Bocconi
AI & Neuroscience
Student
Association

Vision Language Models For Vision Tasks

By:

Hugo Arsénio

Matteo Mello Grand

Riccardo Bertamini

INDEX

TRADITIONAL MACHINE LEARNING TECHNIQUES

Feature engineering (Done manually) → Supervised classification methods such as Support Vector Machines

Problems

- Scalability
- Domain expertise
- Not suitable for complex problems

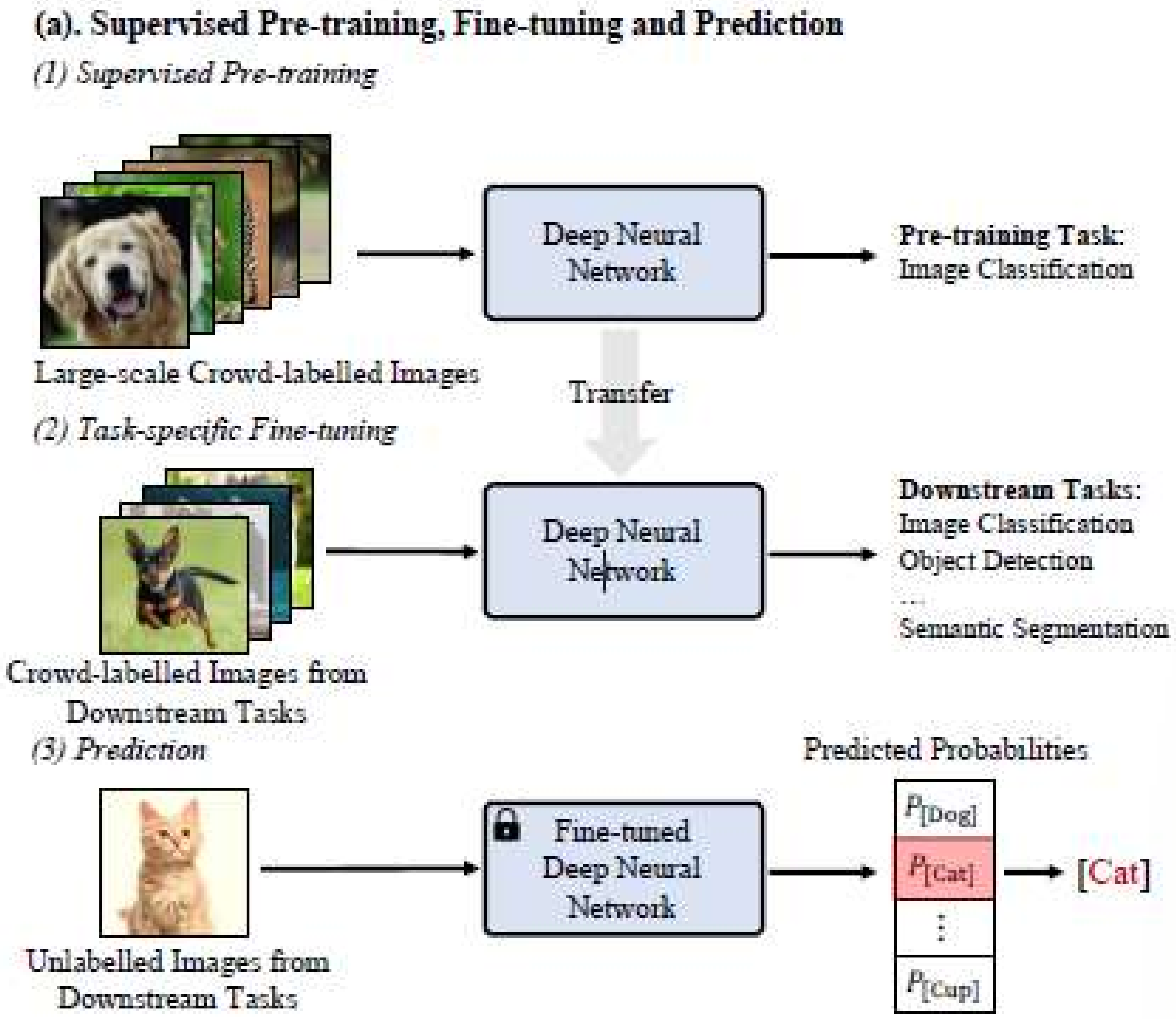
DEEP LEARNING APPROACH

- Allows to avoid feature engineering, solving therefore the problems that traditional ML methods had
- Example: Convolutional neural networks architectures (such as RestNet)

PROBLEMS

- Necessity of large-scale task specific crowd-labelled data
- Slow convergence of DNN training

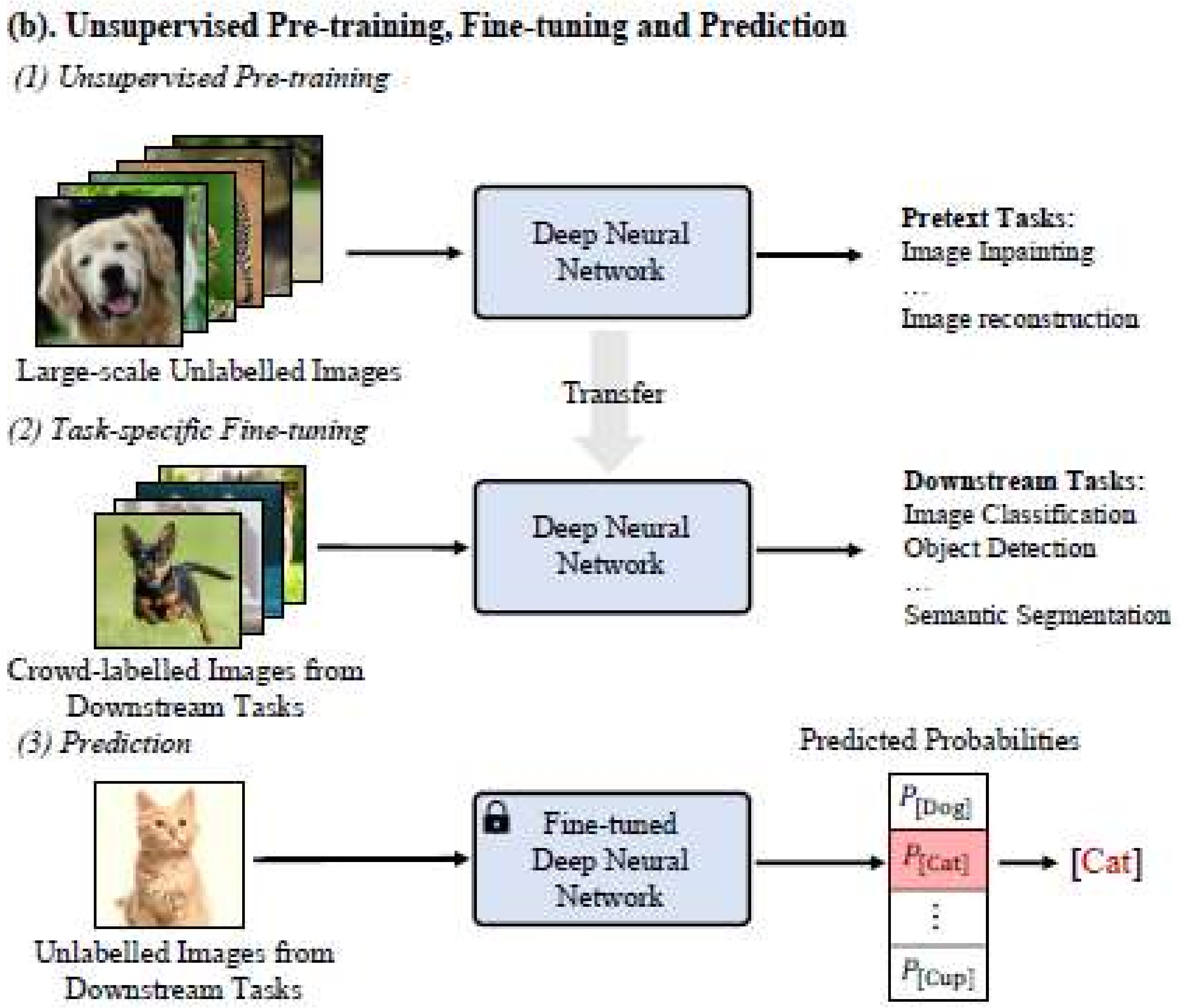
Supervised pre-training approach



UNSUPERVISED PRE-TRAINING APPROACH

Instead of a supervised pre-training, we use a self-supervised pre-training.

We don't need anymore labelled data



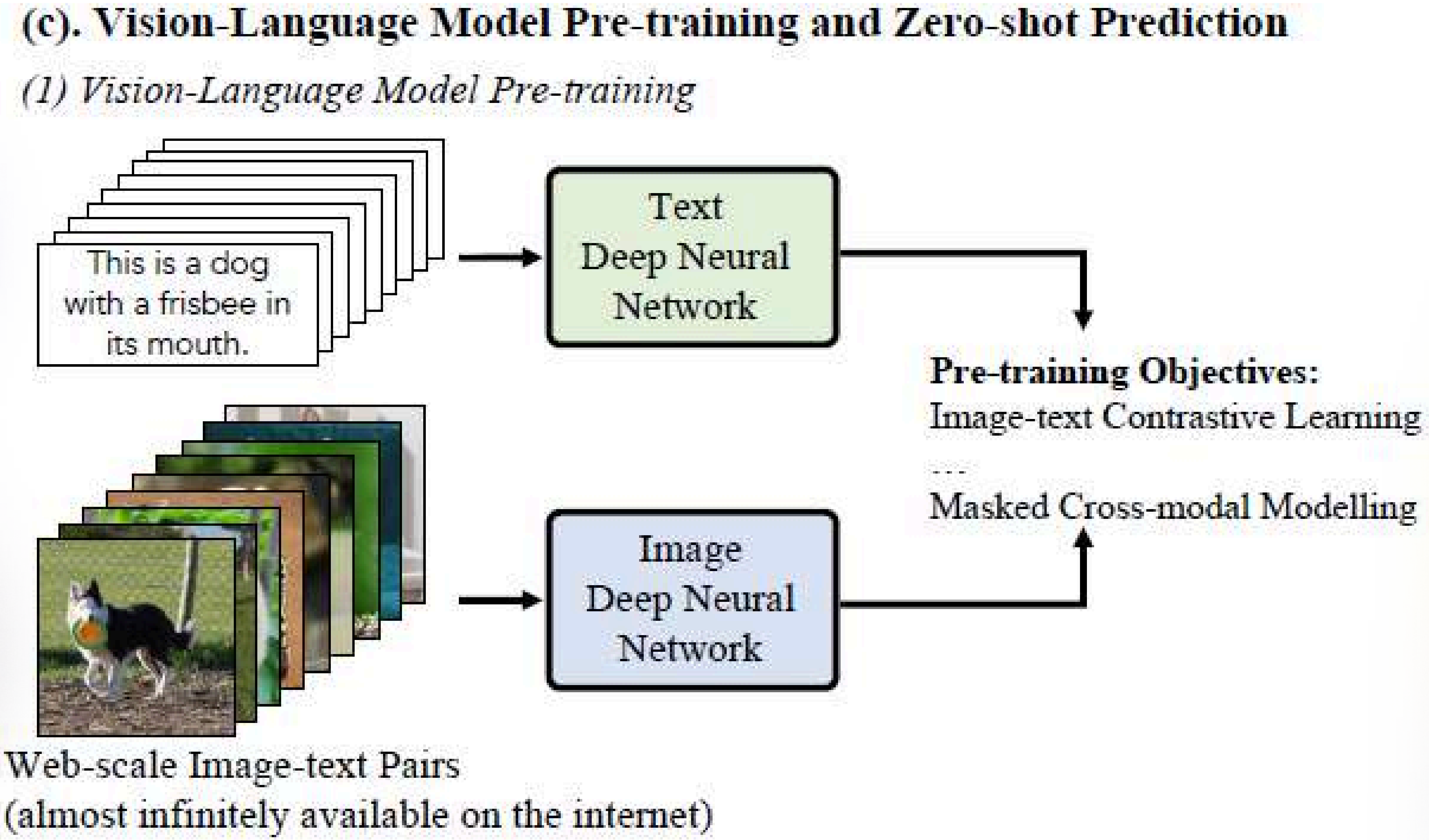
VISION LANGUAGE MODELS

- VLM is pre-trained by a vision-language objective
- It uses image-text paired data, which are in large scale present in the web

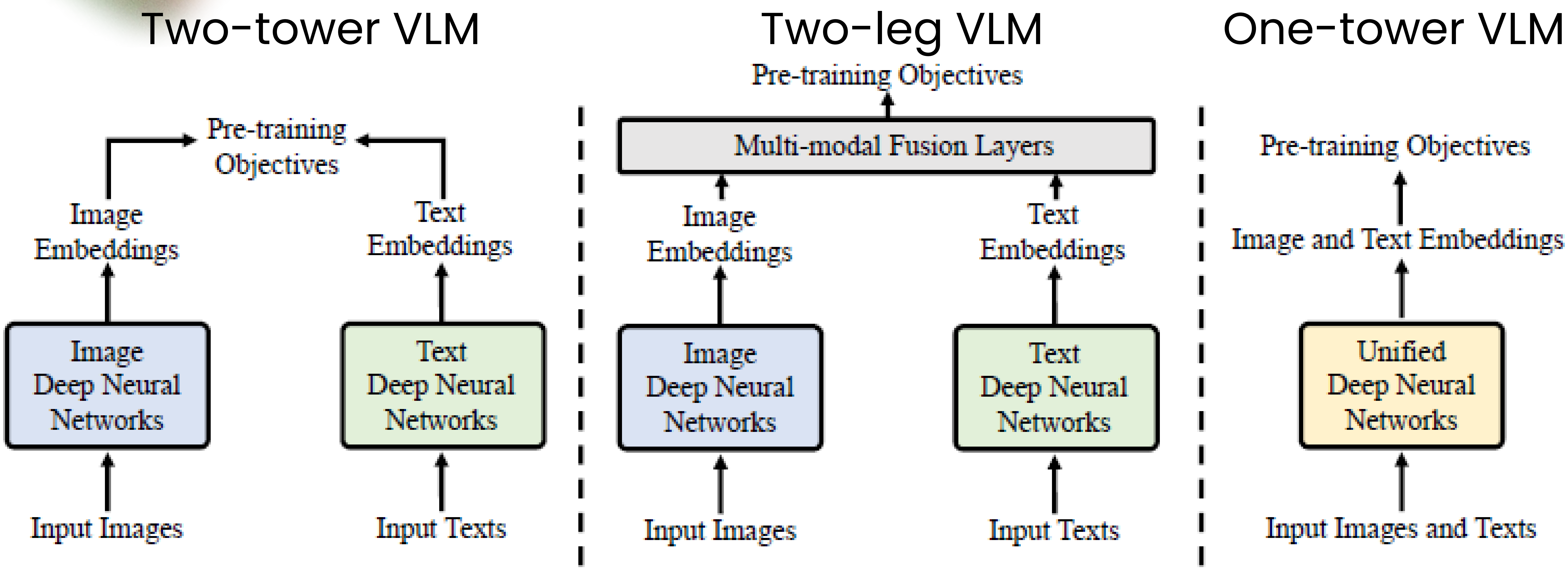
It is easier having enough data for the training

There is no necessity of fine-tuning

Pre-training



VLM FRAMEWORKS



PRE-TRAINING ARCHITECTURES

Image features learning	Text features learning
Convolutional Neural Networks example: RestNet **Transformers** **I**mage is divided into small patches and then fed into the encoder Example: ViT	**Transformers** Standard transformer architecture

Contrastive objectives

Pulls paired images and texts close and pushes others far away in the embedding space

Image Contrastive learning

Images forced to stay near their positive keys and far from their negative keys

For a batch of images B, the objective is:

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

Contrastive objectives

Image Text Contrastive learning

The objective considers both the images and the texts

We define the objective as the sum of the two following functions:

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

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

Generative objectives

The objective learns semantic features by generating data and checking whether are correct or not

Masked cross-modal learning

Integrates masked image learning with masked language learning. For a batch of images B, the loss function is:

$$L_{MCM} = -\frac{1}{B}\sum_{i=1}^{B}\left[log f_{\theta}\left(x_i^I \,\middle|\, x_{i,u}^I, x_{i,u}^T\right) + log f_{\phi}\left(x_i^T \,\middle|\, x_{i,u}^T, x_{i,u}^I\right)\right]$$

Generative objectives

Image to text generation

Aims to predict text autoregressively based on the images paired with that text

$$\mathcal{L}_{ITG} = -\sum_{l=1}^{L} \log\ f_\theta(x^T \mid x^T_{<l}, z^I)$$

Alignment objectives

Learns how to link images and text

Image text matching

Models global correlation between images and text

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

Region-Word matching

Local cross-modal correlations in image text pairs

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

VLM EVALUATION

Zero shot prediction

The pre-trained VLM is applied directly to a task

- Image classification
- Semantic segmentation
- Object detection
- Image-text retrieval

Linear probing

It freezes the pre-trained VLM and train a linear classifier to classify the VLM encoded embeddings to assess the VLM representation

VISION-LANGUAGE MODEL PRE-TRAINING

What is pre-training in VLMs?

Learn general patterns from vast datasets -> ‘zero-shot prediction’

Pre-training objectives:

A. Contrastive Objectives: learn to match images and text.
B. Generative Objectives: learn to fill in missing information.
C. Alignment Objectives: learn to correctly link parts of an image to text.

How these objectives work together?

Identifying objects in unfamiliar scenes, answering questions about images, and performing tasks that require understanding of specific details.

VLM TRANSFER LEARNING

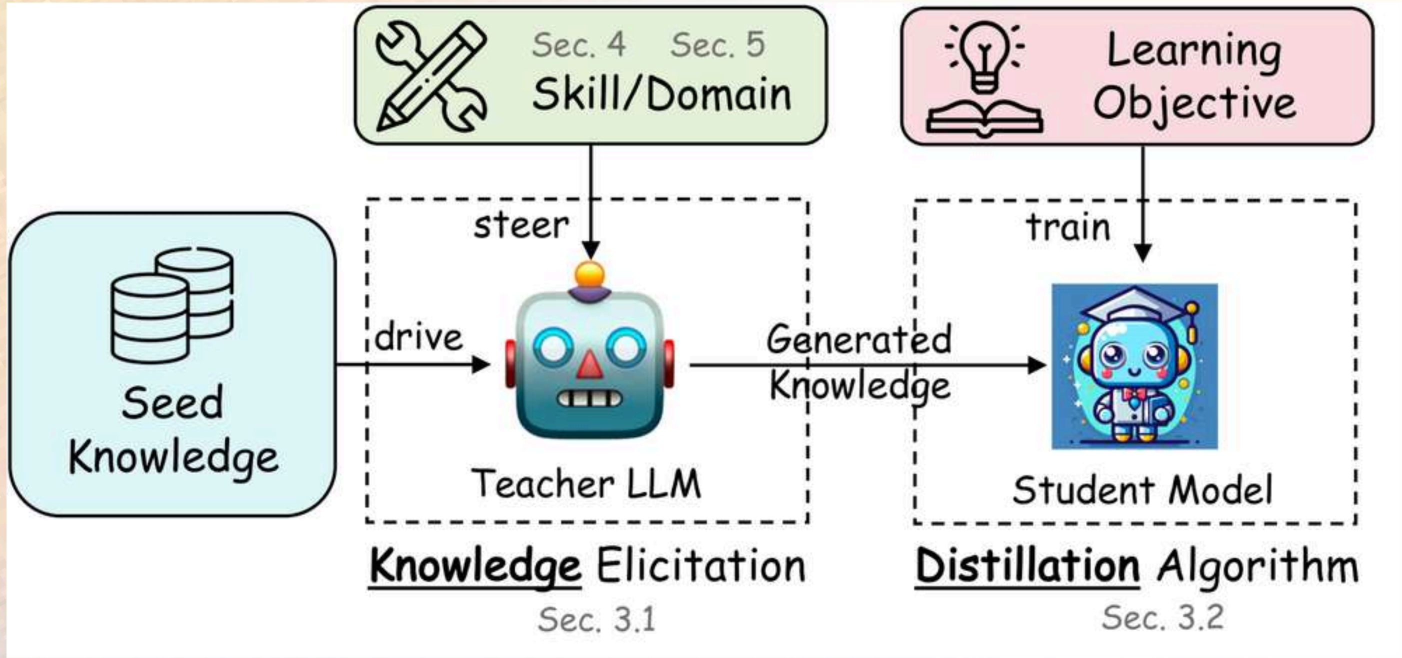
Prompt Tuning: method where specific text prompts are optimized to make VLMs respond accurately to a certain task.

Feature Adapters: add-on module to the model's architecture, allowing task-specific features to be learned without altering the core model.

Cross-Attention Modules: integration of information from different sources at a more granular level -> associate image details with text instructions or queries.

VLM KNOWLEDGE DISTILLATION

- extracts the most important part of the knowledge
- uses task-specific models without any restriction of VLM architecture
- transfers image-level knowledge to region/pixel-level tasks



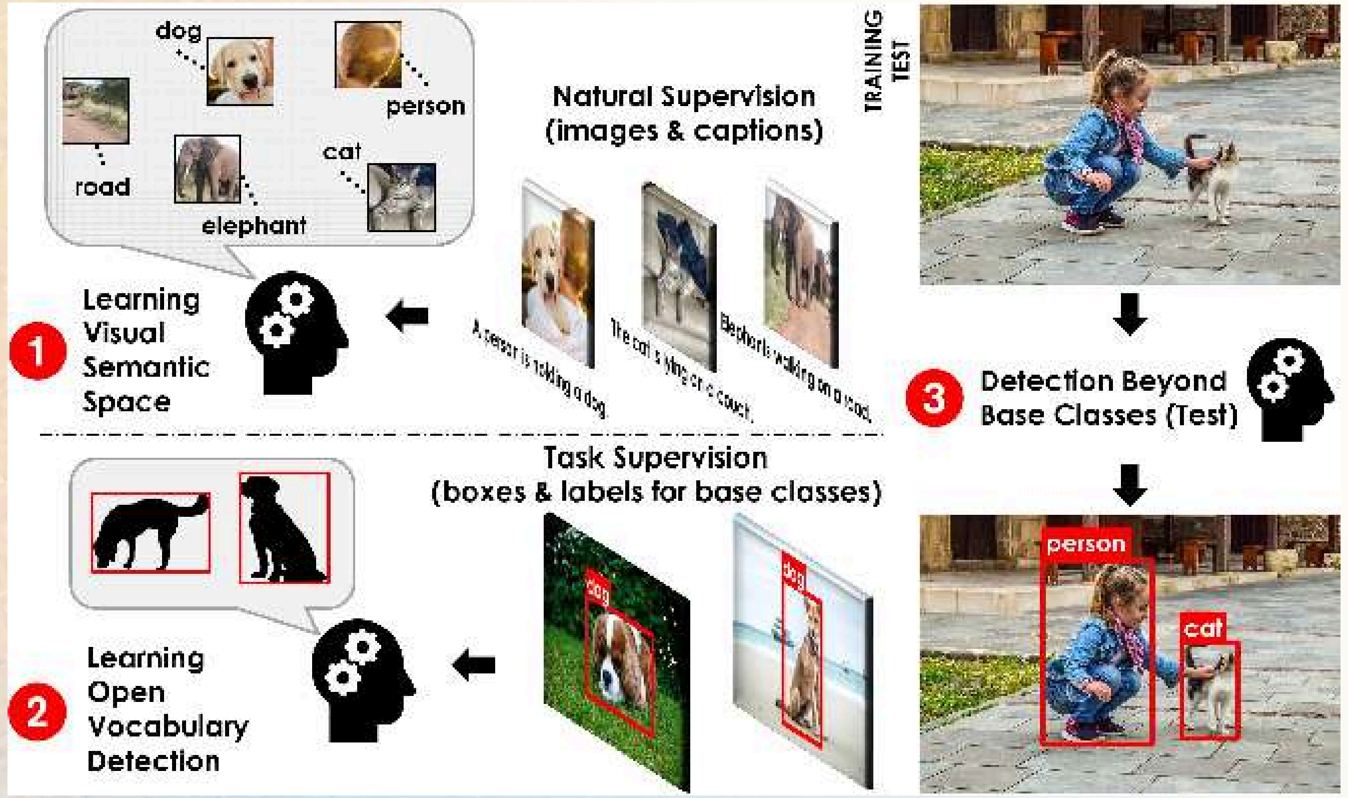
- **ISSUE** -> how does it manage to distil this knowledge while tackling complex dense predictions such as:
  - object detection
  - semantic segmentation

OBJECT DETECTION

AIM -> better align image-level and object-level representations

- Open Vocabulary -> detects objects described by arbitrary texts

- VLM distillation via PL, e.g. CLIP



$$e_{\mathrm{img}} = \mathrm{CLIP}_{\mathrm{img}}(x_i), \quad e_{\mathrm{txt}}^{k} = \mathrm{CLIP}_{\mathrm{txt}}(T(\mathrm{CLS}_k)).$$

$$\mathrm{P}(y = k|x) = \frac{\exp(\cos(e_{\mathrm{img}}, e_{\mathrm{txt}}^{k})/\tau)}{\sum_{i=1}^{K} \exp(\cos(e_{\mathrm{img}}, e_{\mathrm{txt}}^{i})/\tau)}$$

$$T(\mathrm{CLS}_k) = [V]_1 [V]_2 \dots [V]_M [\mathrm{CLS}_k].$$

$$\mathcal{L}_{\mathrm{CE}}(x_i, y_i) = -\sum_{k=1}^{K} \mathbb{1}\{y_i = k\} \log P(y = k|x_i).$$

SEMANTIC SEGMENTATION

Tackles mismatches between image-level and pixel-level representations

- Open-Vocabulary -> aims to segment pixels described by arbitrary texts
- Knowledge distillation for weakly supervised semantic segmentation-> leverages both VLMs and weak supervision for semantic segmentation

VLM COMPARISON

VLM pre-training

ADVANTAGES	DISADVANTAGES

• performance is up to the model size.
• superior generalization attributed to:
  ○ Big Data -> there are many different images on the internet
  ○ Big Model -> adopt larger models compared to traditional visual recognition models
	• if data/model size keeps increasing, eventually the performance saturates
	• extensive computation resources, hundreads of hours of training
	• Extremely expensive

VLM Transfer Learning

ADVANTAGES	DISADVANTAGES
• help in downstream tasks • are able to mitigate domain gaps • “unsupervised transfer”=”few-shot supervised transfer” (in terms of performance) • Has lower overfitting risks	• presence of noisy pseudo labels

---

VLM Knowledge Distillation

ADVANTAGES	DISADVANTAGES
• brings performance improvements	• not enough studies about it

VLM Pre-Training	• achieves remarkable zero-shot prediction • Development for dense visual recognition tasks lags far behind
VLM Transfer Learning	• Has made a lot of progress across image classification datasets • Unsupervised transfer has been neglected
VLM Knowledge Distillation	• Extremely efficient in task-specific environments • Very hard to benchmark fairly

CONCLUSION

FUTURE DIRECTIONS

VLM Pre-Training

- Unification of vision and language learning
- Pre-Training VLMs with multiple languages
- Data-efficient VLM
- Pre-Training VLMs with LLMs

VLM Transfer Learning

- Unsupervised VLM transfer
- VLM tranfer with LLMs

VLM Knowledge Distillation

- Ditilling knowledge from multiple VLMs

CONCLUSION

SOURCES

- https://github.com/jingyi0000/VLM_survey
- https://openaccess.thecvf.com/content/CVPR2024/papers/Bang_Active_Prompt_Learning_in_Vision_Language_Models_CVPR_2024_paper.pdf
- https://arxiv.org/html/2310.08255v2

THANK YOU

If you have any questions, don’t be afraid to ask