too large and compute-intensive for deployment on phones, AR glasses, or IoT devices.
- **Challenges:**
  - High memory, compute, and energy usage
  - Real-time latency constraints
- **Solutions:**
  - Model quantization, distillation, hardware-specific optimization
  - Lightweight architectures (e.g., MobileViT, MobileSAM)
- **Use Cases:**AR assistants, mobile translators, smart wearables, offline captioning tools.

Future Research Directions

Bias Mitigation and Dataset Fairness

- **Why it matters:**VLMs trained on web-scale data often inherit social, gender, and cultural biases.
- **Risks:**
  - Stereotyping, exclusion, misinformation
  - Negative impact in education, hiring, or healthcare
- **Approaches:**
  - Curated datasets with diverse representation
  - Post-hoc debiasing techniques and fairness audits
- **Goal:** Ensure VLMs are ethical, inclusive, and fair.

Future Research Directions

Model Transparency and Explainability

- **Why it matters:** Users and developers need to trust and understand model decisions.
- **Challenges:**
  - Transformer models are complex and act like black boxes
  - Hard to trace how visual cues influence outputs
- **Research Directions:**
  - Attention maps, saliency heatmaps, token attribution
  - Explanation-by-example, interactive debugging tools
- **Real-World Need:** Healthcare, law, education—where decisions must be justified.

Table of Contents

Conclusion

- Vision-Language Models are transforming how AI interacts with the world.
- Applications are rapidly expanding into education, search, creative tools, healthcare, and more.
- Despite advancements, challenges in efficiency, fairness, and transparency remain open.
- Continued research in unified architectures, better alignment, and deployment efficiency is critical.

SEMINAR
Vision-Language Models Survey

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology
Vietnam National University Ho Chi Minh City

Background

With the advent of deep learning, visual recognition research has achieved great success by leveraging end-to-end trainable deep neural networks (DNNs). However, the shift from traditional machine learning to deep learning comes with two new grand challenges:

- The slow convergence of DNN training under the classical setup of Deep Learning from scratch
- The laborious collection of large-scale, task specific, and crowd-labelled data in DNN training.