Case Study:
Vision Language Models

Samrat Mukherjee
PhD Student

Center for Machine Intelligence and Data Science
Indian Institute of Technology, Bombay

07 March, 2024

Outline

- **Vision Language Models (VLMs).**
  - Introduction.
  - Examples.
- **Prompt Learning.**
  - Motivation.
  - Examples.
- **VLMs for Remote-Sensing use-cases.**
  - GeoChat : Grounded Large Vision-Language Model for Remote Sensing.

Introduction

- Deep Neural Networks (DNNs) are data hungry.
  - Rely heavily on manually **labelled data** for training. (Eg. ImageNet, MS-COCO etc)
  - Train a DNN for each single visual recognition task.
- Vision-Language Models (VLMs).
  - Combine natural language understanding with computer vision capabilities.
- VLMs learn rich vision-language correlation.
  - Uses image-text pairs available on the Internet.
  - Enables zero-shot predictions on various visual recognition tasks with a single VLM.

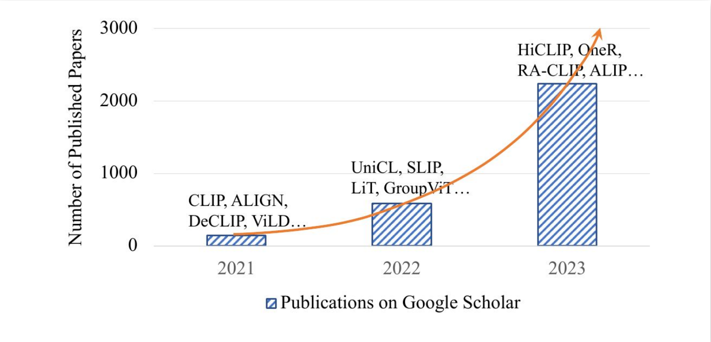


Fig 1: Trend of published work with VLMs [1]

How is it different than previous approaches?

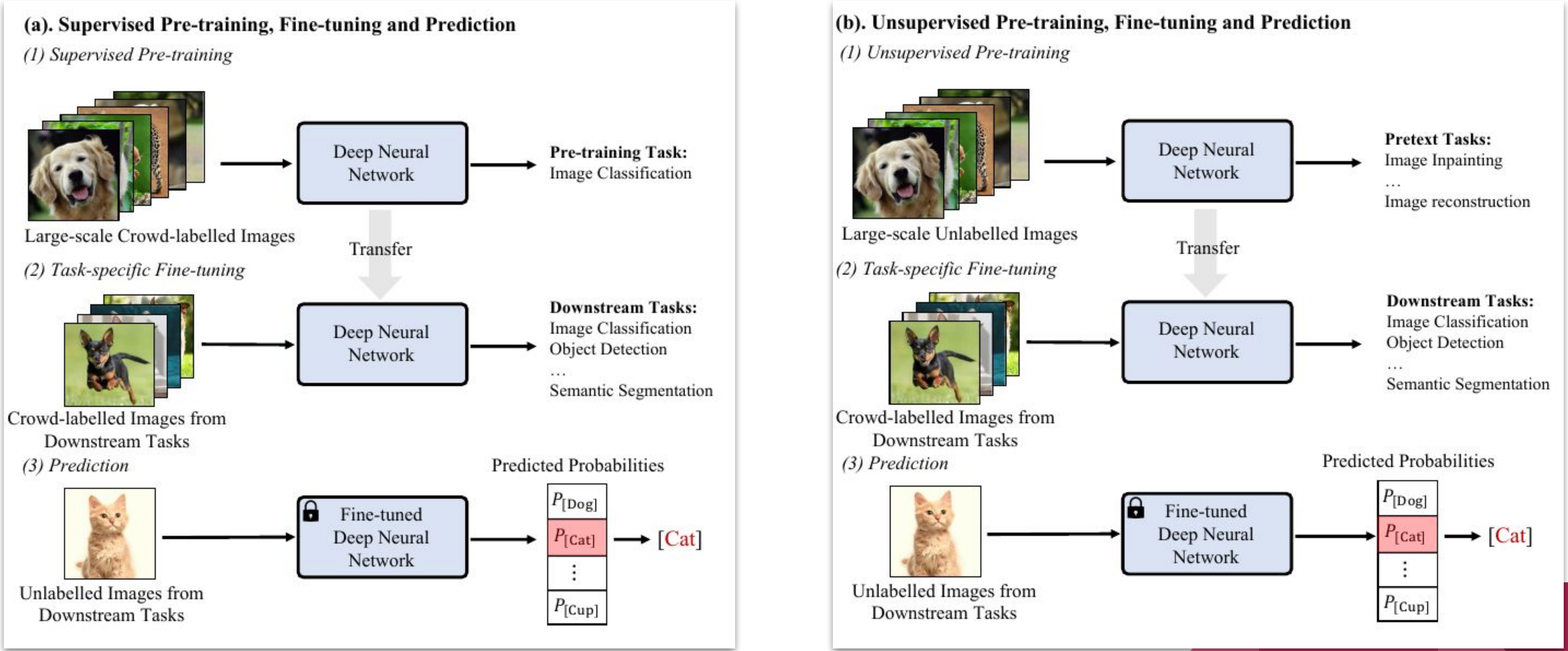

(a). Supervised Pre-training, Fine-tuning and Prediction
(1) Supervised Pre-training
Large-scale Crowd-labelled Images
Deep Neural Network
Pre-training Task:
Image Classification
Transfer
(2) Task-specific Fine-tuning
Crowd-labelled Images from Downstream Tasks
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
(3) Prediction
Unlabelled Images from Downstream Tasks
Fine-tuned Deep Neural Network
Predicted Probabilities
$P_{[Dog]}$
$P_{[Cat]}$
⋮
$P_{[Cup]}$
[Cat]
(b). Unsupervised Pre-training, Fine-tuning and Prediction
(1) Unsupervised Pre-training
Large-scale Unlabelled Images
Deep Neural Network
Pretext Tasks:
Image Inpainting
…
Image reconstruction
Transfer
(2) Task-specific Fine-tuning
Crowd-labelled Images from Downstream Tasks
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
(3) Prediction
Unlabelled Images from Downstream Tasks
Fine-tuned Deep Neural Network
Predicted Probabilities
$P_{[Dog]}$
$P_{[Cat]}$
⋮
$P_{[Cup]}$
[Cat]

Fig.2: Deep Neural Network training paradigm (Supervised (left) and Unsupervised (right)) [1]

How is it different than previous approaches?

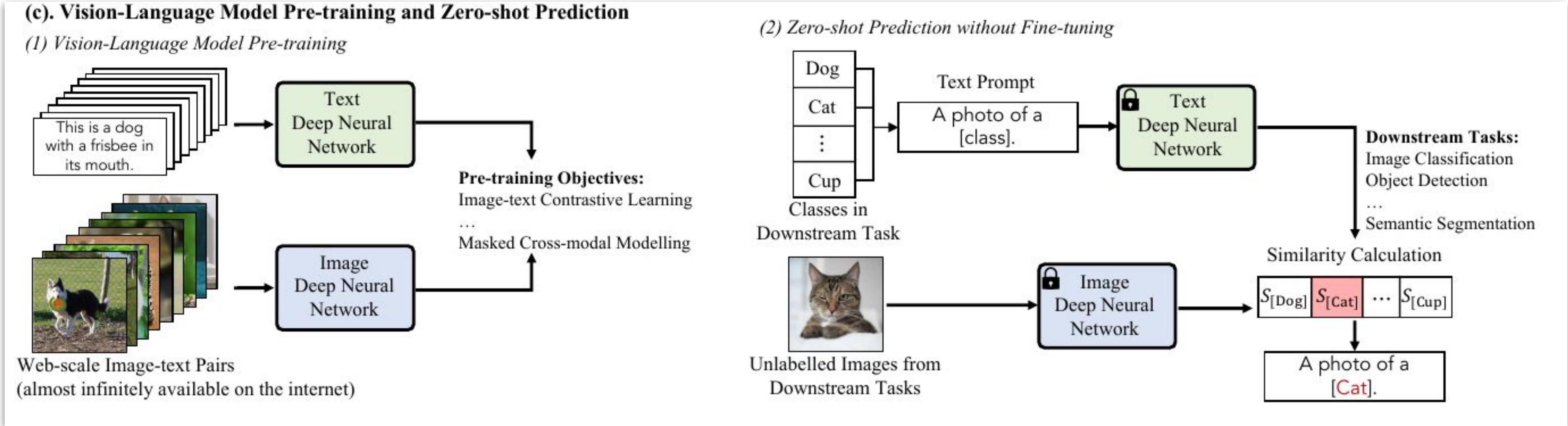

(c). Vision-Language Model Pre-training and Zero-shot Prediction
(1) Vision-Language Model Pre-training
This is a dog with a frisbee in its mouth.
Text Deep Neural Network
Image Deep Neural Network
Pre-training Objectives:
Image-text Contrastive Learning
…
Masked Cross-modal Modelling
Web-scale Image-text Pairs
(almost infinitely available on the internet)
(2) Zero-shot Prediction without Fine-tuning
Dog
Cat
⋮
Cup
Classes in Downstream Task
Text Prompt
A photo of a [class].
Text Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
Similarity Calculation
$S_{[Dog]}$
$S_{[Cat]}$
…
$S_{[Cup]}$
Unlabelled Images from Downstream Tasks
Image Deep Neural Network
A photo of a [Cat].

Fig. 3: DNN Training paradigm of VLMs. Observe the “absence” of Fine-tuning phase. [1]

- Pre-trained by certain vision-language objectives.
  - Captures rich vision-language knowledge.
  - Perform zero-shot predictions by matching the embeddings of images and texts.
    - Without fine-tuning on downstream tasks.

Network architectures in VLMs

- **Architectures for learning Image features.**
  - Convolutional based architectures.
    - ***ResNet***
    - VGG
    - EfficientNet
  - Transformer based architectures.
    - *Vision Transformer.*

- **Architectures for learning Text features.**
  - *Standard Transformer.*
  - GPT-2.
  - BERT.

Pre-training objectives in VLMs

- **Contrastive objective.**
  - Image-text contrastive learning.
  - Image-text-label contrastive learning.

- **Generative objective.**
  - Masked image modelling.
  - Masked language modelling.
  - Masked cross-modal modelling.

- **Alignment objective.**
  - Image-text matching.
  - Region-word matching.

VLM pre-training framework

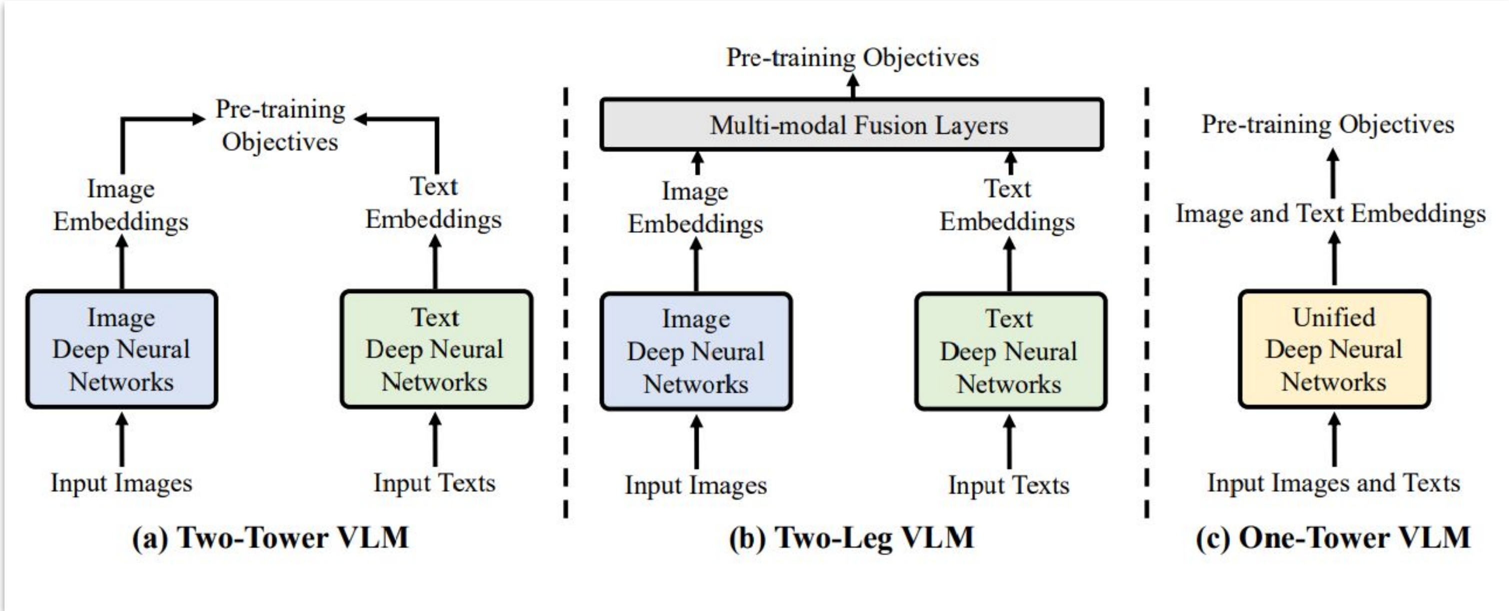
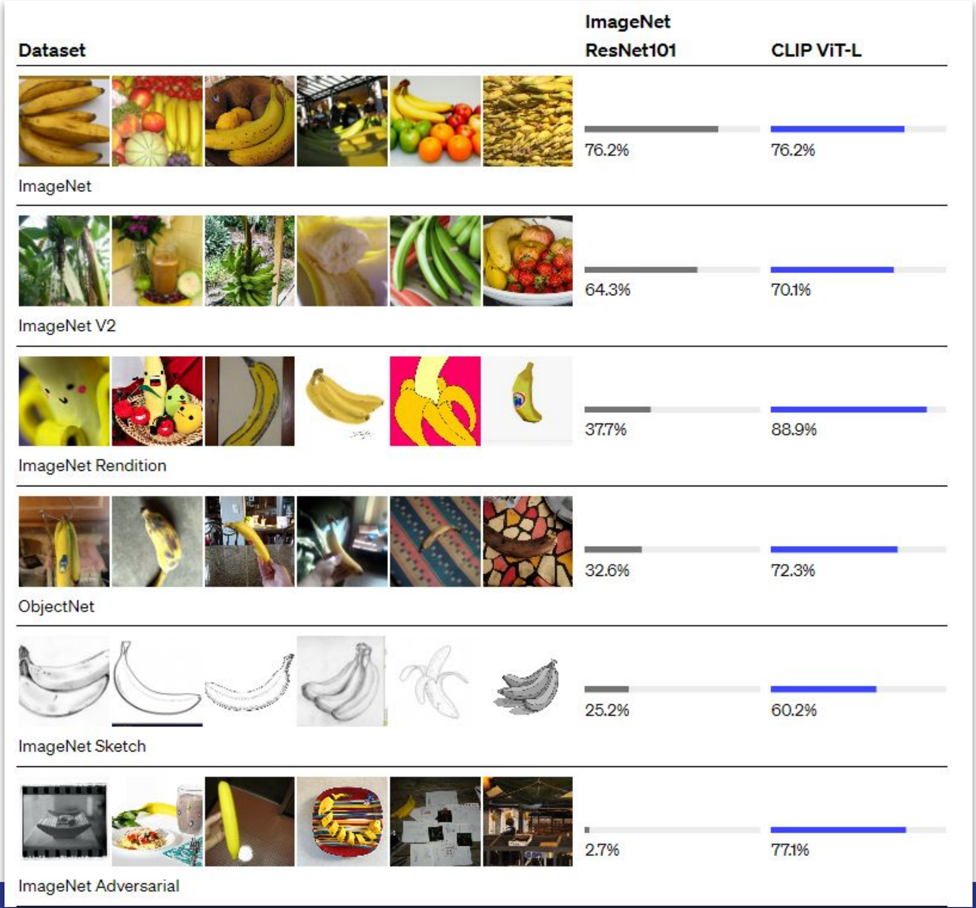


Fig. 4- Different pre-training frameworks [1]

Contrastive Language Image Pre-training (CLIP)

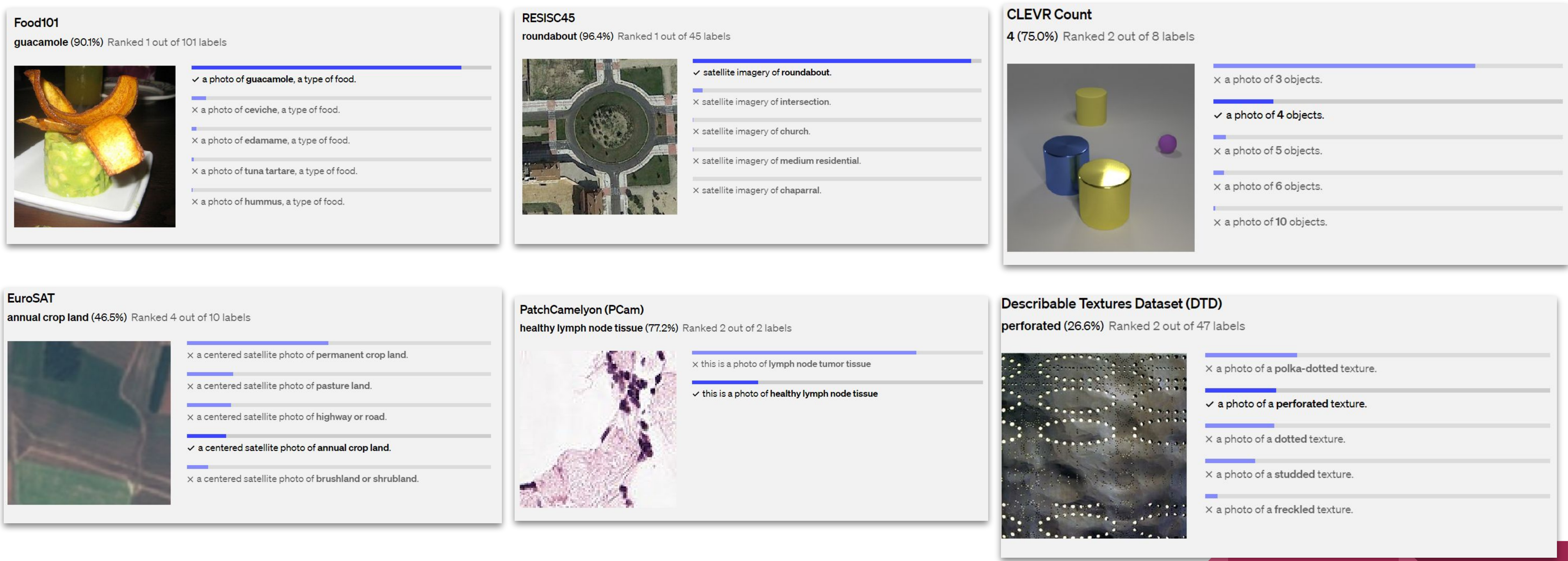


Dataset	ImageNet ResNet101	CLIP ViT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

- Introduced by OpenAI in 2021.
- Consistently performs better against ResNet101 on various dataset.

Fig. 5: Motivation towards CLIP. [2]

Some more examples



Food101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- × a photo of **ceviche**, a type of food.
- × a photo of **edamame**, a type of food.
- × a photo of **tuna tartare**, a type of food.
- × a photo of **hummus**, a type of food.

RESISC45

roundabout (96.4%) Ranked 1 out of 45 labels

- ✓ satellite imagery of **roundabout**.
- × satellite imagery of **intersection**.
- × satellite imagery of **church**.
- × satellite imagery of **medium residential**.
- × satellite imagery of **chaparral**.

CLEVR Count

4 (75.0%) Ranked 2 out of 8 labels

- × a photo of **3** objects.
- ✓ a photo of **4** objects.
- × a photo of **5** objects.
- × a photo of **6** objects.
- × a photo of **10** objects.

EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels

- × a centered satellite photo of **permanent crop land**.
- × a centered satellite photo of **pasture land**.
- × a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- × a centered satellite photo of **brushland or shrubland**.

PatchCamelyon (PCam)

healthy lymph node tissue (77.2%) Ranked 2 out of 2 labels

- × this is a photo of **lymph node tumor tissue**
- ✓ this is a photo of **healthy lymph node tissue**

Describable Textures Dataset (DTD)

perforated (26.6%) Ranked 2 out of 47 labels

- × a photo of a **polka-dotted** texture.
- ✓ a photo of a **perforated** texture.
- × a photo of a **dotted** texture.
- × a photo of a **studded** texture.
- × a photo of a **freckled** texture.

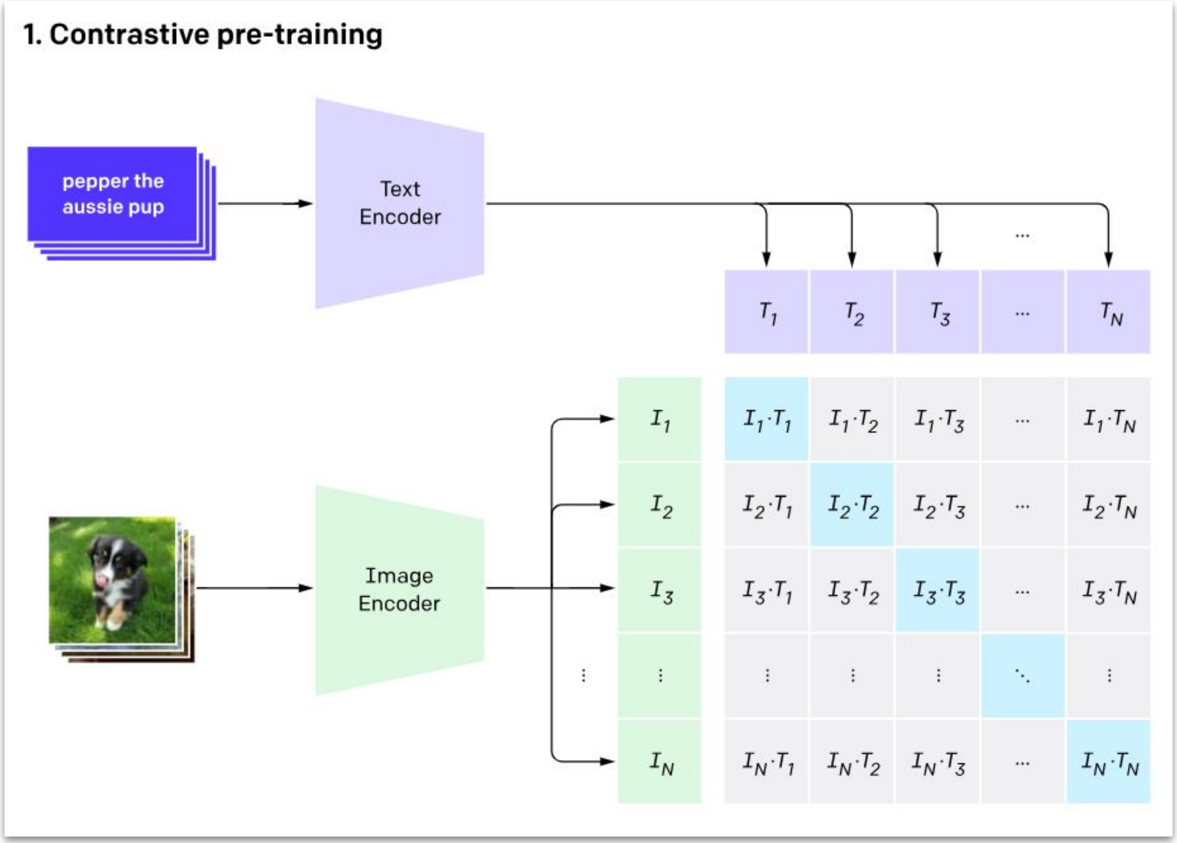
Fig. 6: Some more examples and introduction to prompts [2]

CLIP

- Scaling a simple pre-training task is sufficient to achieve competitive zero-shot performance on a great variety of image classification datasets.
- **Dataset:** the text paired with images found across the internet.
- **Proxy training task for CLIP:** Given an image, predict which out of a set of 32,768 randomly sampled text snippets, was actually paired with it in our dataset.
- To solve the task, CLIP learns broad visual concept recognition associated with their natural texts.
- As a result, CLIP models can then be applied to nearly arbitrary visual classification tasks.

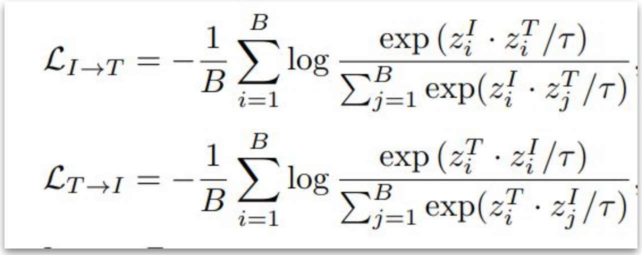
- CLIP was designed to mitigate a number of major problems in the standard deep learning approach to computer vision:
  - Costly datasets
  - Narrow
  - Poor real-world performance

Pre-training of CLIP



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Fig. 7: Pre-training step for CLIP [2]



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

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

Applying CLIP to new task

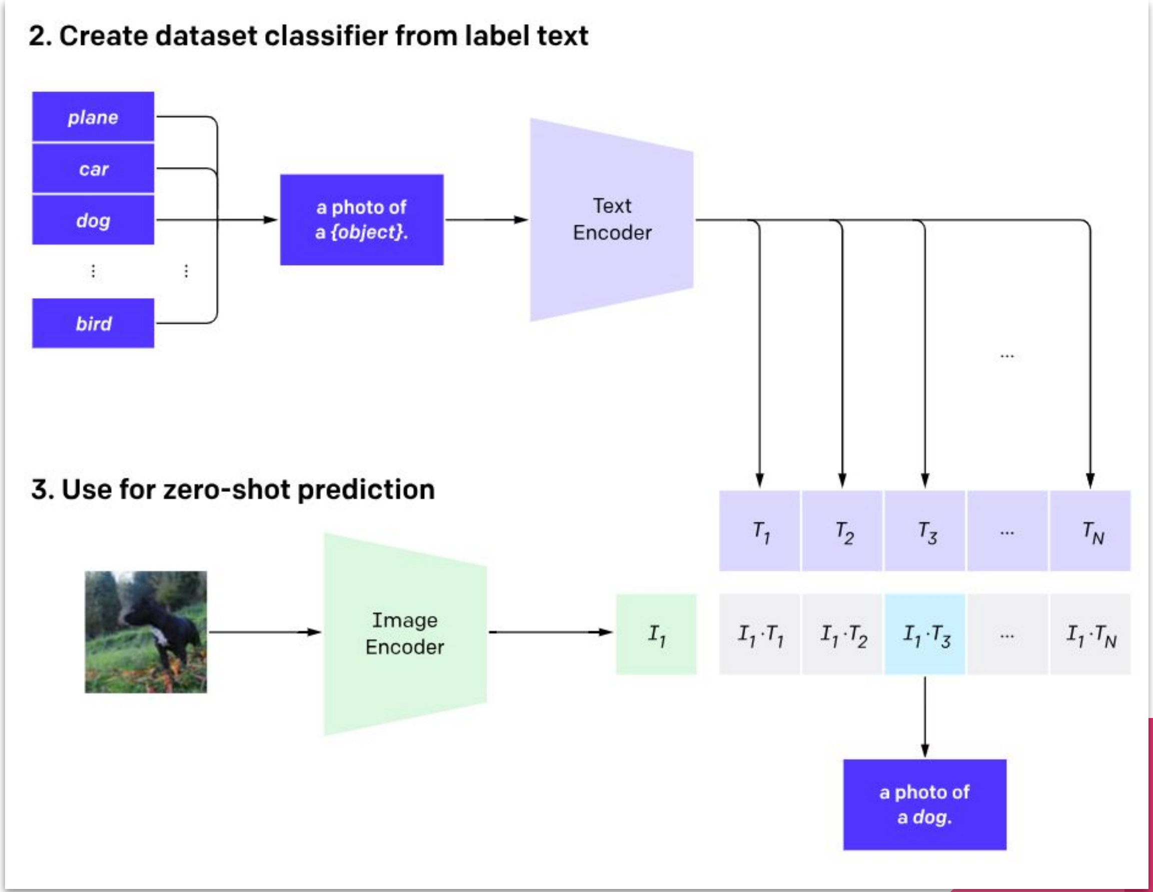


Fig. 8: Apply CLIP to a new task [2]

Limitations

- CLIP struggles on more abstract or systematic tasks.
  - Counting the number of objects in an image
  - Predicting how close the nearest car is in a photo.
  - *On these two datasets, zero-shot CLIP is only slightly better than random guessing.*

- Zero-shot CLIP also struggles compared to task specific models on very fine-grained classification.
  - Such as telling the difference between car models, variants of aircraft, or flower species.

- CLIP is sensitive to wording or phrasing.
  - Require trial and error "prompt engineering" to perform well.

Motivation towards learning the prompts

For pre-trained vision-language models, the text input, known as prompt, plays a key role in downstream datasets.

Manual prompt engineering is a non-trivial task and requires prior knowledge of the domain and the VLM.

Prompt learning research started flourishing lately in Natural Language Processing domain.

CoOp (**Co**ntext **Op**timization)

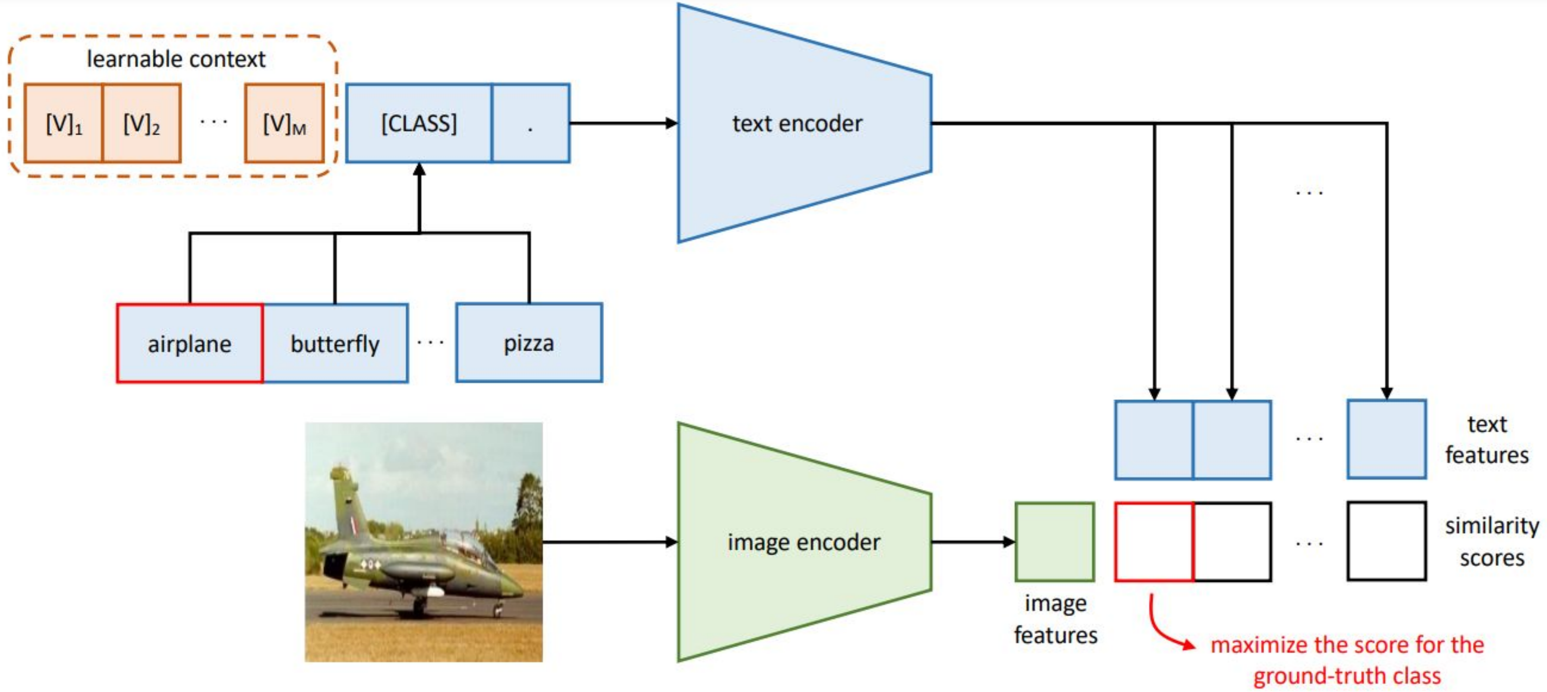
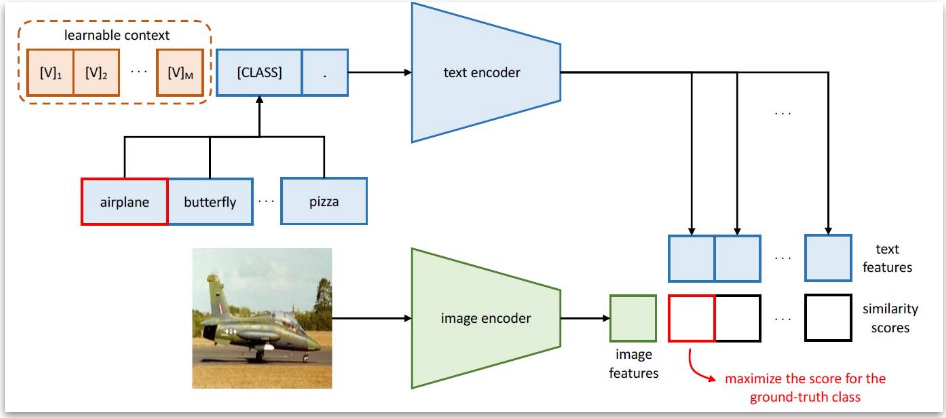


Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

Fig. 9: CoOp learning [3]

CoOp (**Co**ntext **Op**timization)



CoOp models prompt's context words with learnable vectors, which could be initialized with either random values or pre-trained word embeddings.

CLIP prediction

$$p(y=i|\boldsymbol{x})=\frac{\exp(\cos(\boldsymbol{w_i},\boldsymbol{f})/\tau)}{\sum_{j=1}^{K}\exp(\cos(\boldsymbol{w_j},\boldsymbol{f})/\tau)}$$

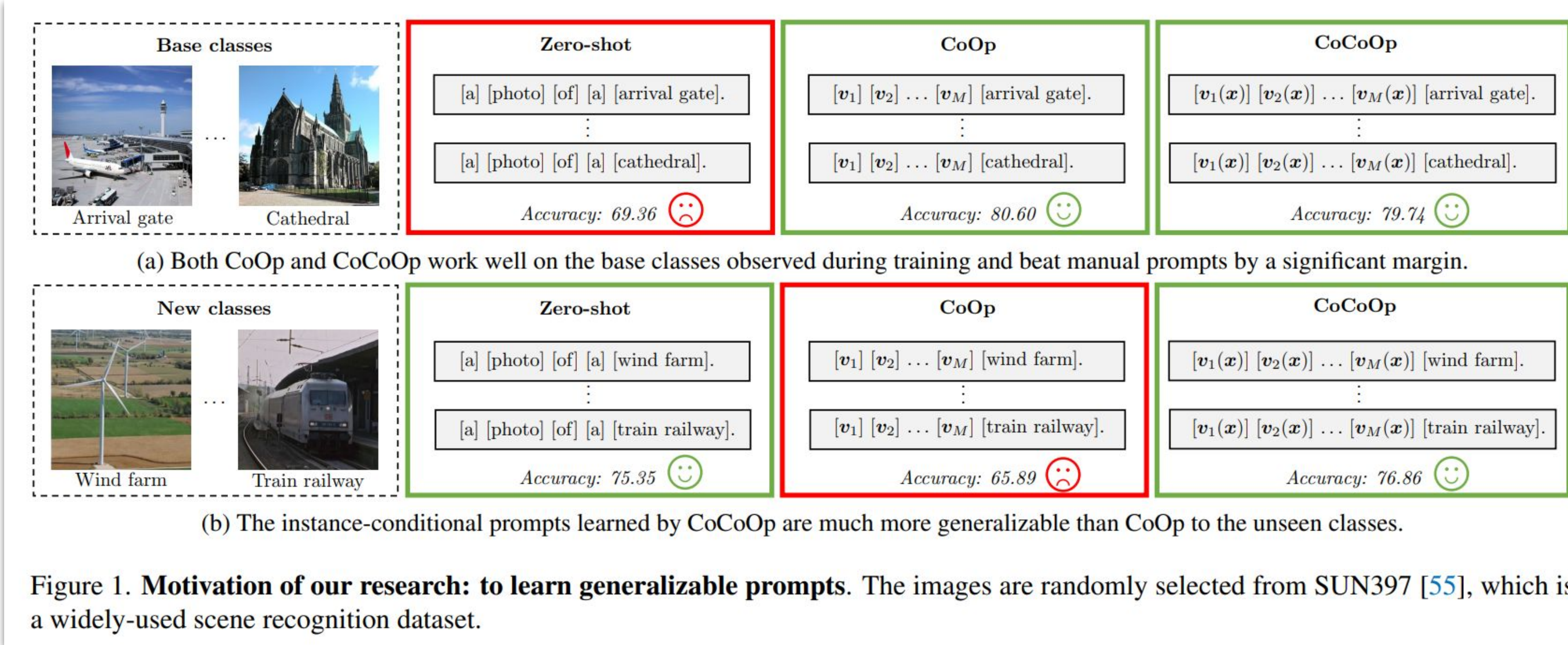
Using CoOp

$$p(y=i|\boldsymbol{x})=\frac{\exp(\cos(g(\boldsymbol{t}_i),\boldsymbol{f})/\tau)}{\sum_{j=1}^{K}\exp(\cos(g(\boldsymbol{t}_j),\boldsymbol{f})/\tau)}$$

$$\boldsymbol{t}=[\mathrm{V}]_1[\mathrm{V}]_2\dots[\mathrm{V}]_M[\mathrm{CLASS}],$$

$$\boldsymbol{t}=[\mathrm{V}]_1\dots[\mathrm{V}]_{\frac{M}{2}}[\mathrm{CLASS}][\mathrm{V}]_{\frac{M}{2}+1}\dots[\mathrm{V}]_M$$

Conditional **Co**ntext **Op**timization (CoCoOp)



Base classes

Arrival gate ... Cathedral

Zero-shot

[a] [photo] [of] [a] [arrival gate].

⋮

[a] [photo] [of] [a] [cathedral].

Accuracy: 69.36 ☹

CoOp

$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [arrival gate].

⋮

$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [cathedral].

Accuracy: 80.60 ☺

CoCoOp

$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [arrival gate].

⋮

$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [cathedral].

Accuracy: 79.74 ☺

(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.

New classes

Wind farm ... Train railway

Zero-shot

[a] [photo] [of] [a] [wind farm].

⋮

[a] [photo] [of] [a] [train railway].

Accuracy: 75.35 ☺

CoOp

$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [wind farm].

⋮

$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [train railway].

Accuracy: 65.89 ☹

CoCoOp

$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [wind farm].

⋮

$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [train railway].

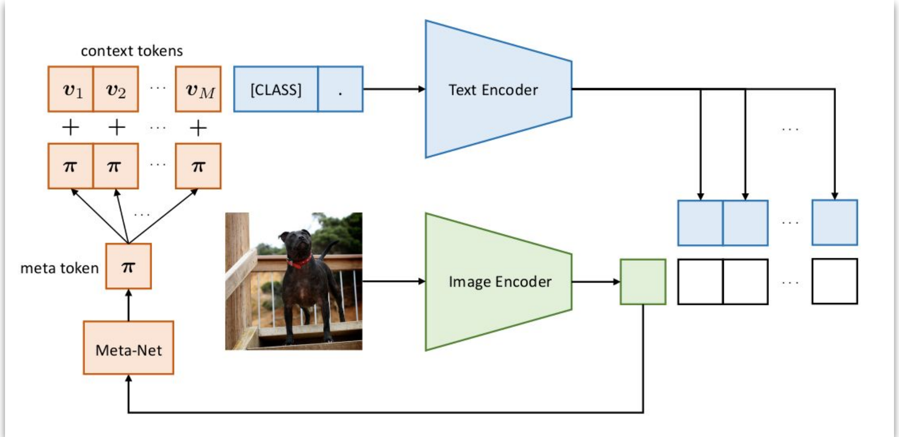
Accuracy: 76.86 ☺

(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

Figure 1. **Motivation of our research: to learn generalizable prompts**. The images are randomly selected from SUN397 [55], which is a widely-used scene recognition dataset.

Fig. 10: Motivation for CoCoOp [4]

Conditional **Co**ntext **Op**timization (CoCoOp)



- Instance-conditional context can generalize better because it shifts the focus away from a specific set of classes—for reducing overfitting—to each input instance. *Inspired from Image Captioning task.*

- For each input generate a conditional token from the Meta-Net which is then combined with the context vector. *(MetaNet : Linear → ReLU → Linear)*

$$p(y|\boldsymbol{x}) = \frac{\exp(\mathrm{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K} \exp(\mathrm{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i(\boldsymbol{x}))/\tau)}.$$

$$\boldsymbol{t}_i(\boldsymbol{x}) = \{\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}), \ldots, \boldsymbol{v}_M(\boldsymbol{x}), \boldsymbol{c}_i\}$$

$$\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi} \qquad \boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

VLMs in Remote Sensing

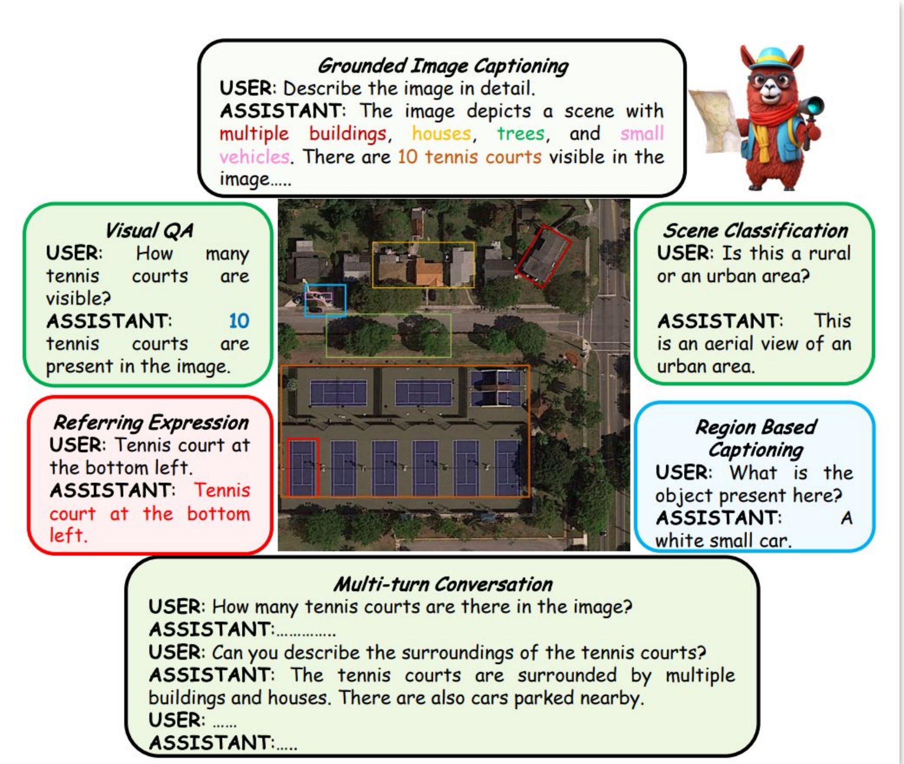
- General-domain VLMs perform poorly for Remote Sensing (RS) scenarios.
  - Fabricated information when presented with RS domain-specific queries.
- Unique challenges introduced by RS imagery.
  - High-resolution RS imagery with diverse scale changes across categories.
  - Many small objects.

- Several VLMs have been proposed for different tasks like-
  - Scene understanding
  - Region based captioning. etc.

But not a generic VLM.....

Geo-Chat

- The first versatile remote sensing VLM that offers multitask conversational capabilities with high-resolution RS images.
- Answer image-level queries but also accepts region inputs to hold region-specific dialogue



Grounded Image Captioning
USER: Describe the image in detail.
ASSISTANT: The image depicts a scene with multiple buildings, houses, trees, and small vehicles. There are 10 tennis courts visible in the image.....

Visual QA
USER: How many tennis courts are visible?
ASSISTANT: 10 tennis courts are present in the image.

Scene Classification
USER: Is this a rural or an urban area?
ASSISTANT: This is an aerial view of an urban area.

Referring Expression
USER: Tennis court at the bottom left.
ASSISTANT: Tennis court at the bottom left.

Region Based Captioning
USER: What is the object present here?
ASSISTANT: A white small car.

Multi-turn Conversation
USER: How many tennis courts are there in the image?
ASSISTANT:............
USER: Can you describe the surroundings of the tennis courts?
ASSISTANT: The tennis courts are surrounded by multiple buildings and houses. There are also cars parked nearby.
USER: ......
ASSISTANT:.....

Fig. 11: GeoChat and its capabilities [5]

Geo-Chat

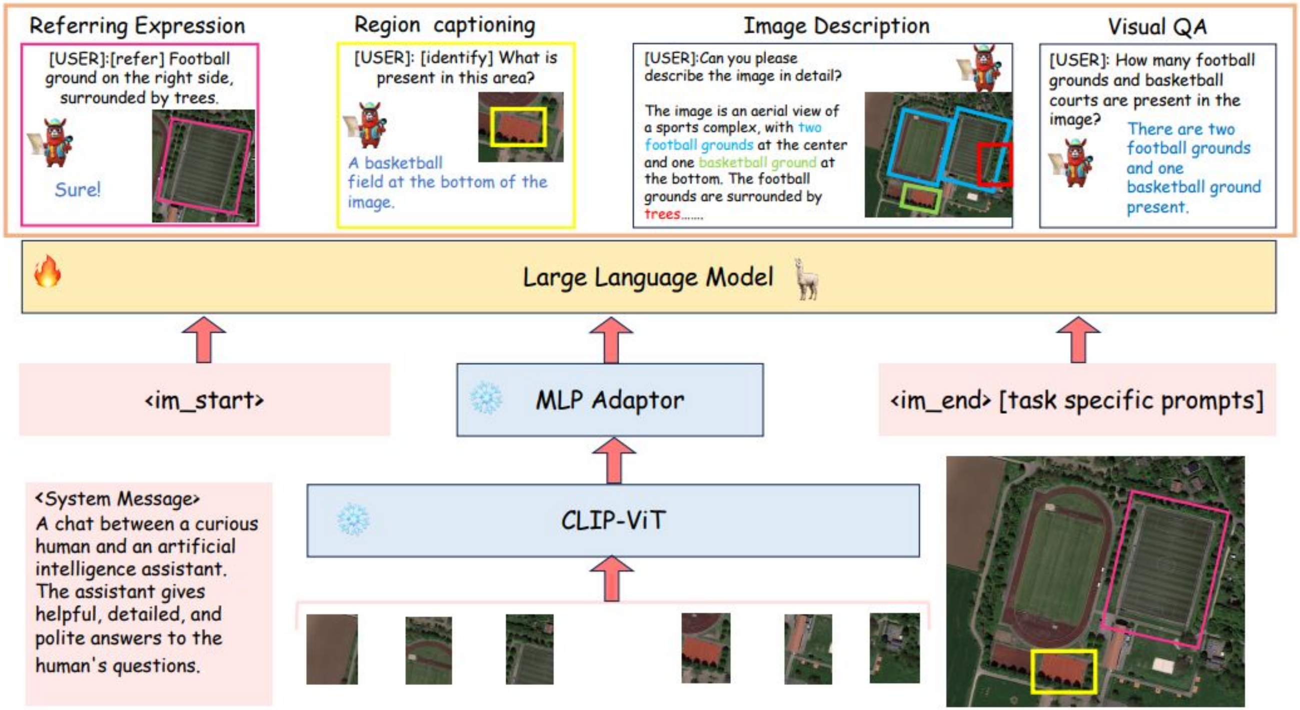

Referring Expression
[USER]:[refer] Football ground on the right side, surrounded by trees.
Sure!
Region captioning
[USER]: [identify] What is present in this area?
A basketball field at the bottom of the image.
Image Description
[USER]:Can you please describe the image in detail?
The image is an aerial view of a sports complex, with two football grounds at the center and one basketball ground at the bottom. The football grounds are surrounded by trees.......
Visual QA
[USER]: How many football grounds and basketball courts are present in the image?
There are two football grounds and one basketball ground present.
Large Language Model
<im_start>
MLP Adaptor
<im_end> [task specific prompts]
<System Message>
A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.
CLIP-ViT

Figure 2. An overview of GeoChat - the first grounded large vision-language model for remote sensing. Given an image input together with a user query, a visual backbone is first used to encode patch-level tokens at a higher resolution via interpolating positional encodings. A multi-layer perceptron (MLP) is used to adapt vision-tokens to language space suitable for input to a Large Language Model (Vicuna 1.5). Besides visual inputs, region locations can also be input to the model together with task-specific prompts that specify the desired task required by the user. Given this context, the LLM can generate natural language responses interleaved with corresponding object locations. GeoChat can perform multiple tasks as shown on top e.g., scene classification, image/region captioning, VQA and grounded conversations.

Fig. 12: GeoChat model.[5]

Reference

1. **Vision-Language Models for Vision Tasks: A Survey** Jingyi Zhang , Jiaxing Huang , Sheng Jin and Shijian Lu, https://arxiv.org/abs/2304.00685
2. **CLIP blog**, OpenAI, https://openai.com/research/clip
3. **Learning to Prompt for Vision-Language Models,** Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu, https://arxiv.org/abs/2109.01134
4. **Conditional Prompt Learning for Vision-Language Models**,Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu, https://arxiv.org/abs/2203.05557
5. **GeoChat : Grounded Large Vision-Language Model for Remote Sensing**, Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, Fahad Shahbaz Khan, https://arxiv.org/abs/2311.15826