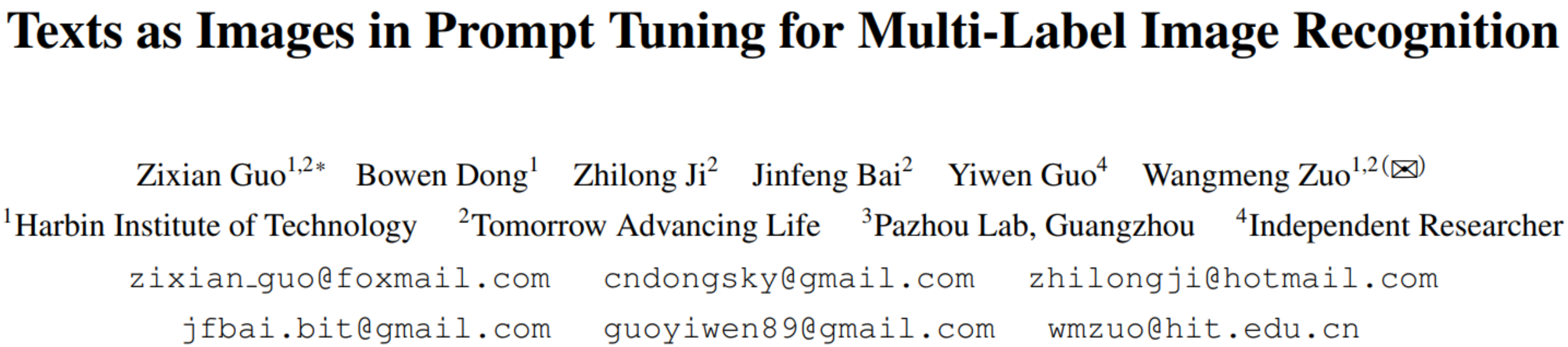


Texts as Images in Prompt Tuning for Multi-Label Image Recognition

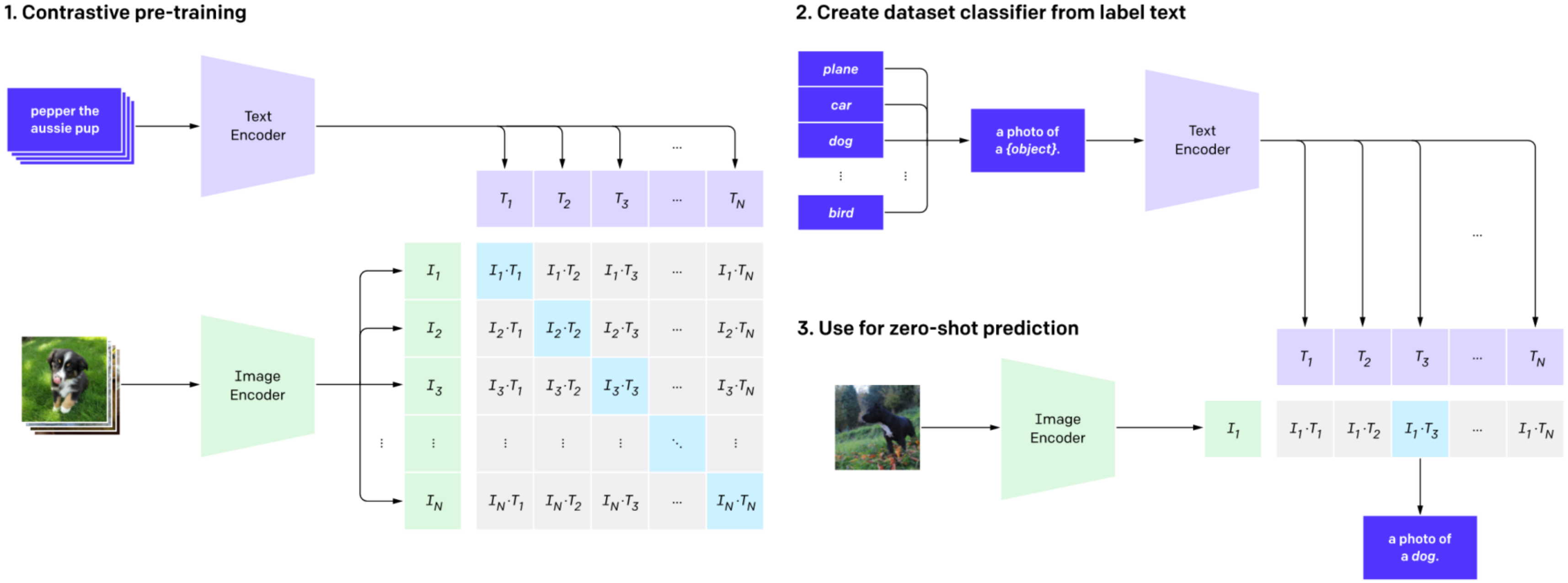
Zixian Guo[1,2]* Bowen Dong[1] Zhilong Ji[2] Jinfeng Bai[2] Yiwen Guo[4] Wangmeng Zuo[1,2](✉)
[1]Harbin Institute of Technology [2]Tomorrow Advancing Life [3]Pazhou Lab, Guangzhou [4]Independent Researcher
zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com
jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn

CVPR 2023

- □ 作者介绍
- □ 研究背景
- □ 本文方法
- □ 实验效果
- □ 总结反思

研究背景

□ CLIP尝试了视觉-语言对齐预训练，利用文本的结构信息提高泛化性



研究背景

- □ CLIP Prompt tuning（with labeled images）
  - ⊙ Context Optimization
  - ⊙ Conditional Context Optimization

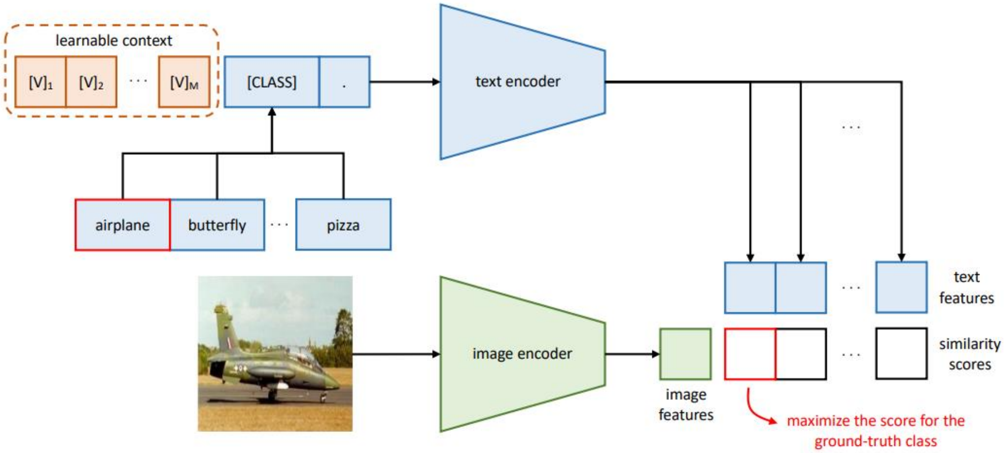


Figure 2: Overview of context optimization (CoOp).

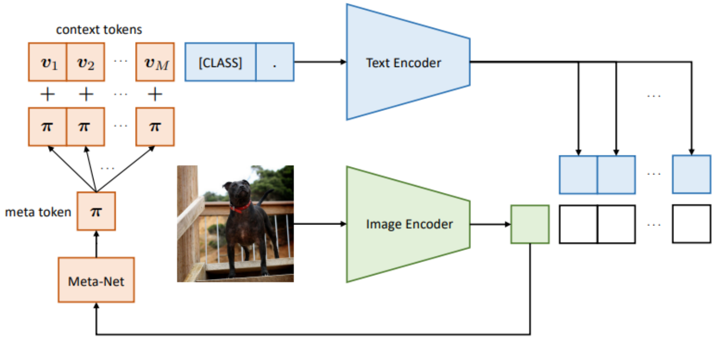


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

研究动机

- CLIP Prompt tuning without image
- 相比图像数据，语言数据丰富、容易获得、容易过滤处理，因此在prompt tuning阶段能否仅使用语言信息？
- 合理性在于，CLIP已经实现了视觉-语言特征对齐，(text + text encoder)与(image + image encoder)可以相互替换。当目标数据集改变，只要把语言特征调整到目标域，再替换image encoder即可提升性能。

文本描述
+text encoder

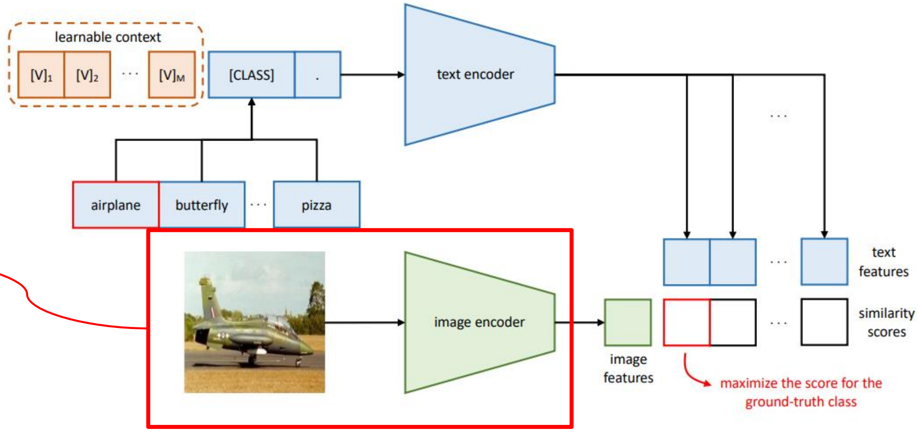


Figure 2: Overview of context optimization (CoOp).

- □ 作者介绍
- □ 研究背景
- □ 本文方法
- □ 实验效果
- □ 总结反思

Architecture

- Overview

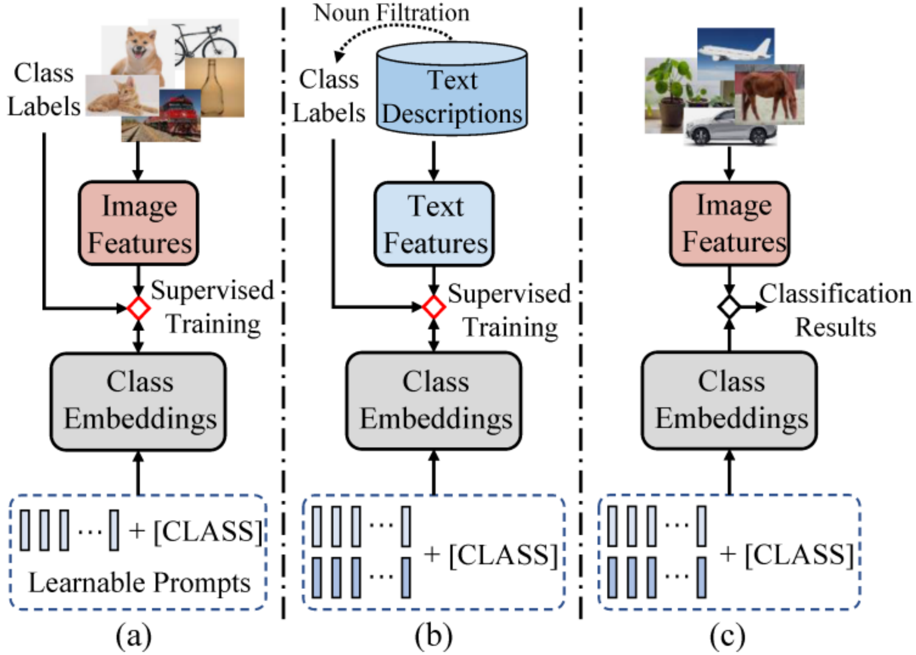


Figure 1. A comparison between prompting from images and our text-as-image (TaI) prompting. (a) Prompting from images (*e.g.*, [41]) uses labeled images of task categories to learn the text prompts. Instead, (b) our TaI prompting learn the prompts with easily-accessed text descriptions containing target categories. (c) After training, the learned prompts in (a) or (b) can be readily applied to test images.

Architecture

- Training: Text as Image prompt tuning
- Testing : Replace text encoder with image encoder

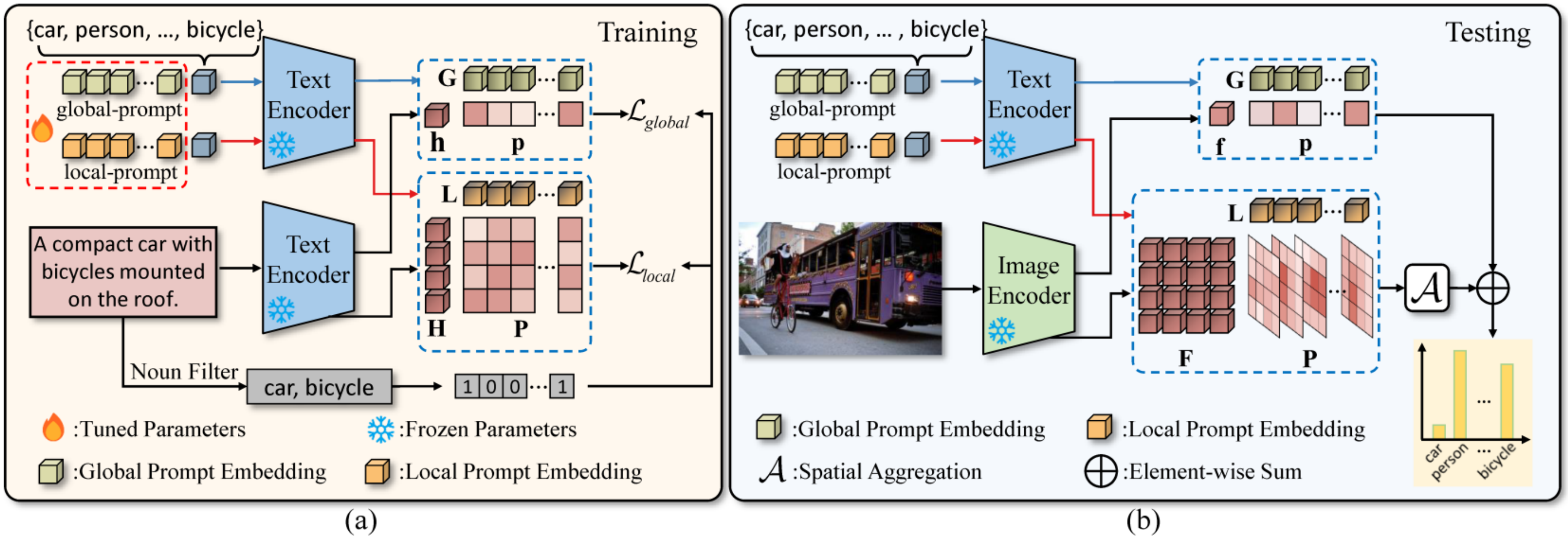


Figure 2. Training and testing pipeline of our proposed Text-as-Image (TaI) prompting, where we use text descriptions instead of labeled images to train the prompts. (a) During training, we use two identical text encoders from pre-trained CLIP to extract the global & local class embeddings (**G**&**L**) and overall & sequential text embeddings (**h**&**H**) respectively from the prompts and text description. The corresponding cosine similarity (**p**&**P**) between the embeddings are guided by the derived pseudo labels with ranking loss. (b) During testing, we replace the input from text descriptions to images. The global and local class embeddings can discriminate target classes from global & local image features (**f**&**F**). The final classification results are obtained by merging the scores of the two branches.

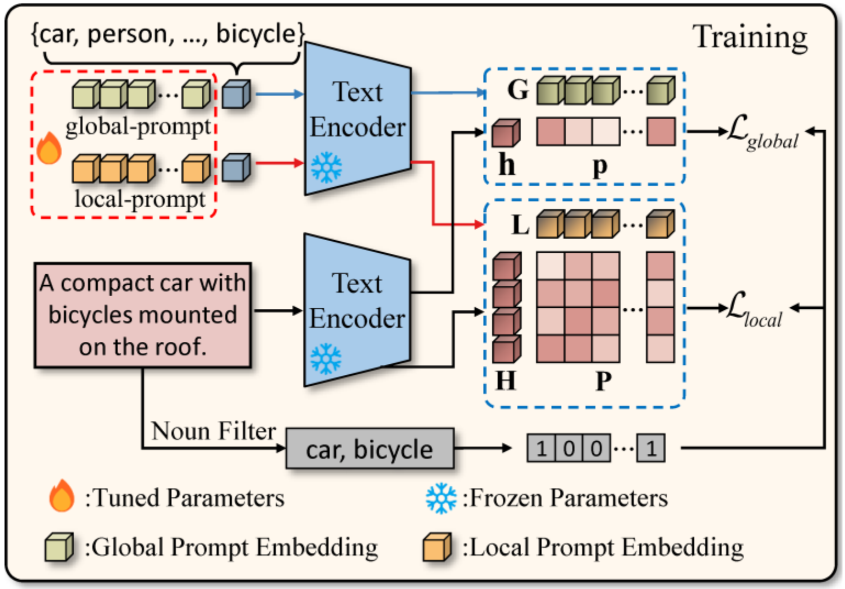
Dual grained prompt

- Prompt definition
  - V： learnable prompt
  - S： class

$$t_i^G = [v_1, v_2, v_3, ..., v_M, s_i],$$
$$t_i^L = [v'_1, v'_2, v'_3, ..., v'_M, s_i],$$

- CLIP： 对视觉特征做pooling
  对语言特征取class token

- 本文增加了细粒度的对齐



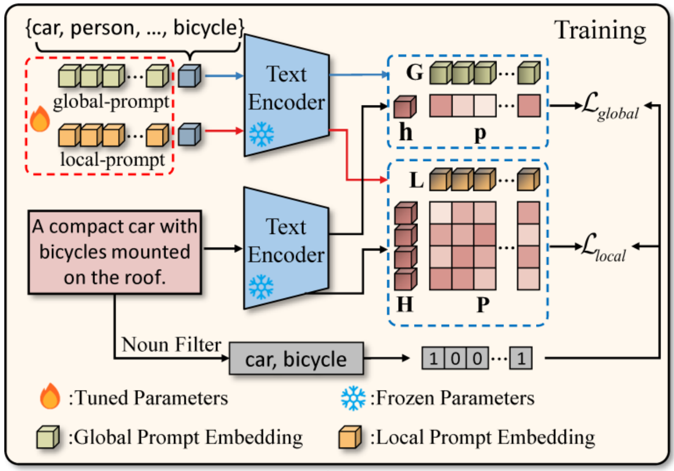
(a)

Objective

- 排序损失

$$\boldsymbol{p}_i = \langle \boldsymbol{u}, \boldsymbol{G}_i \rangle, \boldsymbol{P}_{ij} = \langle \boldsymbol{U}_j, \boldsymbol{L}_i \rangle$$

$$\boldsymbol{p}'_i = \sum_{j=1}^{N} \frac{\exp(\boldsymbol{P}_{ij}/\tau_s)}{\sum_{j=1}^{N} \exp(\boldsymbol{P}_{ij}/\tau_s)} \cdot \boldsymbol{P}_{ij}$$



(a)

$$\mathcal{L}_{global} = \sum_{i\in\{c+\}} \sum_{j\in\{c-\}} \max(0, m - \boldsymbol{p}_i + \boldsymbol{p}_j),$$
$$\mathcal{L}_{local} = \sum_{i\in\{c+\}} \sum_{j\in\{c-\}} \max(0, m - \boldsymbol{p}'_i + \boldsymbol{p}'_j)$$

$$p(y = i|\boldsymbol{x}) = \frac{\exp(\langle \mathrm{Enc}_{\mathrm{T}}(\boldsymbol{t}_i), \mathrm{Enc}_{\mathrm{I}}(\boldsymbol{x})\rangle/\tau)}{\sum_{j=1}^{C} \exp(\langle \mathrm{Enc}_{\mathrm{T}}(\boldsymbol{t}_j), \mathrm{Enc}_{\mathrm{I}}(\boldsymbol{x})\rangle/\tau)}$$

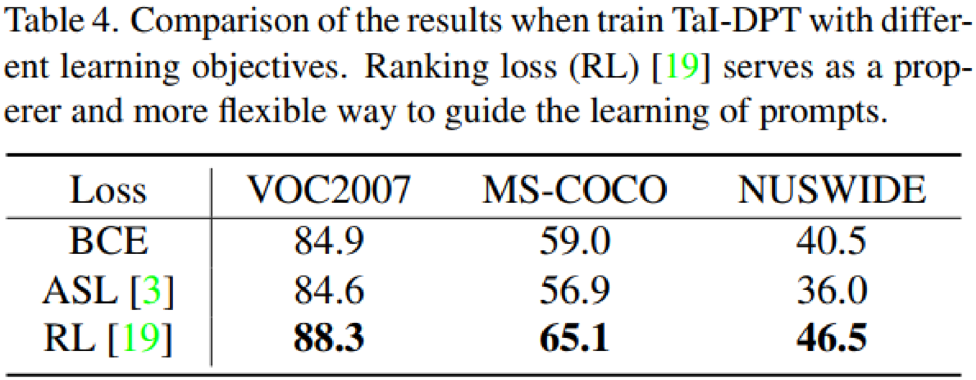
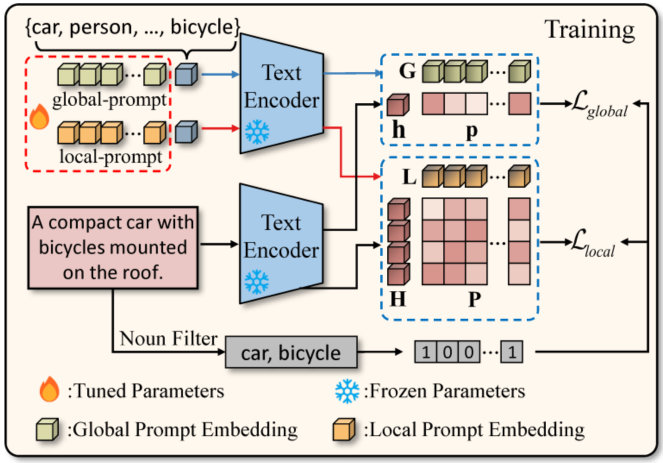


Table 4. Comparison of the results when train TaI-DPT with different learning objectives. Ranking loss (RL) [19] serves as a properer and more flexible way to guide the learning of prompts.

Loss	VOC2007	MS-COCO	NUSWIDE
BCE	84.9	59.0	40.5
ASL [3]	84.6	56.9	36.0
RL [19]	**88.3**	**65.1**	**46.5**

Visualization

- Text prompt tuning（local）



(a)

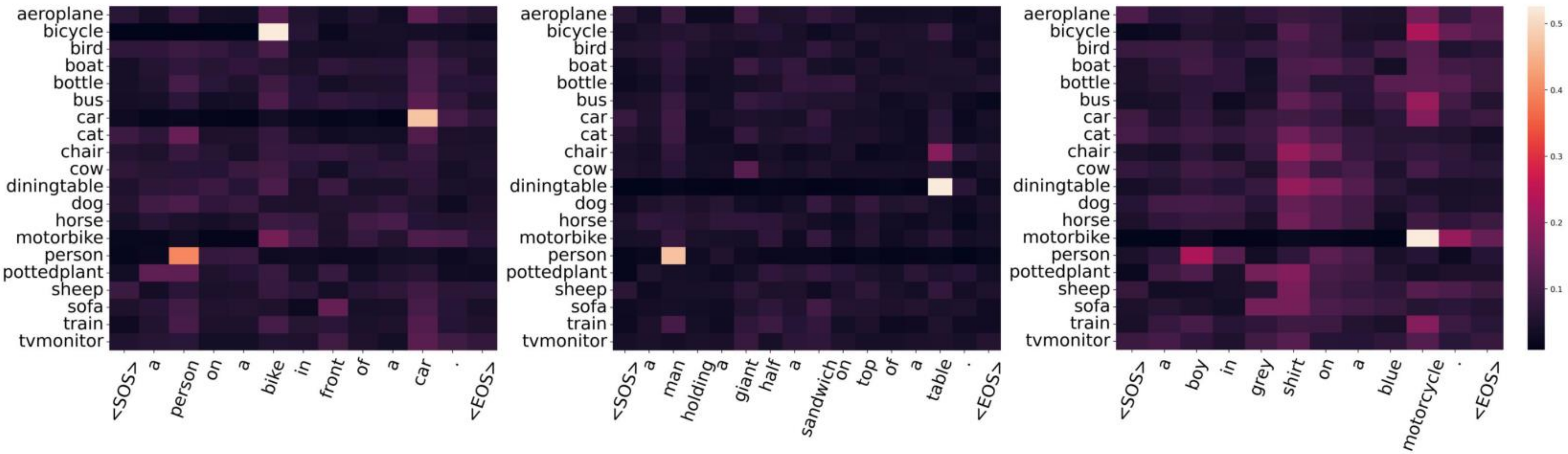
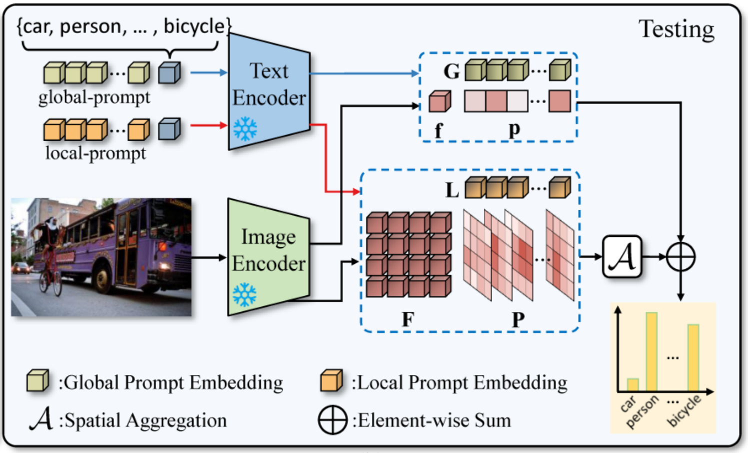


Figure 3. Visualization of correlations $\boldsymbol{P}$ between the local class embedding $\boldsymbol{L}$ and sequential token feature from texts. Each class embedding clearly correlates to words that describe the corresponding class (shown in highlight regions) rather than the global <EOS> token.

Visualization



(b)

- Image inference
  - Local class embeddings can localize objects

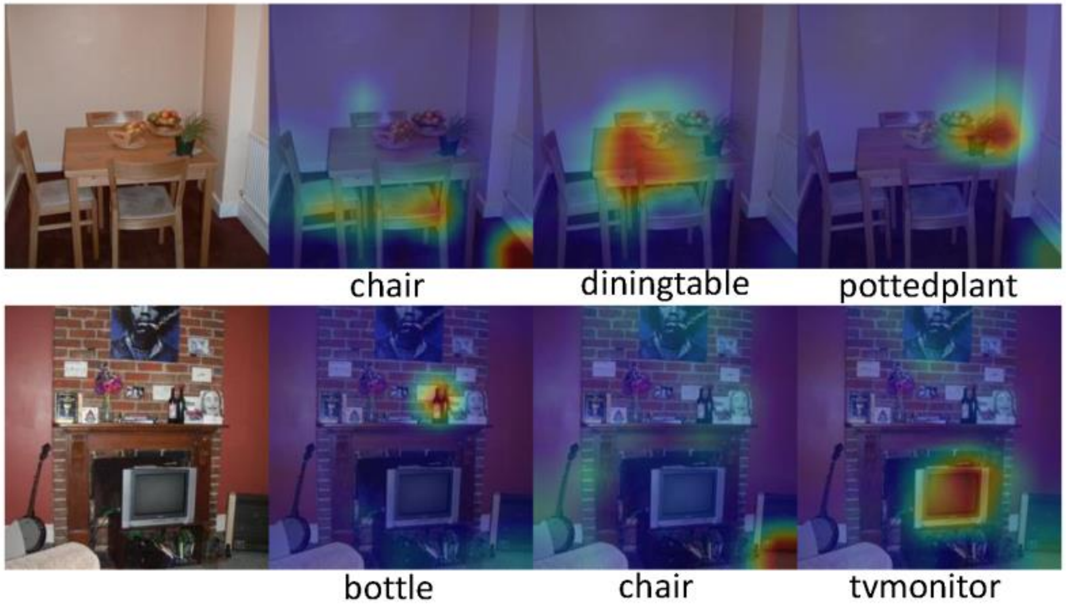
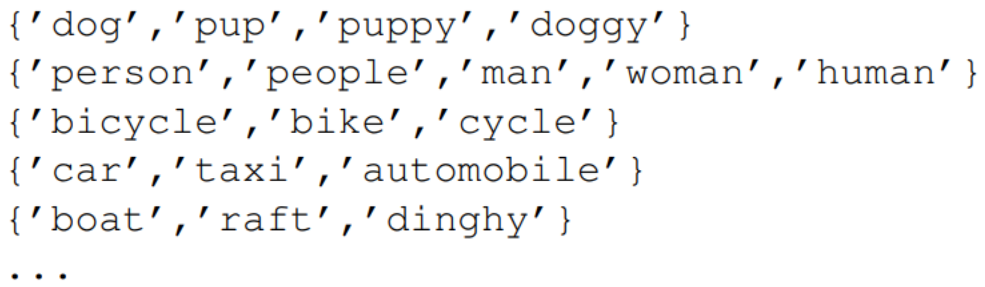


Figure 4. Visualization of correlations between the local class embedding $\boldsymbol{L}$ and dense image feature. The learned class embeddings can focus on the location of the object effectively.

Text Description

- □ 要求
  - ⊙ 完整描述一张图的内容
  - ⊙ 文本描述覆盖目标数据集的所有类别
- □ 直接使用MS-COCO、OpenImage等数据集的caption标注作为语料
- □ 类别名称的近义词词典



```
{'dog','pup','puppy','doggy'}
{'person','people','man','woman','human'}
{'bicycle','bike','cycle'}
{'car','taxi','automobile'}
{'boat','raft','dinghy'}
...
```

- □ 从语料库中筛选出包含至少一个类别名称的句子，用来训练

Incorporate with CoOp

- Text prompt方法可以插入到其他有图像的prompt tuning方法中

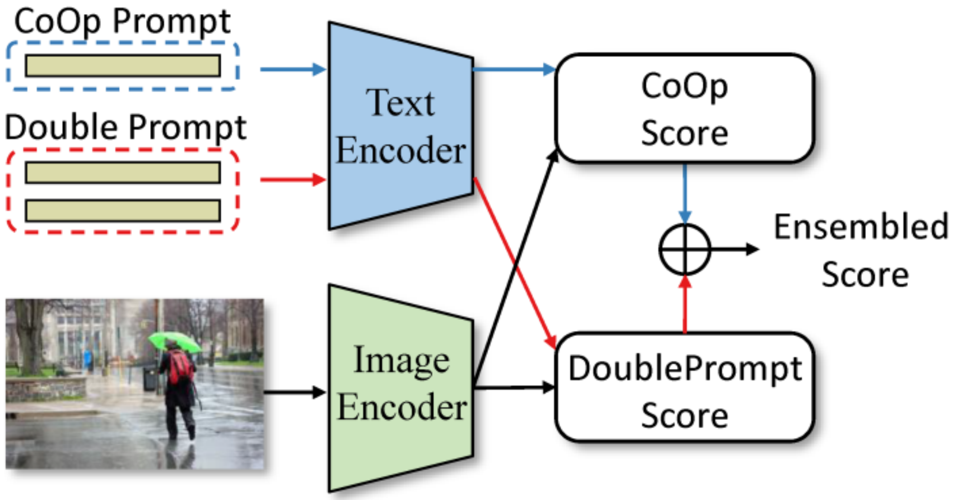


Figure 5. Our learned double-grained prompt tuning is easy to combine with existing prompt tuning methods with ensemble.

- □ 作者介绍
- □ 研究背景
- □ 本文方法
- □ 实验效果
- □ 总结反思

Experiment

- Ablation
  - Dual prompt tuning
  - Number of caption

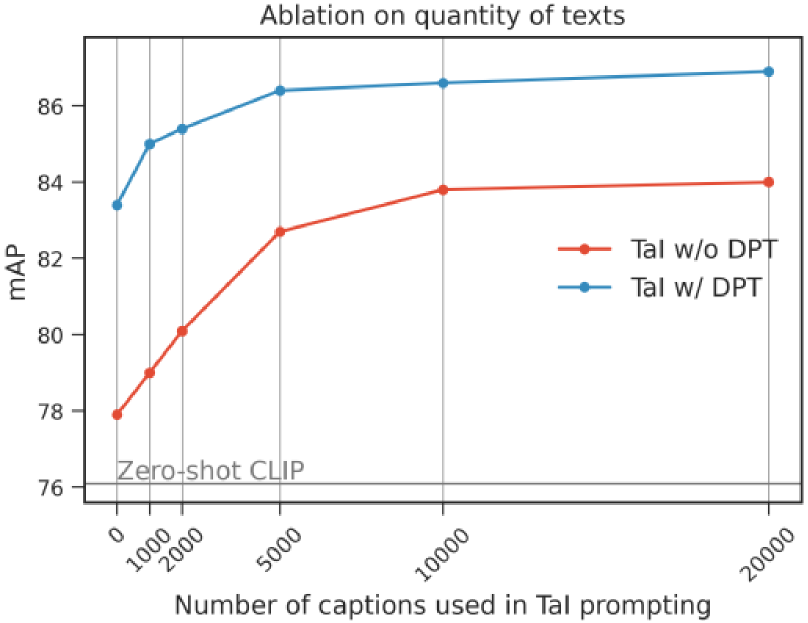


Figure 7. Ablation experiment on the number of texts and performance of TaI prompting on VOC2007

□ Few-shot multi-label learning

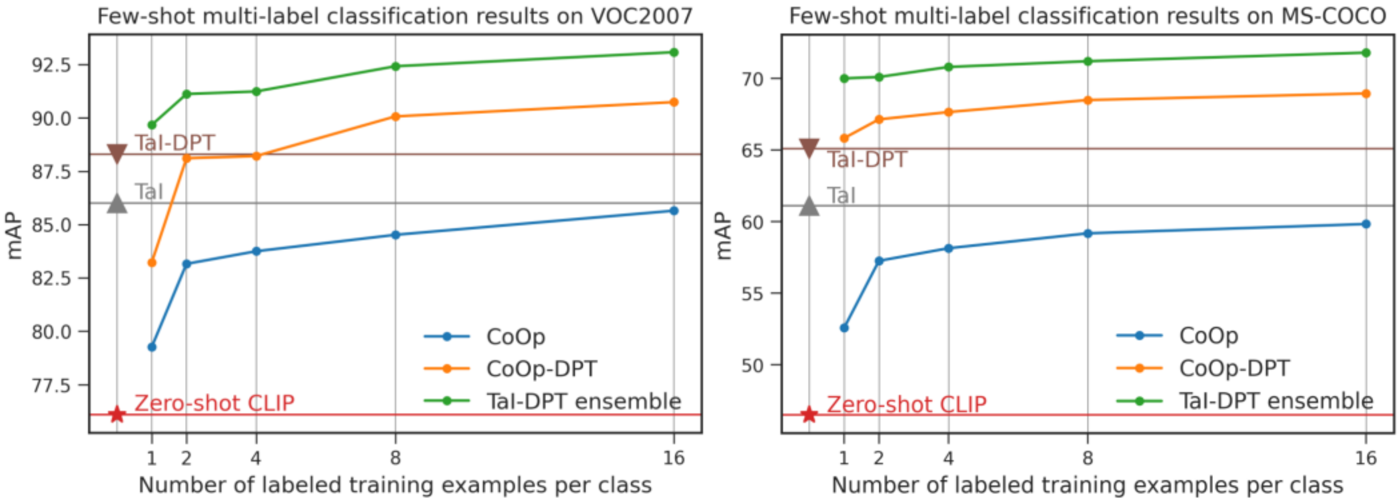


Figure 6. Comparison of different methods in few-shot multi-label recognition on VOC2007 and MS-COCO. Our zero-shot TaI-DPT can achieve comparable results with methods trained by 16-shot labeled image samples. And learned prompt ensemble proofs the complementarity between images and texts.

- □ Zero-shot，few shot multi-label learning

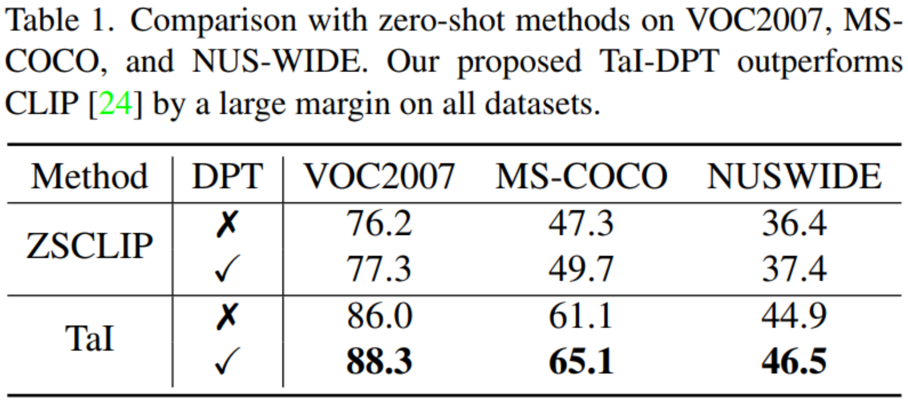


Table 1. Comparison with zero-shot methods on VOC2007, MS-COCO, and NUS-WIDE. Our proposed TaI-DPT outperforms CLIP [24] by a large margin on all datasets.

Method	DPT	VOC2007	MS-COCO	NUSWIDE
ZSCLIP	✗	76.2	47.3	36.4
	✓	77.3	49.7	37.4
TaI	✗	86.0	61.1	44.9
	✓	**88.3**	**65.1**	**46.5**

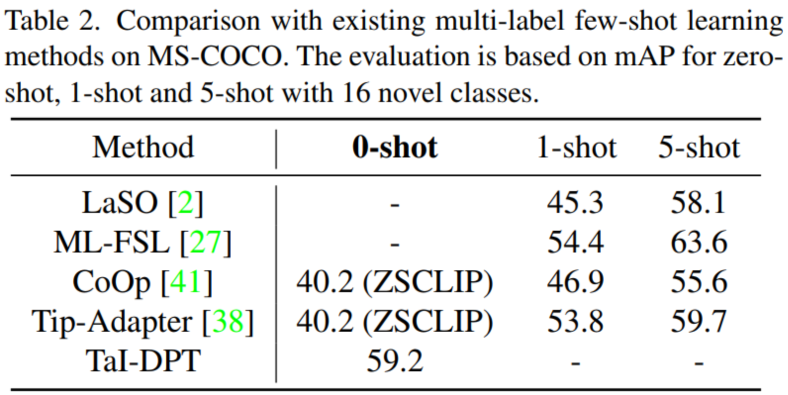


Table 2. Comparison with existing multi-label few-shot learning methods on MS-COCO. The evaluation is based on mAP for zero-shot, 1-shot and 5-shot with 16 novel classes.

Method	**0-shot**	1-shot	5-shot
LaSO [2]	-	45.3	58.1
ML-FSL [27]	-	54.4	63.6
CoOp [41]	40.2 (ZSCLIP)	46.9	55.6
Tip-Adapter [38]	40.2 (ZSCLIP)	53.8	59.7
TaI-DPT	59.2	-	-

□ Partial label

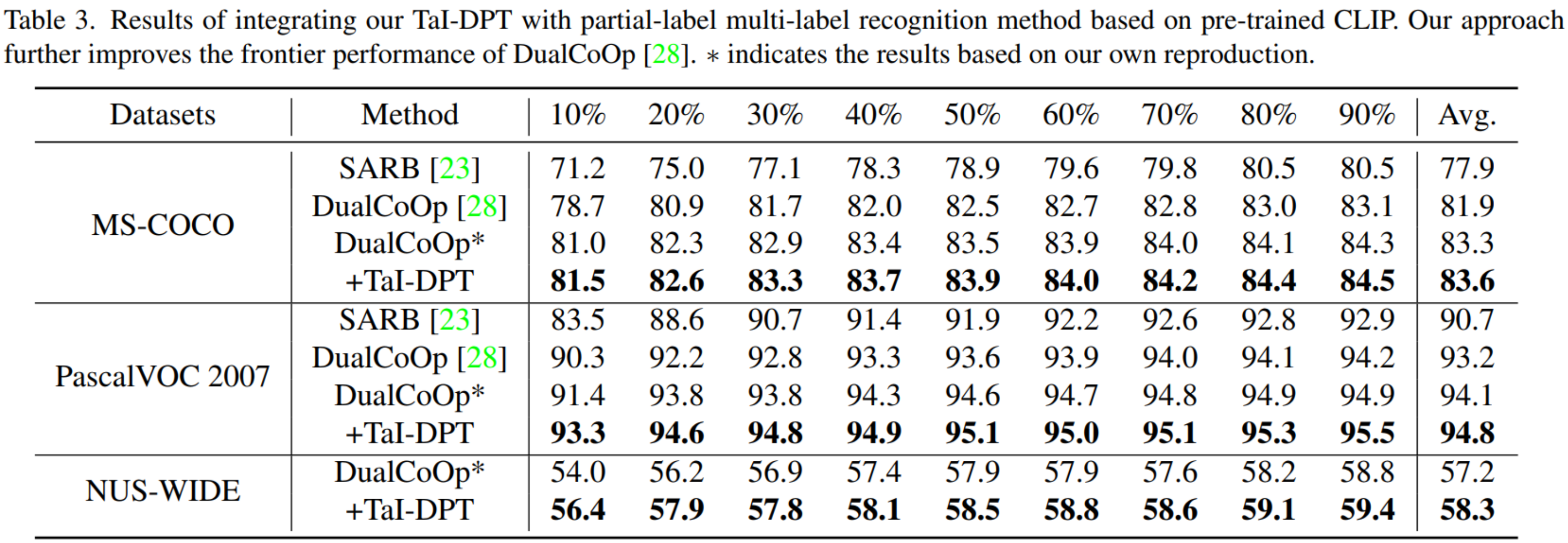


Table 3. Results of integrating our TaI-DPT with partial-label multi-label recognition method based on pre-trained CLIP. Our approach further improves the frontier performance of DualCoOp [28]. ∗ indicates the results based on our own reproduction.

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
MS-COCO	SARB [23]	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoOp [28]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
	DualCoOp*	81.0	82.3	82.9	83.4	83.5	83.9	84.0	84.1	84.3	83.3
	+TaI-DPT	**81.5**	**82.6**	**83.3**	**83.7**	**83.9**	**84.0**	**84.2**	**84.4**	**84.5**	**83.6**
PascalVOC 2007	SARB [23]	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
	DualCoOp [28]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
	DualCoOp*	91.4	93.8	93.8	94.3	94.6	94.7	94.8	94.9	94.9	94.1
	+TaI-DPT	**93.3**	**94.6**	**94.8**	**94.9**	**95.1**	**95.0**	**95.1**	**95.3**	**95.5**	**94.8**
NUS-WIDE	DualCoOp*	54.0	56.2	56.9	57.4	57.9	57.9	57.6	58.2	58.8	57.2
	+TaI-DPT	**56.4**	**57.9**	**57.8**	**58.1**	**58.5**	**58.8**	**58.6**	**59.1**	**59.4**	**58.3**

- □ 作者介绍
- □ 研究背景
- □ 本文方法
- □ 实验效果
- □ 总结反思

总结

- 方法创新性比较强，证明对于CLIP模型，单纯做text prompt tuning也能提高视觉分类能力。
- Text prompt tuning与视觉域的方法正交，可以联合使用

总结

I can't believe there's no images!
Learning Visual Tasks Using only Language Data

Sophia Gu* Christopher Clark* Aniruddha Kembhavi
Allen Institute for Artificial Intelligence
{sophiag, chrisc, anik}@alleni.org

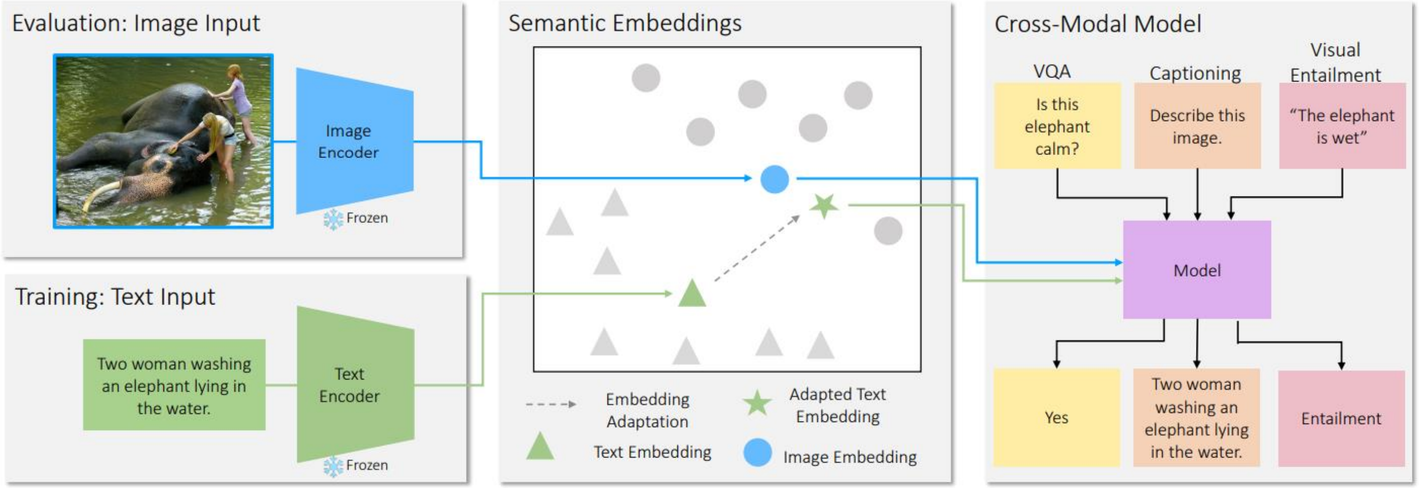
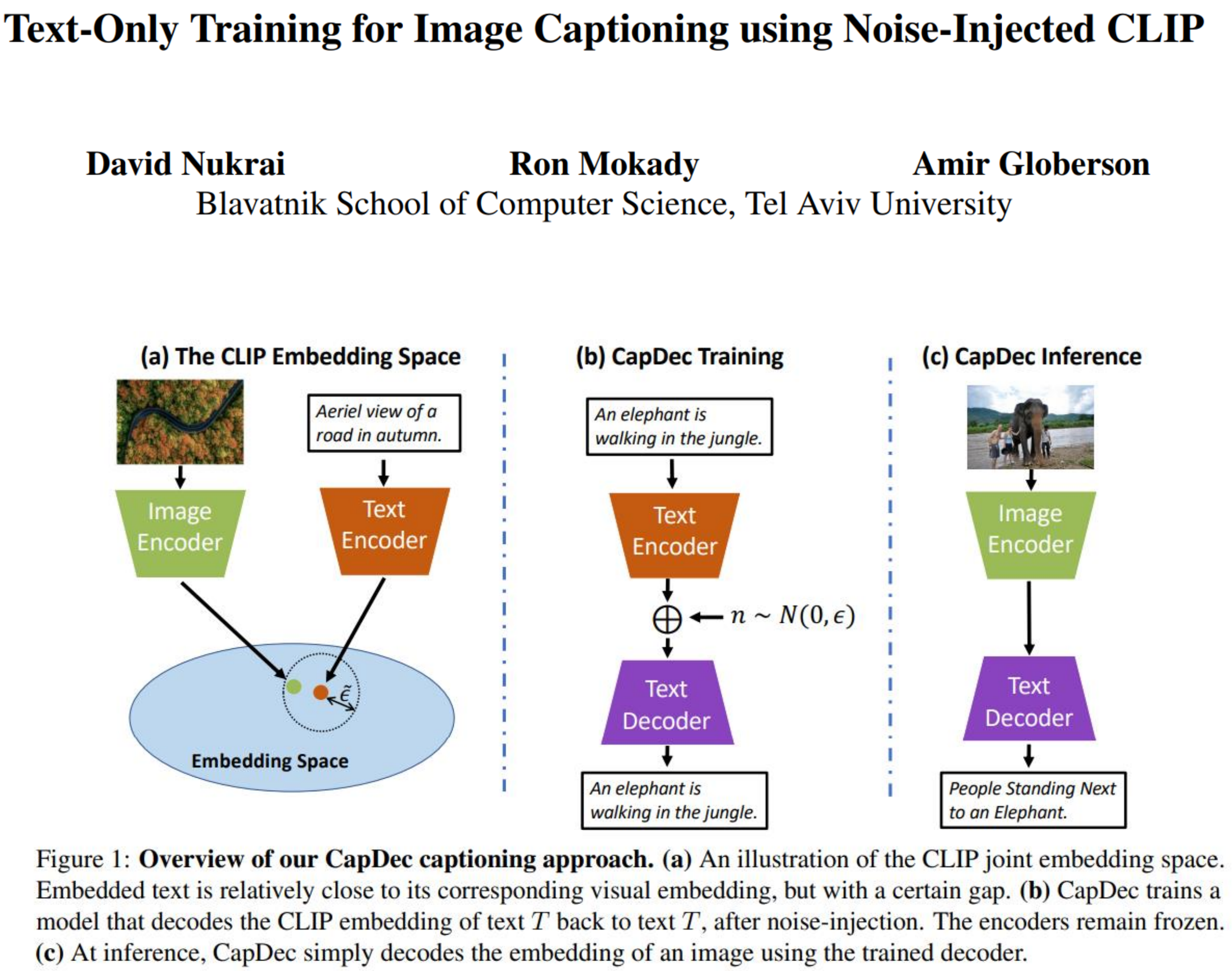


Figure 1: Overview of CLOSE. During training, input text is encoded into a vector with a text encoder and adapted with an adaptation method. A model learns to use the vector to perform a task such as VQA, captioning, or visual entailment. During testing, an input image is encoded with an image encoder instead to allow cross-modal transfer.

总结



Text-Only Training for Image Captioning using Noise-Injected CLIP

David Nukrai Ron Mokady Amir Globerson
Blavatnik School of Computer Science, Tel Aviv University

(a) The CLIP Embedding Space
Aeriel view of a road in autumn.
Image Encoder
Text Encoder
$\tilde{\epsilon}$
Embedding Space
(b) CapDec Training
An elephant is walking in the jungle.
Text Encoder
$\oplus \leftarrow n \sim N(0, \epsilon)$
Text Decoder
An elephant is walking in the jungle.
(c) CapDec Inference
Image Encoder
Text Decoder
People Standing Next to an Elephant.

Figure 1: **Overview of our CapDec captioning approach.** **(a)** An illustration of the CLIP joint embedding space. Embedded text is relatively close to its corresponding visual embedding, but with a certain gap. **(b)** CapDec trains a model that decodes the CLIP embedding of text T back to text T, after noise-injection. The encoders remain frozen. **(c)** At inference, CapDec simply decodes the embedding of an image using the trained decoder.

Thanks!