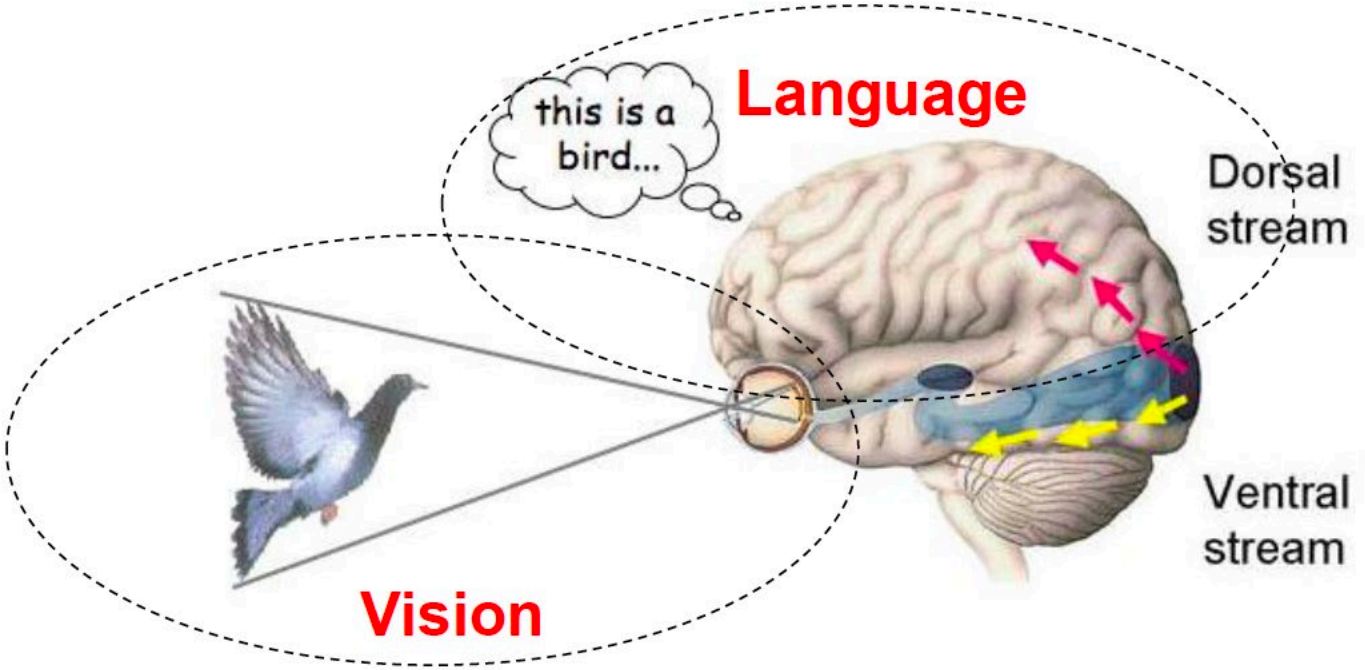
Prompt Pre-training with Over Twenty-thousand Classes for Open-Vocabulary Visual Recognition

Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, Xu Sun

renshuhuai007@gmail.com
https://renshuhuai-andy.github.io/

Background: Visual Recognition



Visual object recognition serves as a gateway **from vision to cognitive processes such as categorization, language and reasoning**.

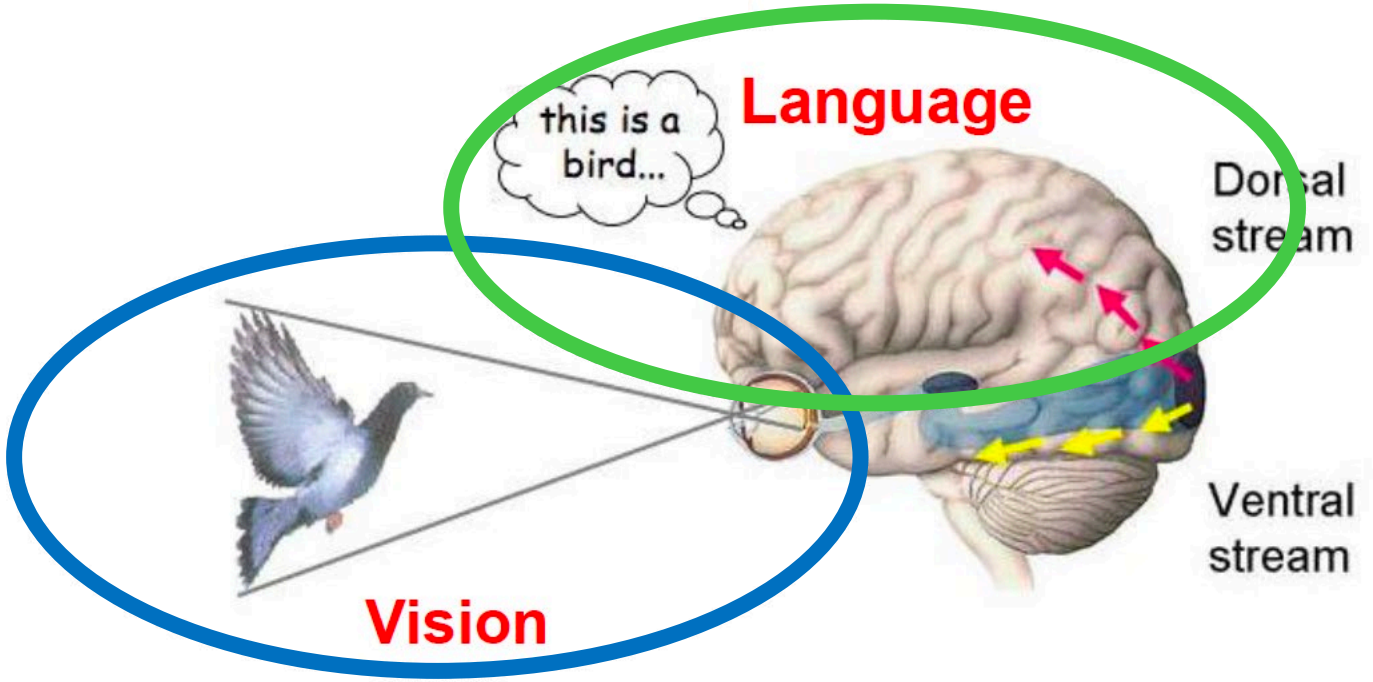
-***Object Recognition***. *John E. Hummel*

Visual Recognition: Observe visual object and map it with semantic concept

Background: Visual Recognition

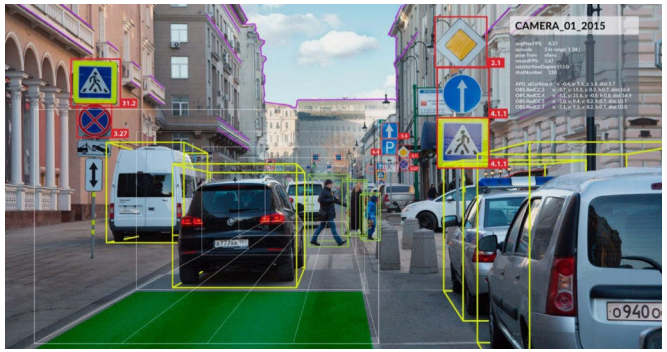
Visual Recognition: Observe visual object and map it with semantic concept

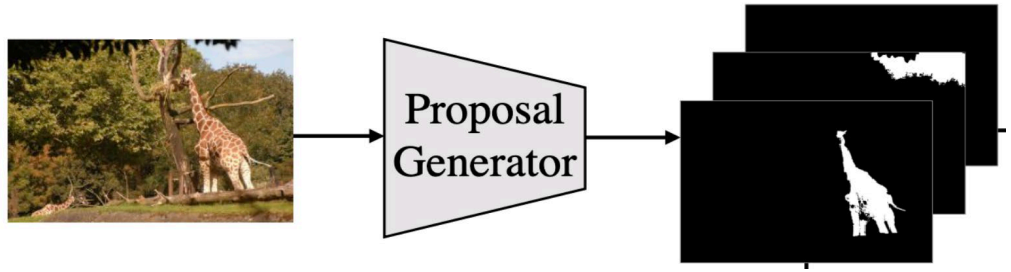
-> Visual recognition can be modeled as a **two-stage** task (**Observation** then **Classification**)



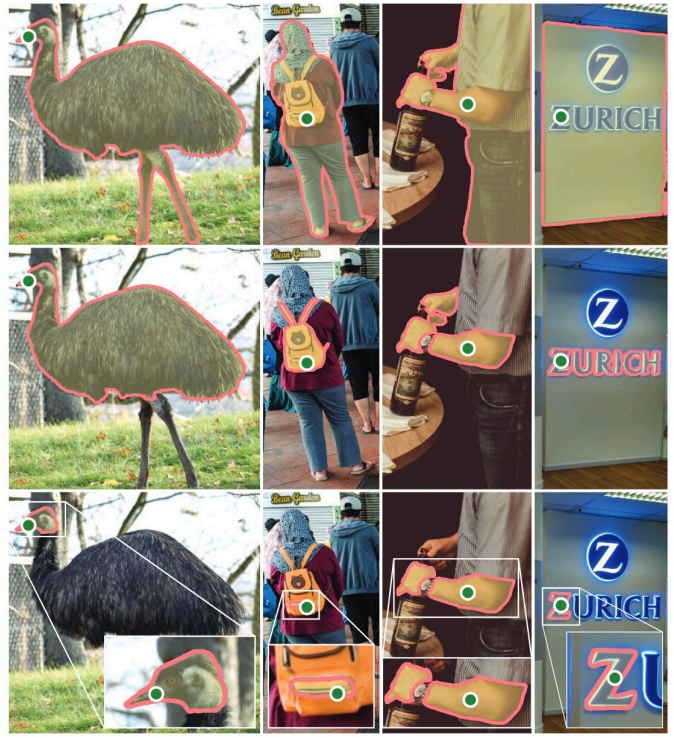
Background: Visual Recognition

Visual Recognition: **Observe visual object** and map it with semantic concept

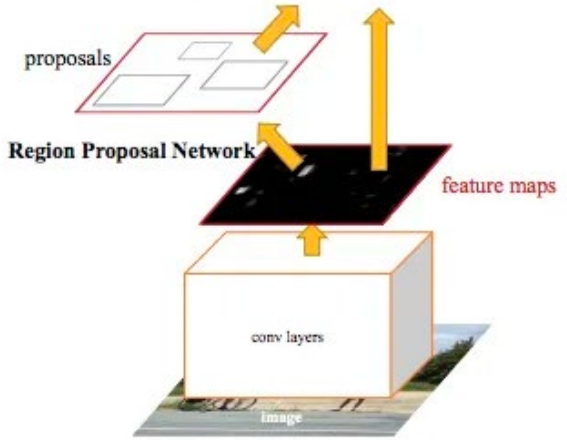




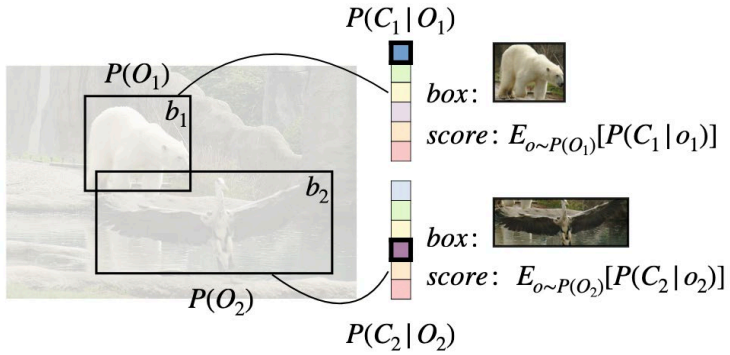
MaskFormer. 2022



SAM. 2023



Faster RCNN. 2015



CenterNet2. 2022

Generate region/mask proposal

Background: Visual Recognition

Visual Recognition: Observe visual object and **map it with semantic concept**

Supervised Learning

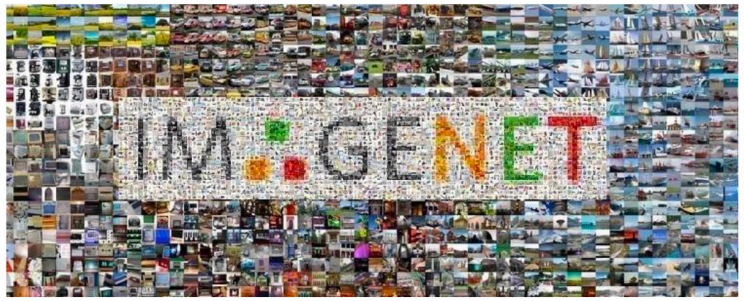
Image → Label (Concept)

( ) → "2" (Apple)

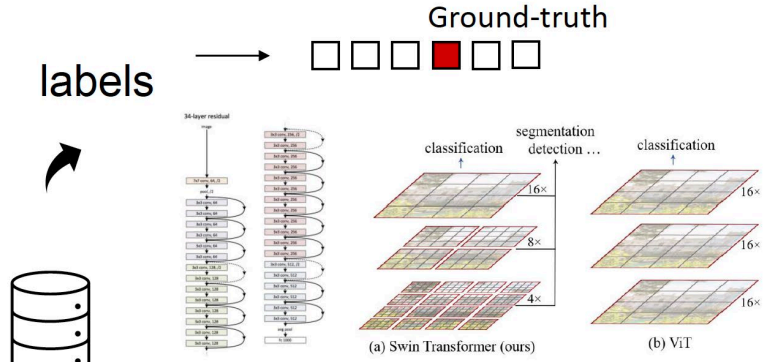
Map an image to a discrete label which is associated a visual concept

MNIST. LeCun *et al.*

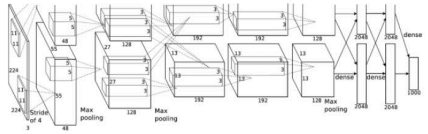
CIFAR-10. Krizhevsky *et al.*



ImageNet. Deng *et al.*



ResNet. He *et al.* **Swin**. Liu *et al.*



AlexNet. Krizhevsky *et al.*

Cons: Requires a lot of human effort & Limited number of categories & Label remap for different datasets

Background: Visual Recognition

Visual Recognition: Observe visual object and **map it with semantic concept**

Zero-Shot Learning (Canonical)

Image		Descriptions (Concept)
	→	Fruit, Red, Sphere (Apple)
	→	Fruit, Yellow (Orange)

Map an image to attribute composition of a visual concept

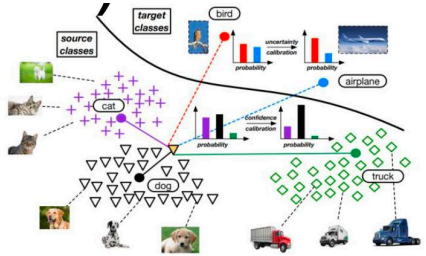
CUB-200-2011. Wah *et al.*

AwA2. Xian *et al.*

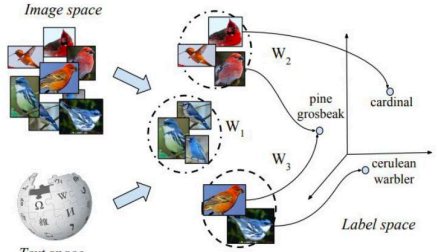
aPY. Farhadi *et al.*

Label & descriptions

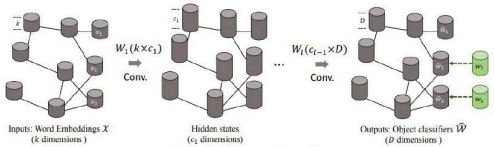
images



Liu *et al.*



Xian *et al.*



Wang *et al.*

Cons: Small scale with limited vocabulary

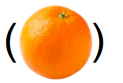
Background: Visual Recognition

Visual Recognition: Observe visual object and **map it with semantic concept**

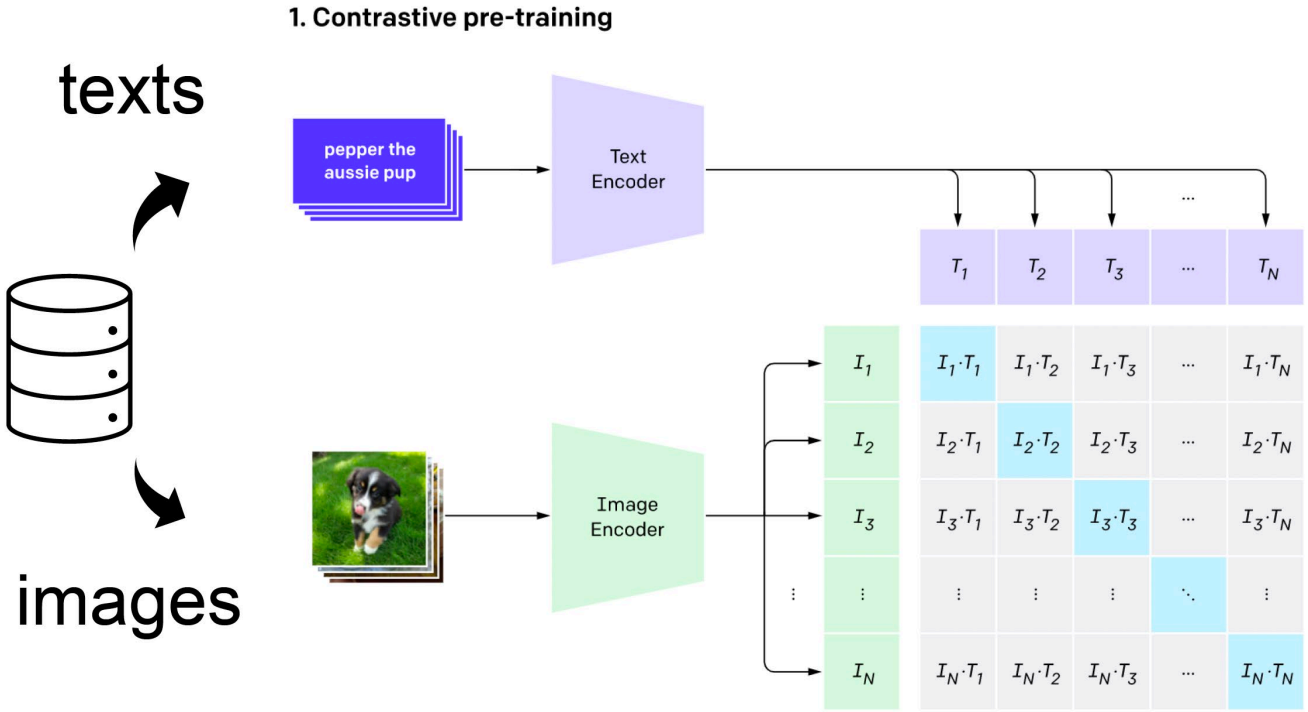
Contrastive Vision-Language Learning

Image ⟷ Text (Align)

() ⟷ A red apple

() ⟷ One yellow orange

Map an image to
its textual description



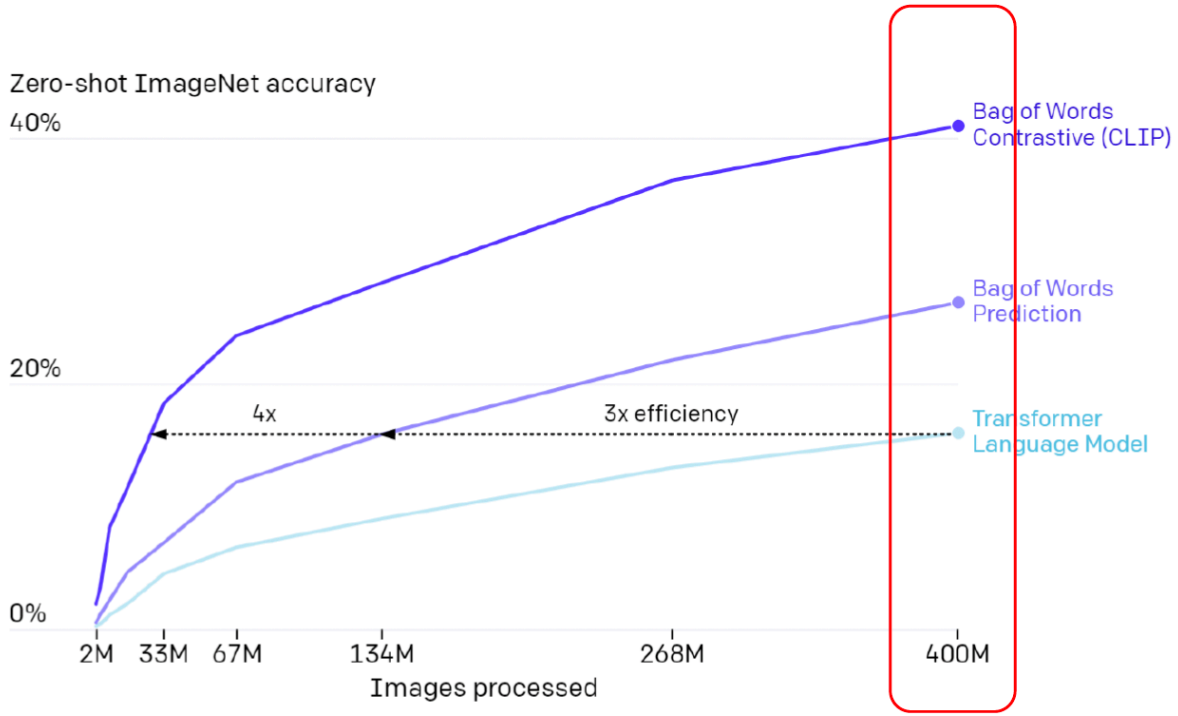
❖Learning from 400M web-crawled image-text pairs
❖Image recognition as an image-text matching problem

End-to-end learning on large-scale corpus

Background: Visual Recognition

Visual Recognition: Observe visual object and **map it with semantic concept**

Contrastive Vision-Language Learning



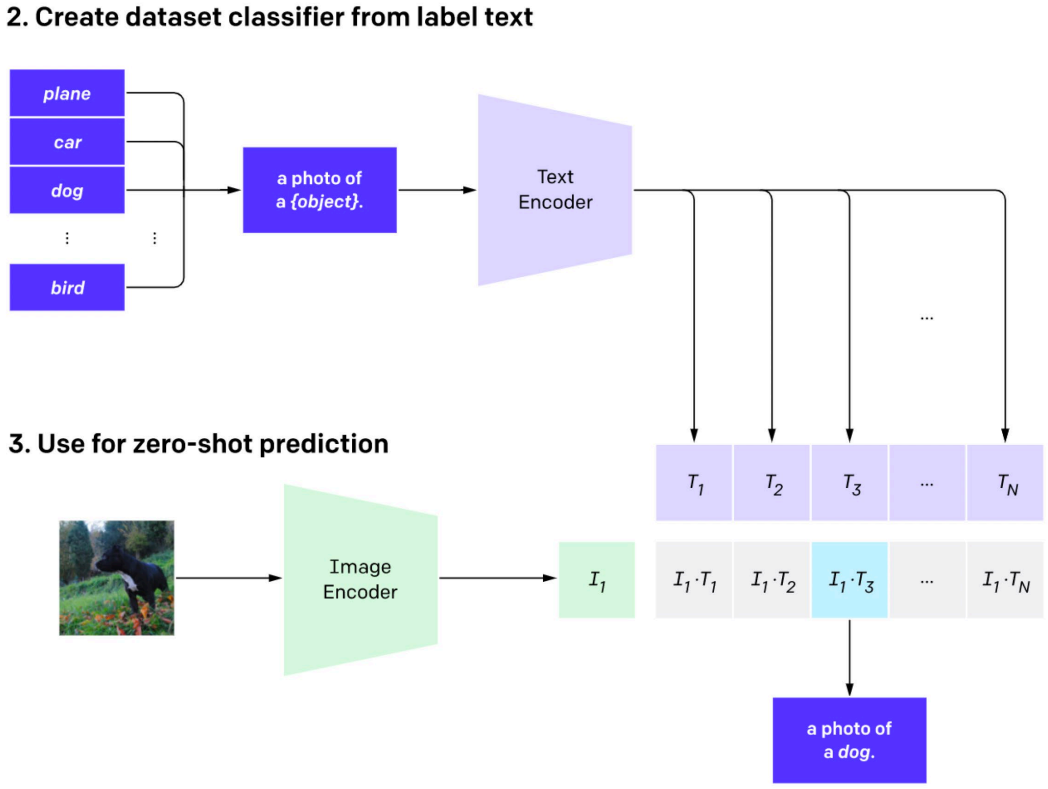
Contrastive learning is more effective than generative learning

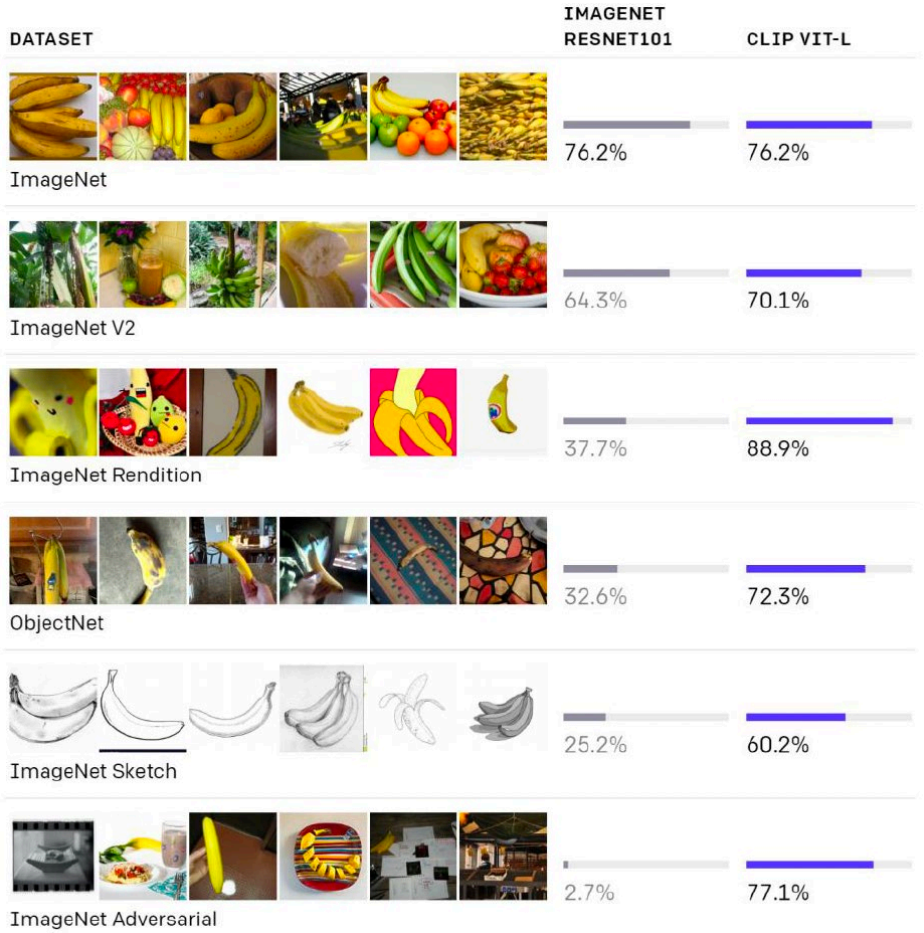
Data really matter

Background: Visual Recognition

Visual Recognition: Observe visual object and **map it with semantic concept**

Contrastive Vision-Language Learning





Profiting from the flexibility of natural language, CLIP open a new direction: open-vocabulary visual recognition

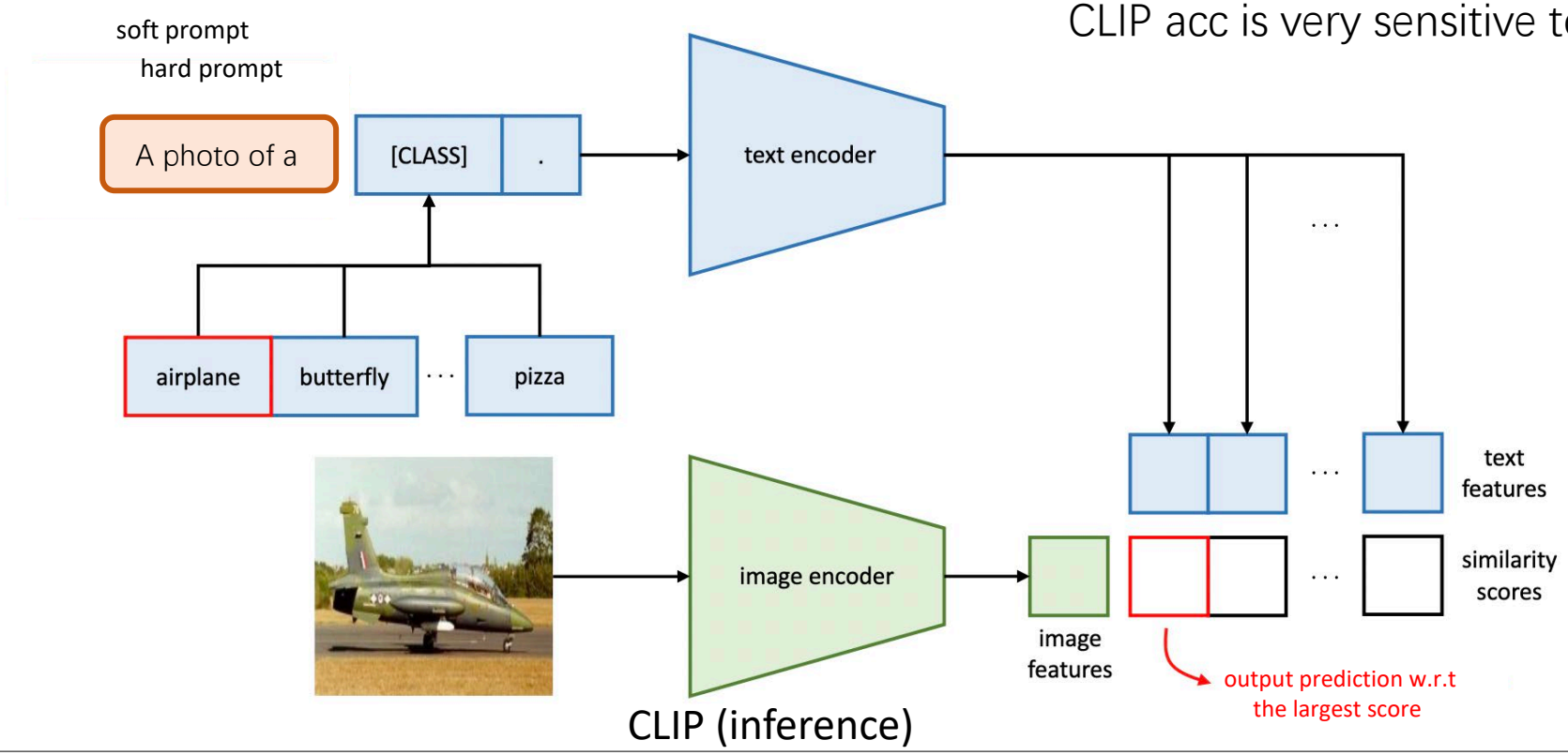
Background: Prompt Tuning for CLIP

How to fine-tune CLIP for a downstream dataset?

- Replace the hard prompt with a soft prompt (learnable token embeddings)
- Freeze both text and image encoder, only fine-tune the soft prompt
- Parameter-efficient (only update 0.012M parameters) and effective

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

CLIP acc is very sensitive to the prompt



CLIP (inference)

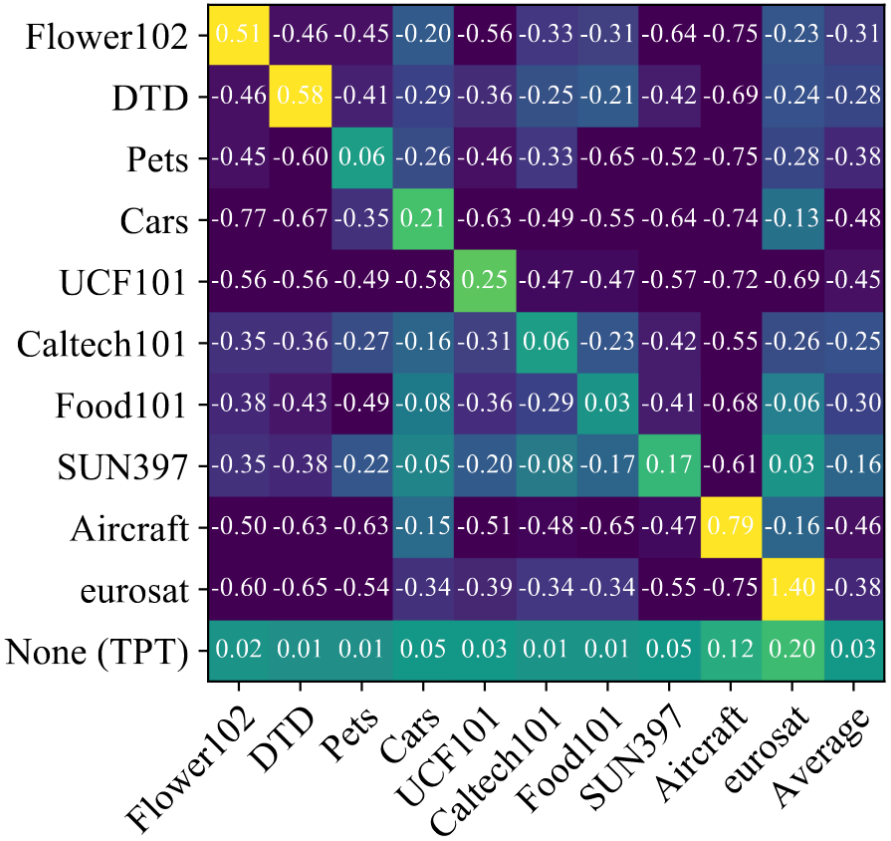
- Learning to Prompt for Vision-Language Models, IJCV 2022

Research Question

Traditional prompt tuning methods:

- Fine-tune on task-specific datasets with limited number of class labels
- Difficult to generalize to novel classes and across tasks

	DTD	Flowers102
Hard prompt	44.5	61.8
Prompt tuning on DTD	63	33.4

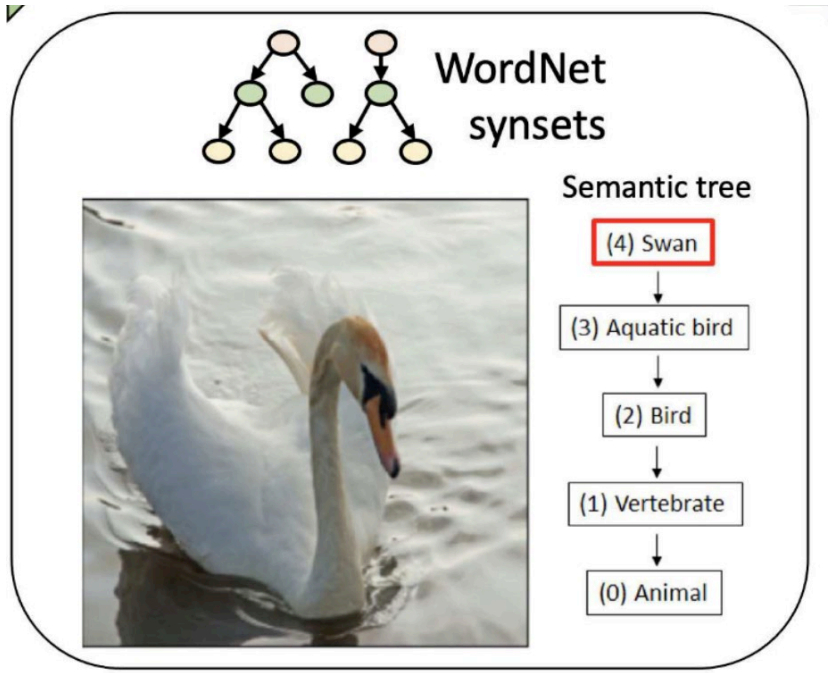


normalized relative improvement

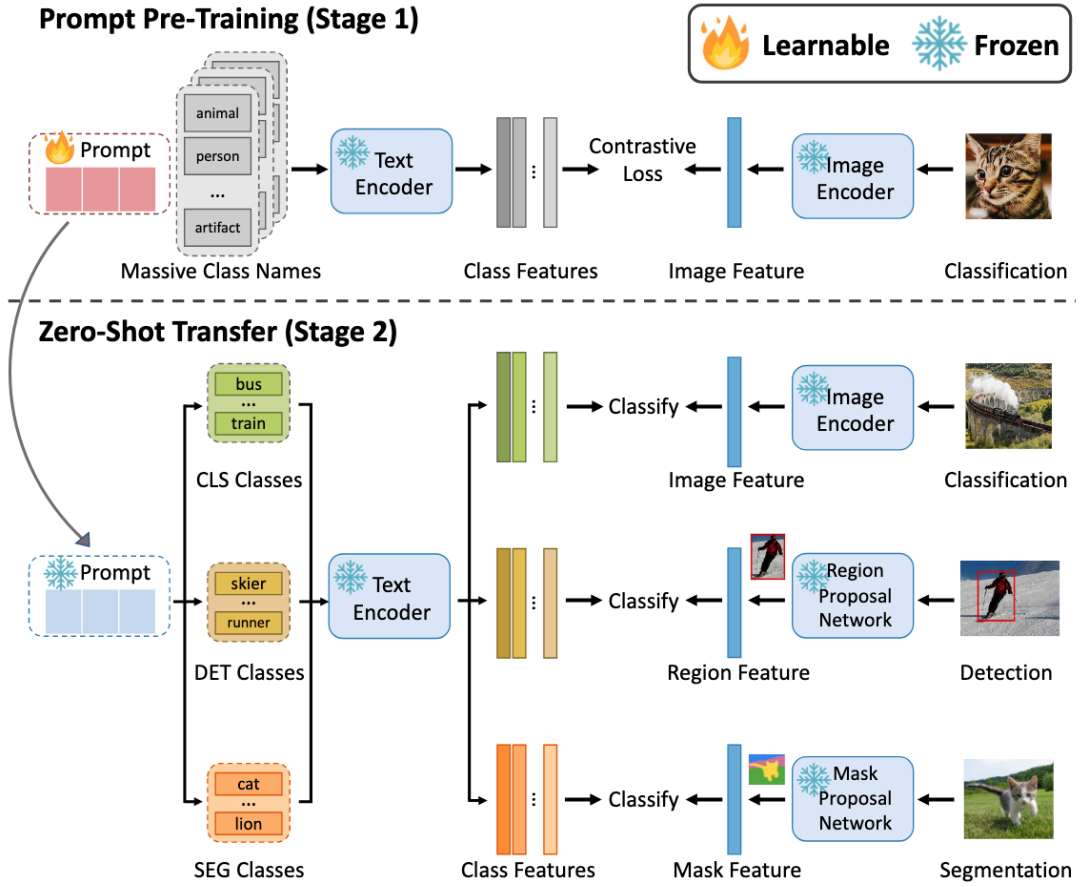
Research Question

Scaling up prompt learning on massive classes to condense semantic information for universal visual discrimination

- Pre-training soft prompt on large-scale dataset with massive classes (ImageNet-21k)
- Transfer to other datasets or tasks to conduct zero-shot inference



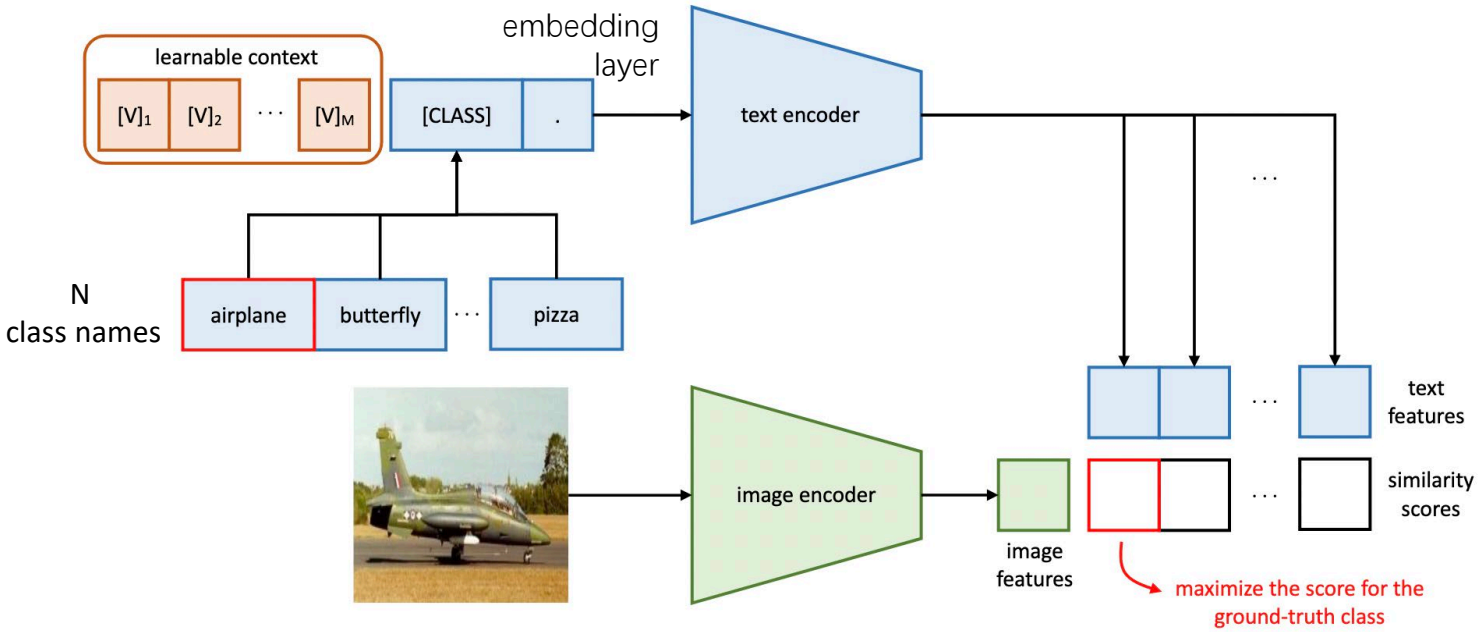
ImageNet-21k dataset

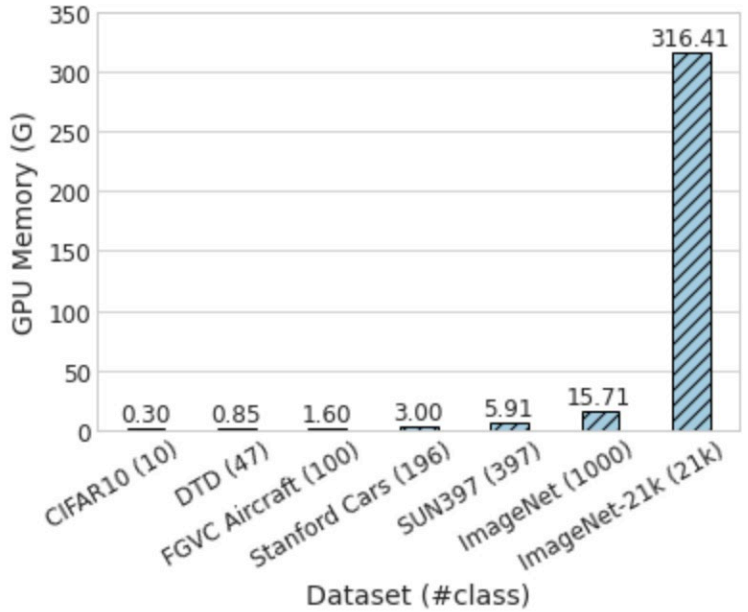


Research Question

Huge challenge of training cost:

- The loss is calculated at the output layer (last layer), The learnable parameters are at the embedding layer (first layer)
- For every class: need to store the activation of the whole text encoder, the gradient should pass through the whole text encoder
- When N is large, the states of forward and the gradient of backward will cost huge GPU memory, training is very slow
- pre-training prompts on ImageNet-21K requires over 300 GB GPU memory with CoOp





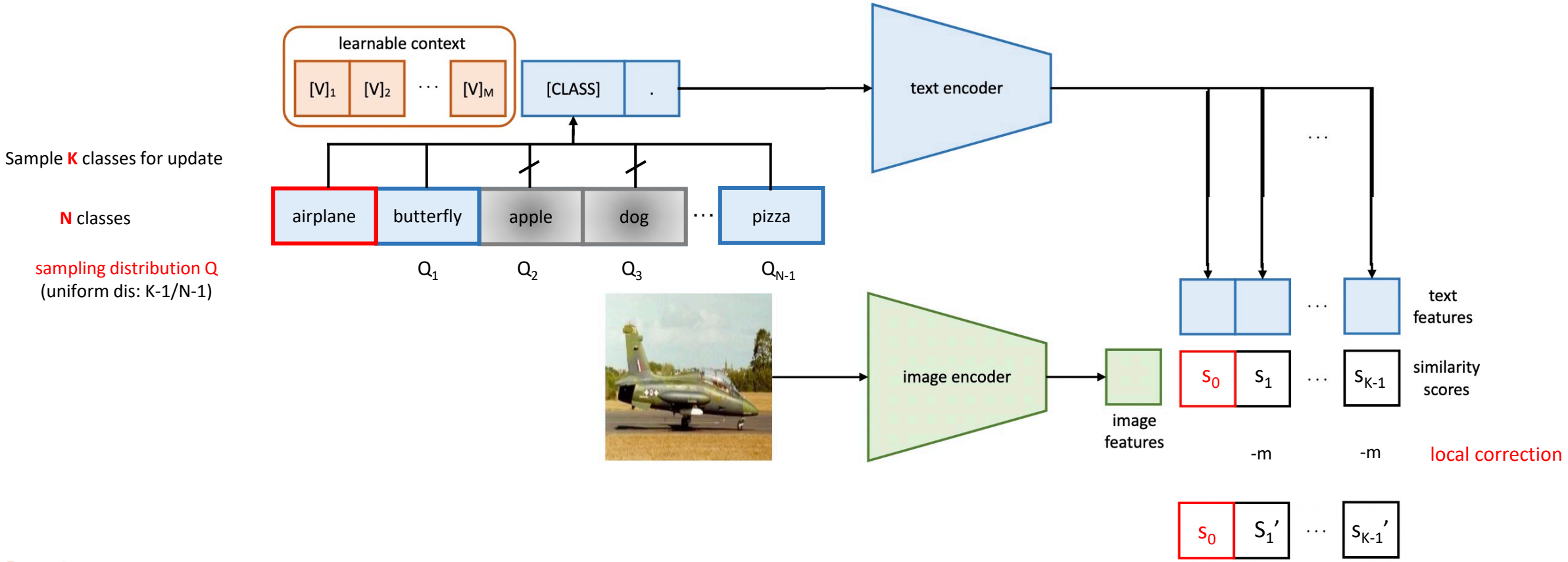
unaffordable training cost when the classes are scaling
(the actual cost will be larger due to PyTorch memory caching)

- Learning to Prompt for Vision-Language Models, IJCV 2022

Methodology: Prompt Pre-training (POMP)

Key technology: Local Contrast & Local Correction

- For a dataset with N classes, only sample K (<N) classes base on a sampling distribution Q for update
- Local correction for negative scores: $s_i' = s_i - m$



Methodology: Prompt Pre-training (POMP)

Key technology: Local Contrast & Local Correction

- For a dataset with N classes, only sample K (<N) classes from a sampling distribution Q for update
- Local correction for negative scores: $s_i = s_i - m$

Final prediction probability of POMP:

$$\tilde{P}(y \mid \mathbf{x}; \mathbf{\Theta}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_y^{(\mathbf{\Theta})}/\tau)}{\exp(\mathbf{x}^\top \mathbf{w}_y^{(\mathbf{\Theta})}/\tau) + \sum_{i \sim \mathcal{N}} \exp(\mathbf{x}^\top \mathbf{w}_i^{(\mathbf{\Theta})}/\tau + m_i)}$$

Local correction term:

$$m_i = -\log\left((K-1)/(N-1)\right)$$

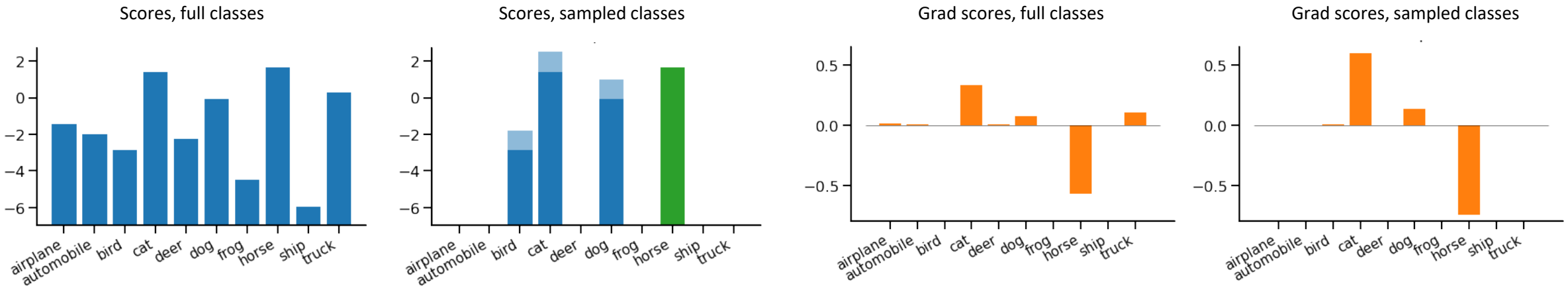
encourages the positive logit to be larger than the negative logits by a certain margin, resulting in a more stringent decision boundary

$$C_+ : \mathbf{x}^\top \mathbf{w}_y^{(\mathbf{\Theta})}/\tau \geq \mathbf{x}^\top \mathbf{w}_i^{(\mathbf{\Theta})}/\tau + m_i, \quad i \neq y.$$

Methodology: Prompt Pre-training (POMP)

Key technology: Local Contrast & Local Correction

- For a dataset with N classes, only sample K (<N) classes from a sampling distribution Q for update
- Local correction for negative scores: $s_i = s_i - m$

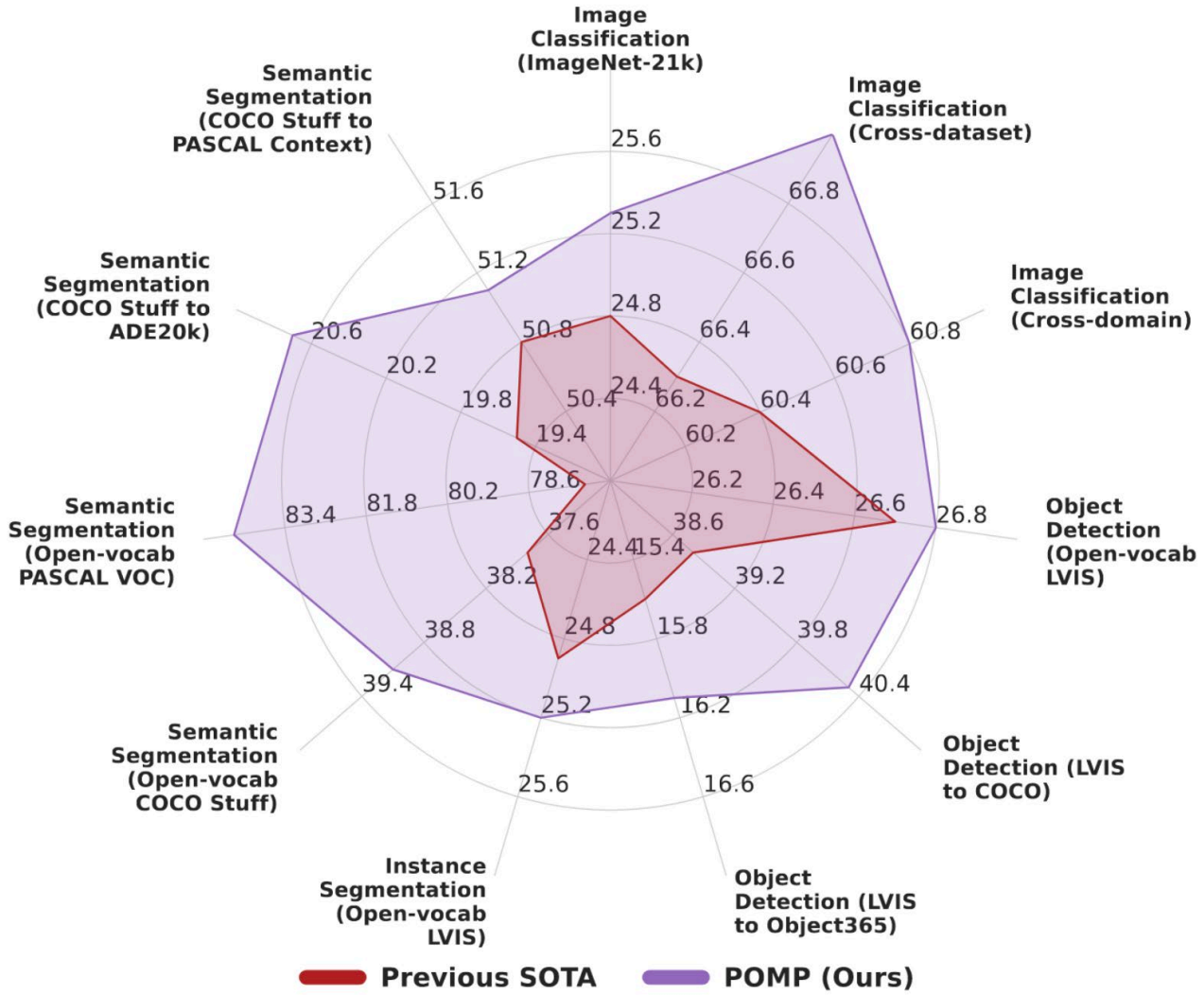


Each neg class score (blue) has two parts:

1. the part from the model (dark blue)
2. a part from the correction term –m (light blue)

Enlarge the gradient of pos and sampled neg class

Experiment

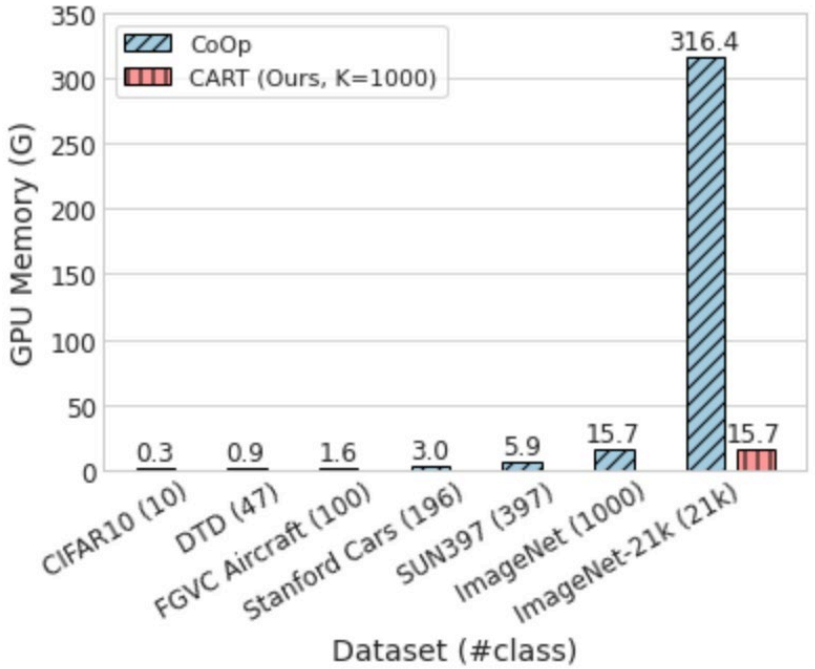


"Scaling up prompt learning on ImageNet-21K achieves SOTA on 21 downstream datasets."

Experiment

Prompt Pre-training on ImageNet-21k

- Shots=16, epoch=20, bsz=32
- Prompt length=16



Method	ResNet50	ViT-B/32	ViT-B/16
ZeroshotCLIP Radford et al. [2021]	17.5	19.8	21.8
Prompt Ensemble Radford et al. [2021]	18.8	20.9	23.5
CoOp Zhou et al. [2022a]	16.6	18.1	20.8
MaPLe Khattak et al. [2022]	-	21.6	24.2
Linear Probing Radford et al. [2021]	6.5	18.2	20.9
VPT Derakhshani et al. [2022]	-	21.8	24.8
POMP (Ours)	**20.2**	**22.2**	**25.3**

- POMP significantly reduces the GPU memory, making prompt tuning on massive classes possible.
- Surpasses ZeroshotCLIP by **3.5%** and Linear Probe by **4.4%**.

Experiment

北京大学 PEKING UNIVERSITY

Image Classification: Cross-dataset & Cross-domain Transfer

- Cross-dataset: Zero-shot inference on other 10 target datasets
  - outperforming CoOp by **3.1%** and surpassing the previous SOTA in **7/10** datasets
  - provide a more expressive context for fine-grained / long-tail visual concepts such as specific **objects** and **scenes**
  - **StanfordCars** (+1.2%) and **Aircraft** (+0.9%), as well as **SUN397** (+0.7%) and **EuroSAT** (+4%).
- Cross-domain: zero-shot inference on ImageNet from other 4 domains
  - More robust to domain shift, achieves a new SOTA with 60.8%

	Target (cross-dataset)											Target (cross-domain)				
	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	***Average***	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R	***Average***
CoOp Zhou et al. [2022a]	93.7	89.1	64.5	68.7	85.3	18.5	64.2	41.9	46.4	66.6	63.9	64.2	48.0	49.7	75.2	59.3
CoCoOp Zhou et al. [2022c]	94.4	90.1	65.3	71.9	86.1	22.9	67.4	45.7	45.4	68.2	65.7	64.1	48.8	50.6	76.2	59.9
LASP Bulat and Tzimiropoulos [2022]	94.5	89.4	64.8	70.5	86.3	23.0	67.0	45.5	48.3	68.2	65.8	63.8	49.0	50.7	77.1	60.1
VPT Derakhshani et al. [2022]	93.7	**90.6**	65.0	70.9	86.3	24.9	67.5	46.1	45.9	68.7	66.0	**64.2**	49.2	51.3	77.0	60.4
MaPLe Khattak et al. [2022]	93.5	90.5	65.6	72.2	86.2	24.7	67.0	**46.5**	48.1	**68.7**	66.3	64.1	49.2	50.9	77.0	60.3
POMP (Ours)	**95.0**	89.5	**66.8**	**72.4**	**86.3**	**25.6**	**67.7**	46.2	**52.1**	68.5	**67.0**	63.8	**49.8**	**51.6**	**77.9**	**60.8**

Experiment

Image Classification: Training efficiency

- Tuning on ImageNet-1K
- Shots=16, epoch=50, bsz=32, prompt length=16

Method	Acc. (%)	GPU Mem. (GB)	Training Time (h)
CoOp	71.9	28.2	5.9
CoCoOp	70.1	28.3	27.5
POMP ($K = 128$)	71.2	5.3	2.7
POMP ($K = 256$)	71.4	8.8	3.3
POMP ($K = 512$)	71.6	15.9	4.2

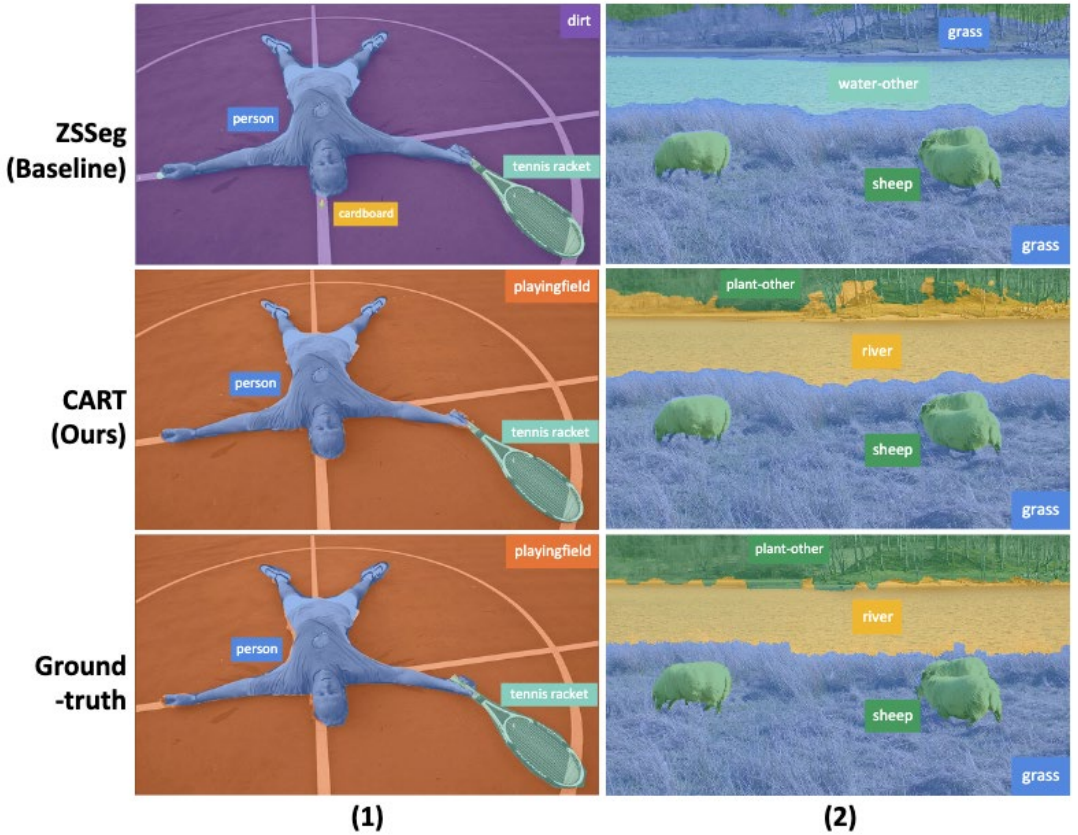
- POMP (K = 128) achieves competitive accuracy on ImageNet-1K while using less than **19%** of GPU memory and **50%** of training time

Experiment

Semantic Segmentation: Open-vocab COCO Stuff and Open-vocab Pascal VOC

- Fine-tune visual backbone on base classes, zero-shot inference on novel classes

Method	Open-Vocab COCO Stuff			Open-Vocab Pascal VOC		
	hIoU	mIoU		hIoU	mIoU	
		seen	unseen		seen	unseen
SPNet Xian et al. [2019]	16.8	20.5	14.3	21.8	73.3	15.0
ZS3 Bucher et al. [2019]	15.0	34.7	9.5	28.7	77.3	17.7
CaGNet Gu et al. [2020]	18.2	35.5	12.2	39.7	78.4	25.6
ZegFormer Ding et al. [2021]	34.8	36.6	33.2	73.3	86.4	63.6
ZSSeg Xu et al. [2021]	37.8	39.3	36.3	77.5	83.5	72.5
POMP (Ours)	**39.1**	**39.9**	**38.2**	**84.4**	**93.6**	**76.8**



compared to SOTA: **+1.3 hIoU** on open-vocab COCO Stuff, **+6.9 hIoU** on open-vocab Pascal VOC

stronger ability to **distinguish background categories**

Experiment

北京大学 PEKING UNIVERSITY

Semantic Segmentation: Cross-Dataset Transfer

- Fine-tune visual backbone on Standard COCO Stuff, zero-shot inference on ADE20k & PASCAL Context

Method	Source Dataset: Standard COCO Stuff			Target Dataset: ADE20K			Target Dataset: PASCAL Context		
	mIoU	fwIoU	pACC	mIoU	fwIoU	pACC	mIoU	fwIoU	pACC
ZSSeg	40.8	49.0	62.7	19.5	48.7	60.0	50.8	64.1	75.7
POMP (Ours)	**41.1**	**49.2**	**62.9**	**20.7**	**51.5**	**63.7**	**51.1**	**65.4**	**76.1**

compared to SOTA: **+0.3 mIoU** on standard COCO Stuff, **+1.2 mIoU** on ADE20k, **+0.3 mIoU** on PASCAL Context

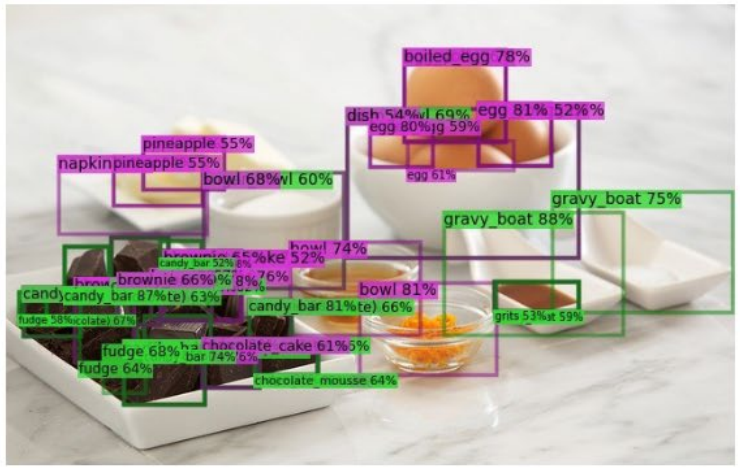
Experiment

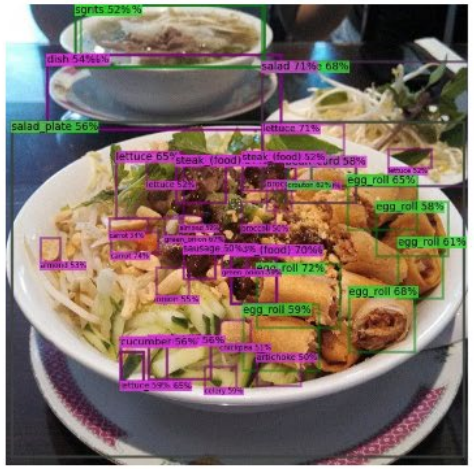
Object Detection: Open-vocab LVIS

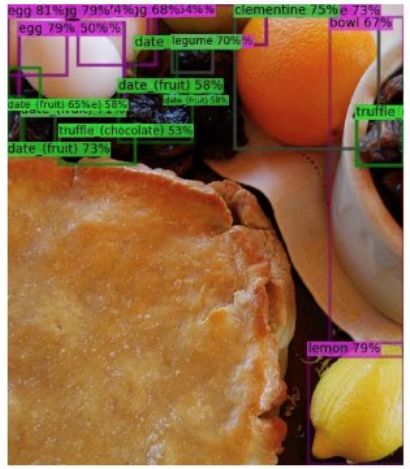
- Fine-tune visual backbone on LVIS-base (#class=866), zero-shot inference on LVIS-novel (#class=337)

Method	Detection				Instance segmentation			
	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD Gu et al. [2021]	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
DetPro Du et al. [2022]	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
PromptDet Feng et al. [2022]	-	-	-	-	21.4	23.3	29.3	25.3
Detic Zhou et al. [2022b]	26.7	36.4	40.3	36.3	24.9	32.5	35.6	32.4
POMP (Ours)	**26.8**	36.4	40.4	36.2	**25.2**	33.0	35.6	32.7

compared to SOTA:
+0.3 APr on instance segmentation,
+0.1 APr on object detection







Is able to detect the **novel classes**, except for the **base classes**

Experiment

Object Detection: Cross-Dataset Transfer

- Fine-tune visual backbone on Standard LVIS, zero-shot inference on COCO & Object 365

Method	Source Dataset: Standard LVIS						Target Dataset: COCO						Target Dataset: Objects365					
	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ViLD* Gu et al. [2021]	27.5	41.8	29.3	20.6	35.9	43.4	34.1	52.3	36.5	21.6	38.9	46.1	11.5	17.8	12.3	4.2	11.1	17.8
DetPro Du et al. [2022]	28.4	42.9	30.3	21.0	36.7	44.1	34.9	53.8	37.4	22.5	39.6	46.3	12.1	18.8	12.9	4.5	11.5	18.6
Detic Zhou et al. [2022b]	36.8	50.7	38.6	26.1	46.7	51.7	38.8	56.0	41.9	25.6	42.2	50.0	15.6	22.1	16.8	6.1	15.6	23.8
POMP (Ours)	**37.2**	**51.1**	**39.3**	**26.5**	**47.2**	**52.6**	**40.3**	**57.9**	**43.6**	**28.3**	**43.9**	**50.6**	**16.1**	**22.9**	**17.3**	**6.2**	**16.3**	**24.7**

compared to SOTA: **+0.4 AP** on standard LVIS, **+1.5 AP** on COCO, **+0.5 AP** on Object 365

Ablation & Analysis

Method	ImageNet-21K	Cross-dataset (10 Avg.)	Cross-domain (4 Avg.)
POMP ($K = 100$)	24.1	65.5	59.5
POMP ($K = 500$)	24.9	66.5	60.0
POMP ($K = 1000$)	25.3	67.0	60.8
- local correction	25.0 **(-0.3)**	65.8 **(-1.2)**	59.8 **(-1.0)**

	m=0	m=-0.5	m=-1	m=-1.5
K=319	65.2	65.7	66.2	**66.5 (Ours)**
K=1000	65.8	66.1	**67.0 (Ours)**	66.4

varying m values of {0, -0.5, -1, -1.5} and report the cross-dataset accuracy

- Local contrast:
  the performance of POMP improves as K increases
  balances accuracy and cost by adjusting K

- Local correction:
  significantly improves the generalization of the pre-trained prompt
  Our adaptive margin approach achieves optimal performance under different K values

Ablation & Analysis

Alignment: the image feature and its ground-truth class feature are supposed to stay closed

$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\in\mathcal{D}} \left\| \mathbf{x} - \mathbf{w}_y^{(\Theta)} \right\|^2$$

Uniformity: all the class features should be uniformly distributed to preserve maximal information and make the categories more distinguishable

$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{\substack{1\leqslant i,j\leqslant N,\\ i\neq j}} \exp(-2\|\mathbf{w}_i^{(\Theta)} - \mathbf{w}_j^{(\Theta)}\|^2)$$

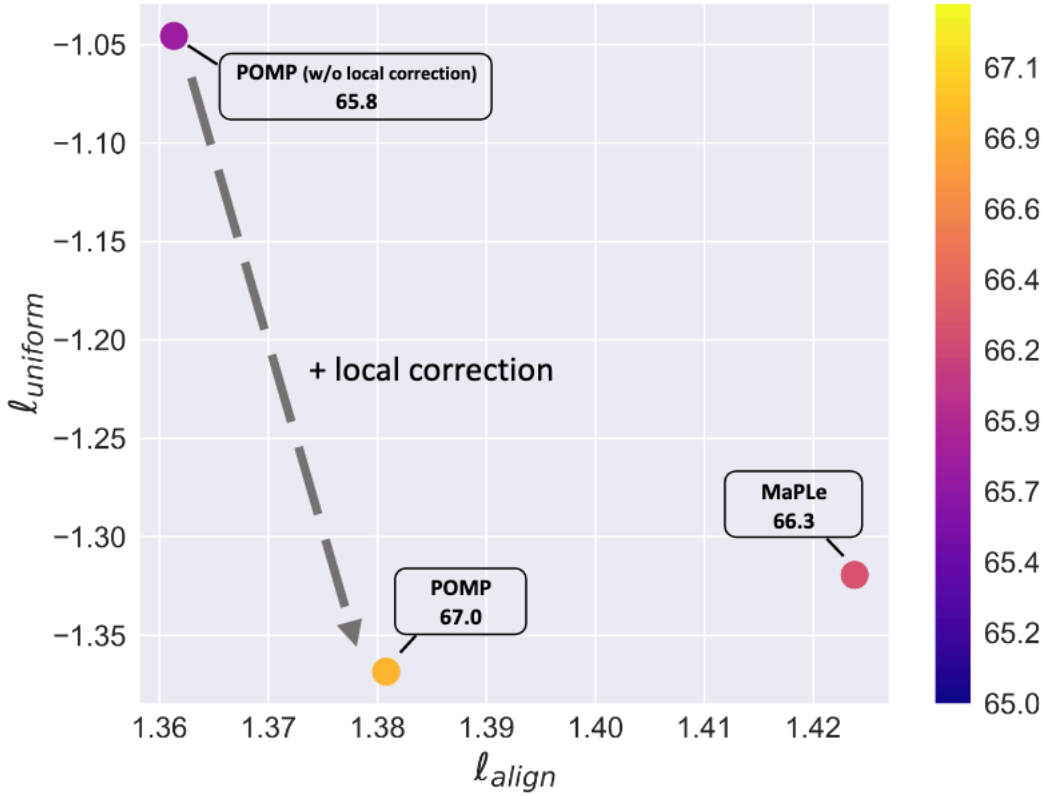


Figure 5: ℓ_{align} and ℓ_{uniform} of POMP. For both measures, lower numbers are better. The color of circles and the numbers in the boxes denote the average accuracy of cross-dataset evaluation on 10 datasets (higher is better).

With local correction: POMP significantly reduces the uniformity loss at only a slight expense of alignment

- Delving into the Openness of CLIP, 2022

Summary

Target:

- Scaling up prompt tuning on massive classes to to condense semantic information for universal visual discrimination

Method:

- Local contrast: sample class from a distribution Q for update, and add a local correction to enlarge the margin

Achievement:

- Enable prompt tuning on massive classes
- Achieve comparable accuracy on ImageNet-1k with 19% GPU memory and 50% training time
- Achieve new SOTAs on open-vocab various visual recognition datasets and tasks

Thanks

Q&A

Paper: https://arxiv.org/abs/2304.04704
Code: https://github.com/amazon-science/prompt-pretraining