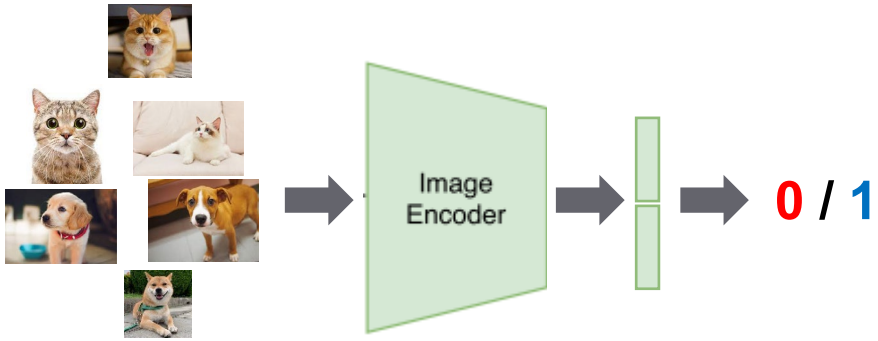
Rethinking the openness of CLIP
CLIP的开放性研究

Reporter：任抒怀

背景：什么是视觉识别中的开放性?

传统的视觉识别任务是**封闭**的（closed-set）：

- 要识别的图片类别是**预定义**好的
- 类别集合是**有限**的、**静态**的
- cifar10上训练的分类器，永远只能处理10个类别对应的图片输入



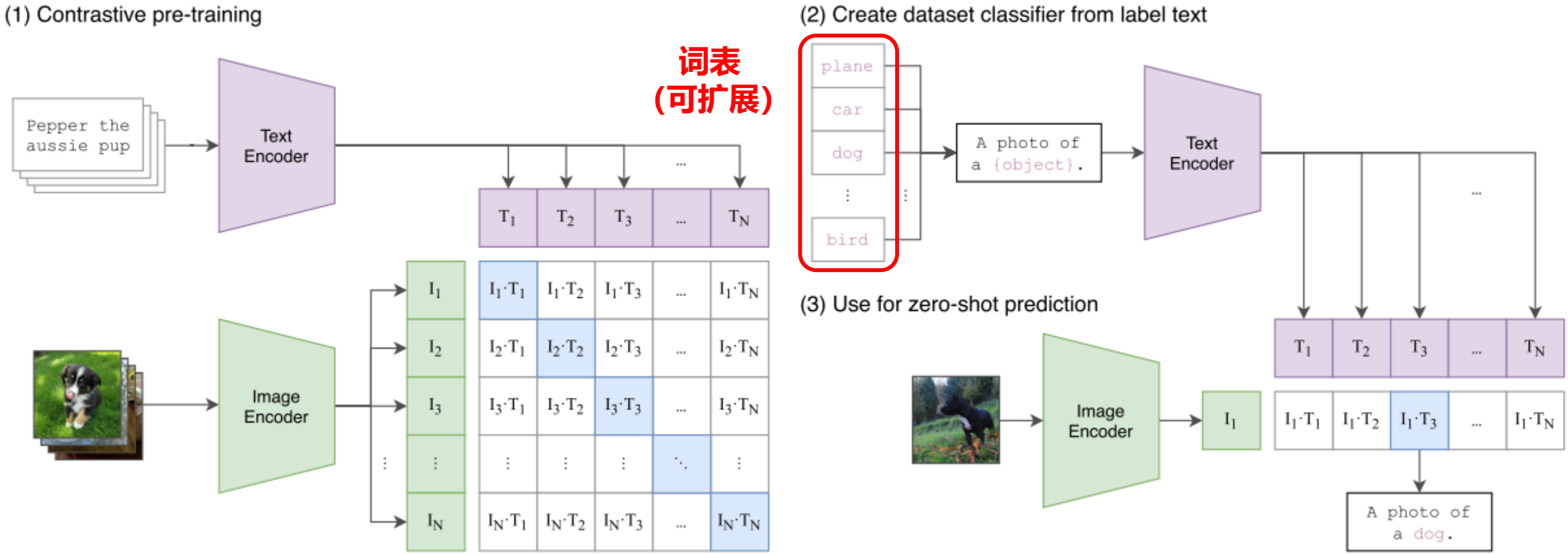
真实世界的视觉识别任务是**开放**的（open-world）：

- 输入的图片与其类别依据实际需求**动态变化**，无法预定义
  - 具体应用包括：医疗诊断、自动驾驶、视觉审查、商品分类...
- 期待模型具有**扩展性**，能灵活地适应**目标类别**的改变

背景：有模型接近我们对开放性的期待吗？

北京大学 PEKING UNIVERSITY

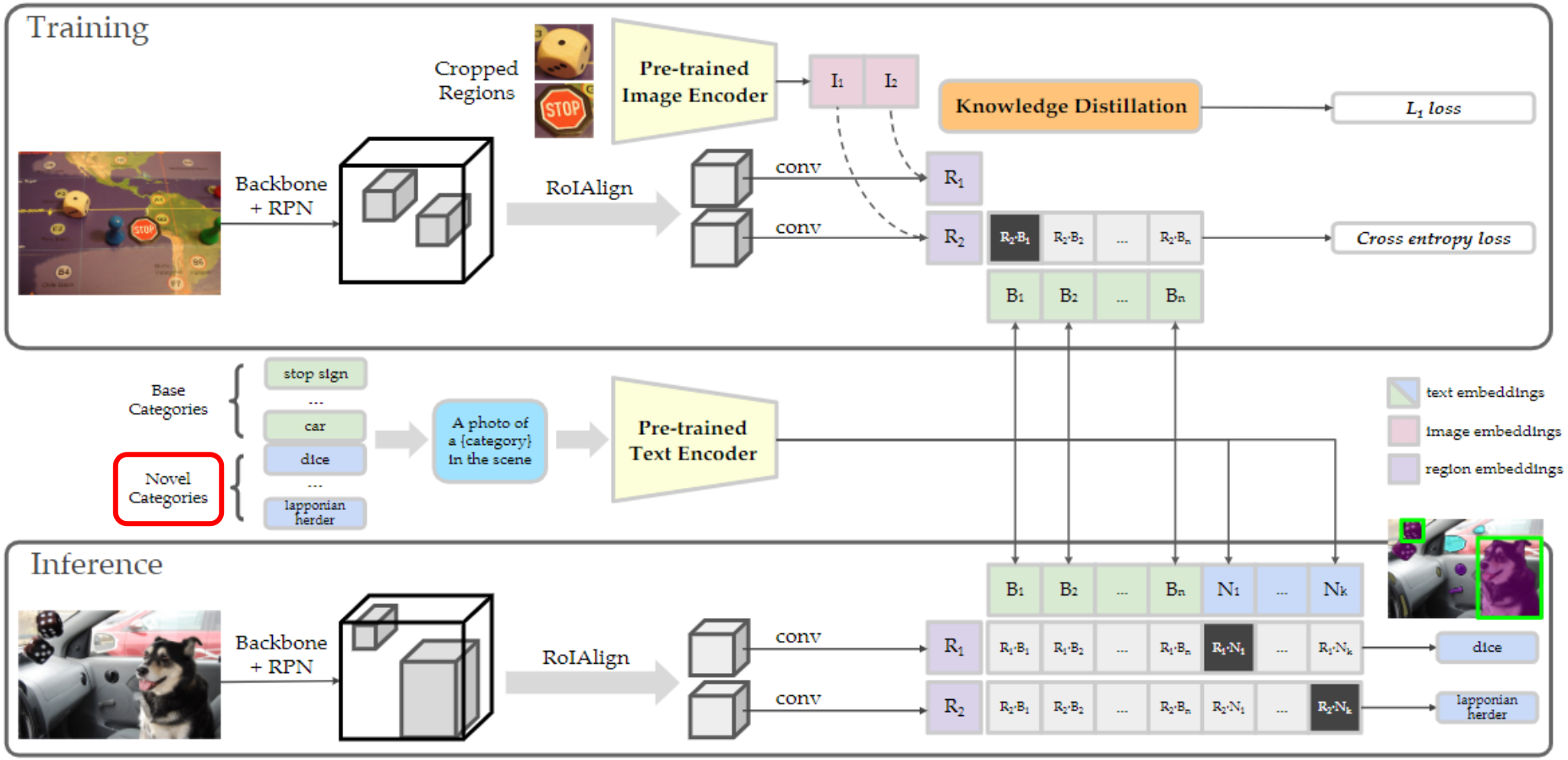
有，CLIP (Contrastive Language-Image Pretraining)

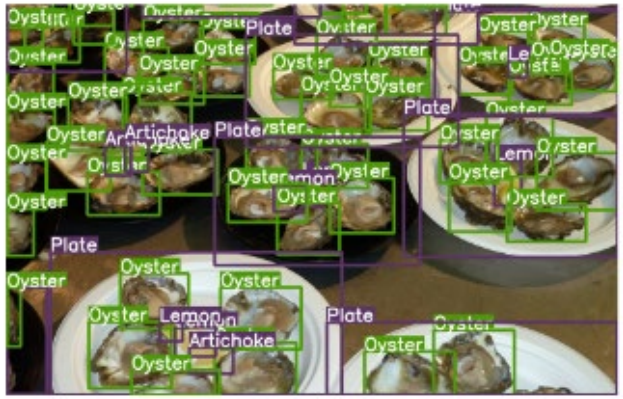


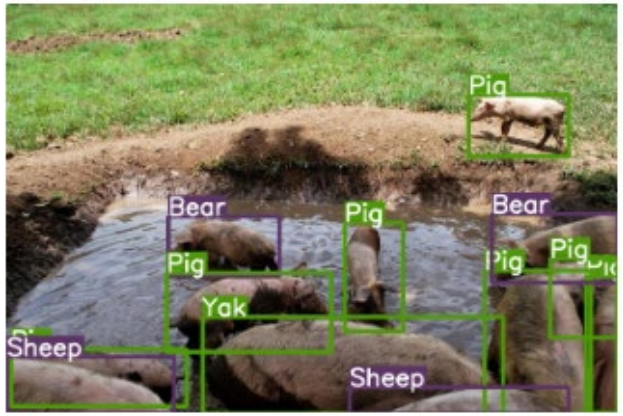
Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

背景：open-vocabulary范式为其它视觉模型注入开放性

Object Detection中的词表扩展：Novel Object Detection







Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, ICLR 2022
Detecting Twenty-thousand Classes using Image-level Supervision, 2022

探究：如何量化 CLIP 的开放性?

“CLIP has a wide range of capabilities due to its ability to carry out arbitrary image classification tasks”

现有评价体系的局限和矛盾：

- CLIP 在操作上可以向任意词表开放，可以灵活地适应词表扩展，

  但目前所有 CLIP-like models 都仅在封闭词表上进行评估

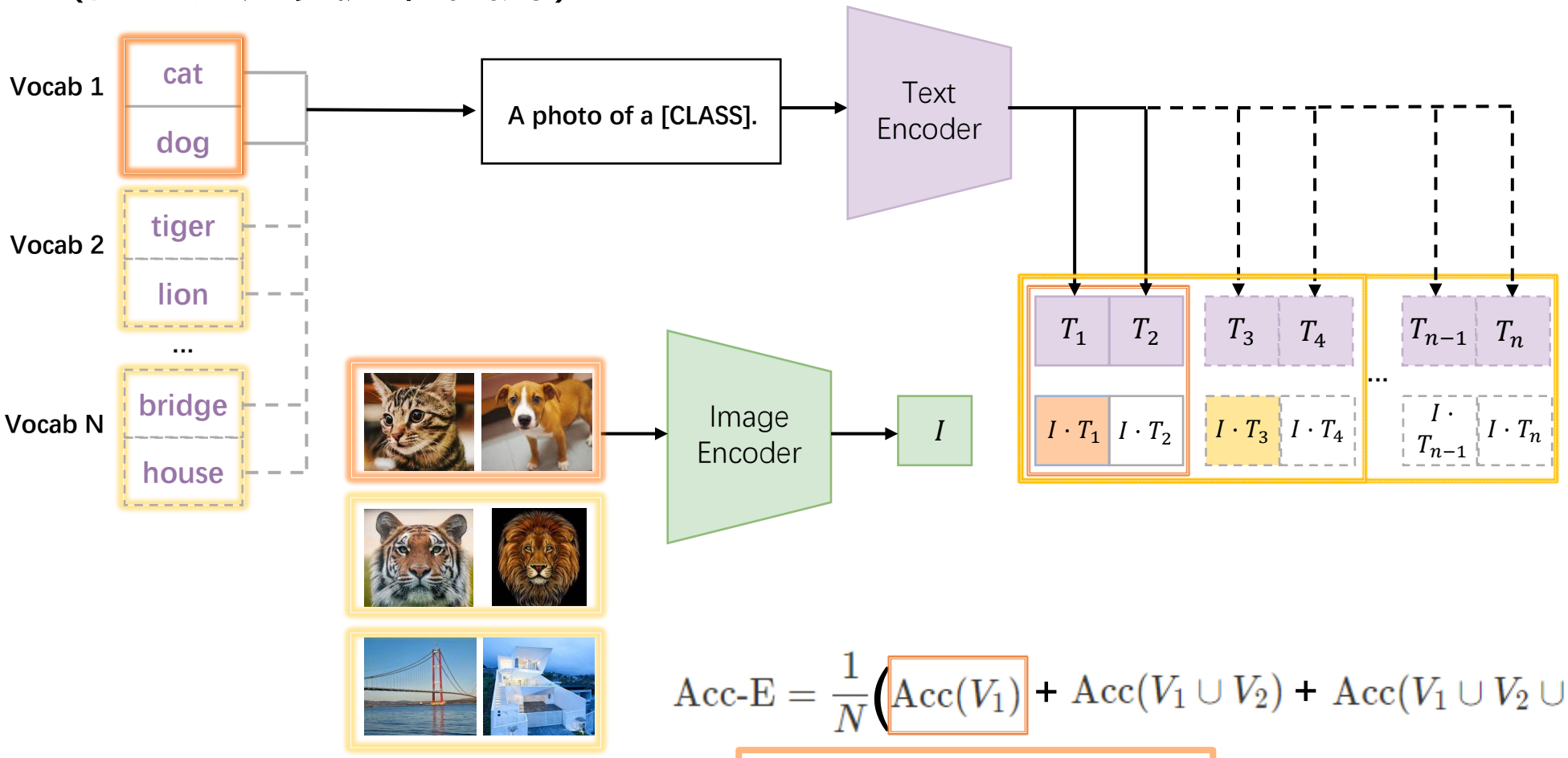
  没有显式建模词表扩展的这一动态性，无法反映 CLIP 面对类别扩展时真实的开放性

科学问题：

- 如何设计 evaluation protocol，反映 CLIP 在开放任务上的实际效果?

探究：通过渐进式的词表扩展量化 CLIP 的扩展性

扩展性（同时添加类别和图片）



$$\text{Acc-E} = \frac{1}{N}\left(\text{Acc}(V_1) + \text{Acc}(V_1 \cup V_2) + \text{Acc}(V_1 \cup V_2 \cup \cdots \cup V_N)\right)$$

Acc-C：closed-set下的模型表现

探究：通过渐进式的词表扩展量化 CLIP 的扩展性

词表怎么划分?

依据数据集自有的超类-子类结构

CIFAR100: 20个词表，每个词表中有5个类

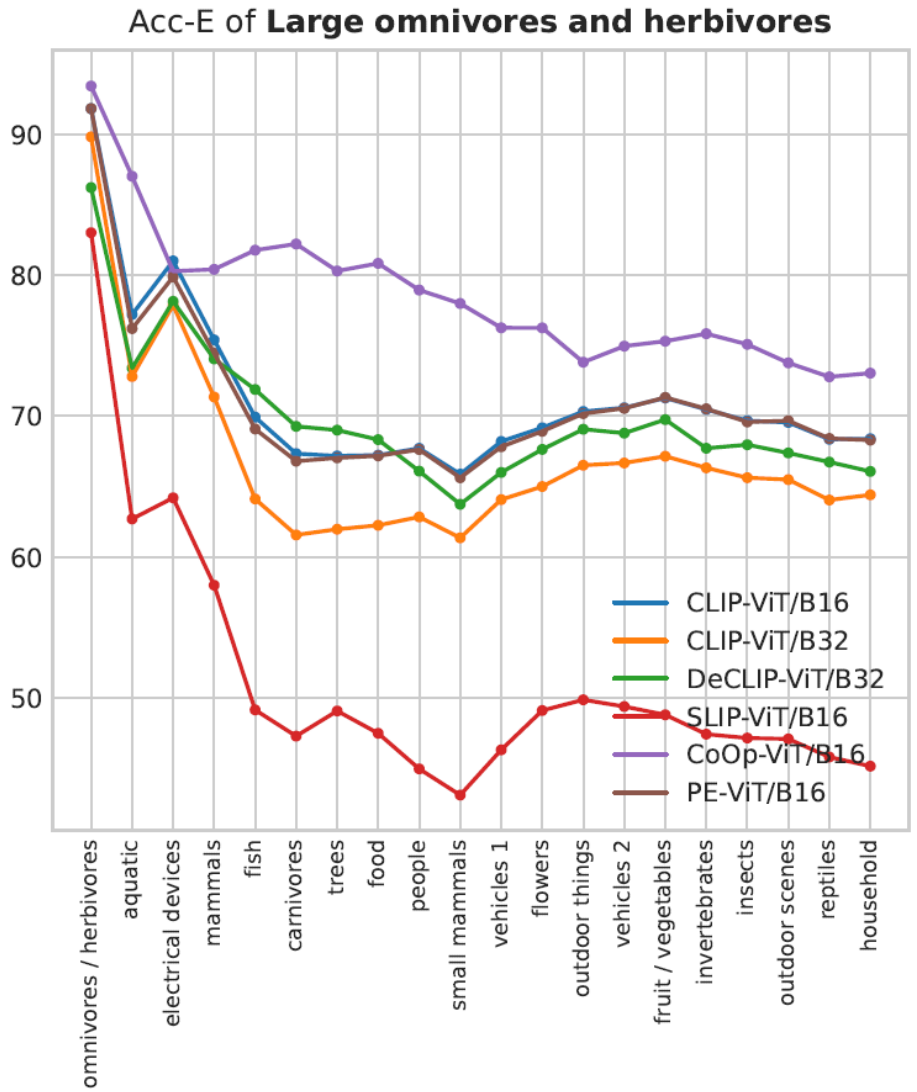
Vocabulary (Superclass)	Classes
aquatic	mammals beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food	containers bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household	furniture bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

探究：通过渐进式的词表扩展量化 CLIP 的扩展性

Table 1: Extensibility and stability of CLIP and its variants on CIFAR100 and ImageNet datasets (BREEDS benchmark). Δ refers to the decline of Acc-E/Acc-S (%) compared to Acc-C (%). PE denotes Prompt Ensemble. CoOp is prompt-tuned on all classes with 16 shots.

Model	CIFAR100					ImageNet (Entity13)					ImageNet (Living17)				
	Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability	
		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ
CLIP (RN101)	68.3	55.4	-12.9	54.9	-13.4	80.4	77.4	-3.0	77.3	-3.1	77.6	74.5	-2.9	74.4	-3.0
CLIP (ViT-B/32)	78.0	69.6	-8.4	68.9	-9.1	80.8	78.0	-2.8	77.8	-3.0	78.0	74.4	-3.6	75.0	-3.0
CLIP (ViT-B/16)	**79.7**	**72.6**	**-7.1**	72.0	-7.7	**83.5**	**81.1**	**-2.4**	81.0	-2.5	**79.5**	**77.9**	**-1.6**	77.6	-1.9
SLIP (ViT-B/16)	63.9	51.1	-12.8	50.4	-13.5	65.7	62.3	-3.4	62.0	-3.7	65.7	62.6	-3.1	62.5	-3.2
DeCLIP (ViT-B/32)	**78.7**	**70.8**	**-7.9**	70.4	-8.3	**81.9**	**79.2**	**-2.7**	79.1	-2.8	**82.1**	**80.2**	**-1.9**	80.0	-2.1
PE (ViT-B/32)	78.3	70.3	-8.0	69.9	-8.4	81.9	79.4	-2.5	79.2	-2.7	78.7	76.0	-2.7	75.8	-2.9
PE (ViT-B/16)	**79.6**	**72.6**	**-7.0**	72.0	-7.6	**85.3**	**83.2**	**-2.1**	83.1	-2.2	**79.6**	**78.2**	**-1.4**	78.0	-1.6
CoOp (ViT-B/16)	83.6	76.9	-6.7	76.7	-6.9	87.5	85.3	-2.2	85.5	-2.0	82.7	82.6	-0.1	81.3	-1.4

探究：通过渐进式的词表扩展量化 CLIP 的扩展性



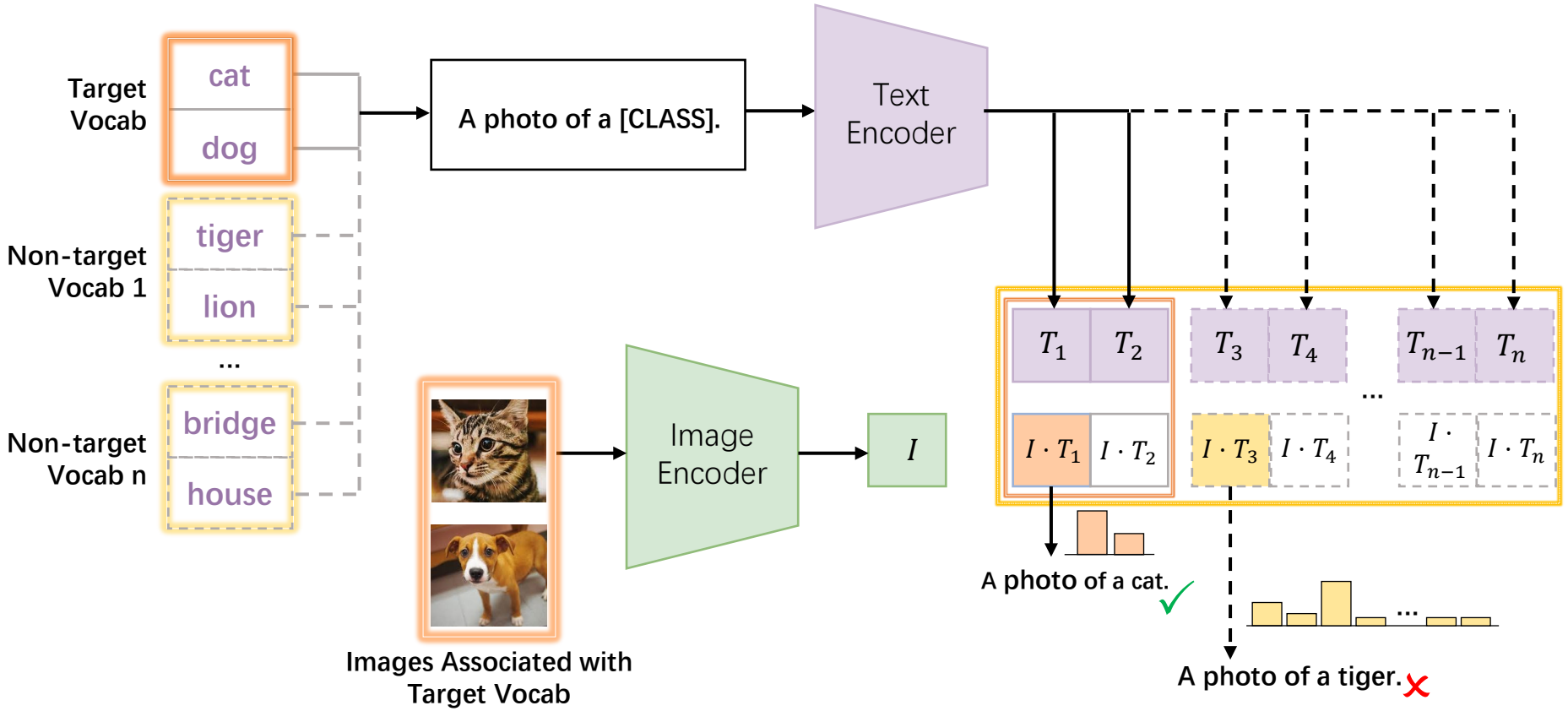
准确率（Acc-E）的下降说明：

1. 模型在新词表上的表现很差
2. 对于之前的图片输入，模型本来能分对，但引入新类别后，分错了（预测成新的类别了）

CLIP的预测稳定性很差？

探究：引入非目标词表量化 CLIP 面对扩展时的稳定性

稳定性（只加类别，不加图片）

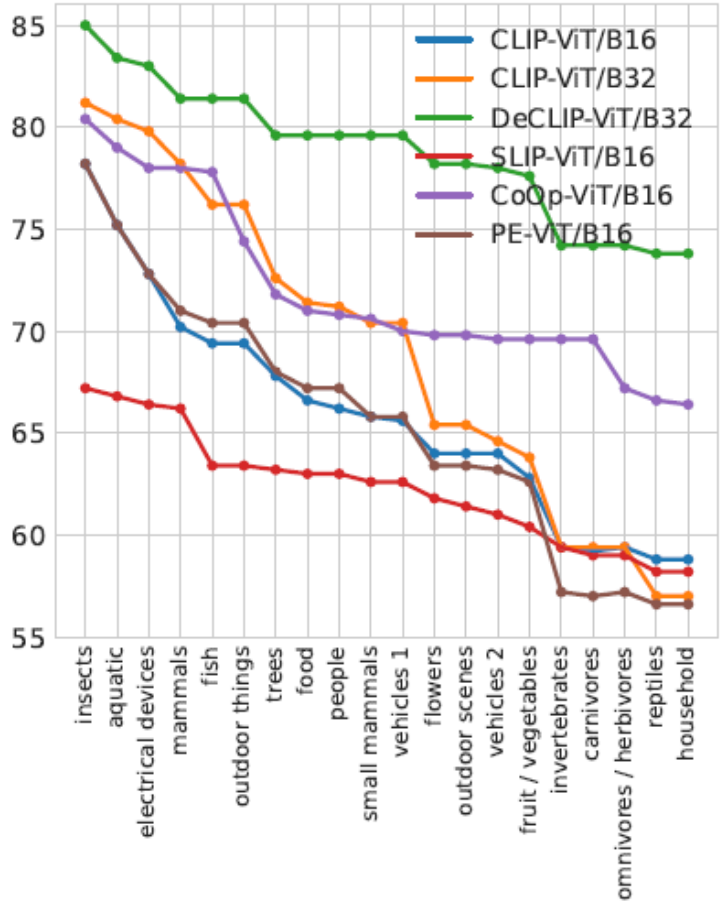


探究：引入非目标词表量化 CLIP 面对扩展时的稳定性

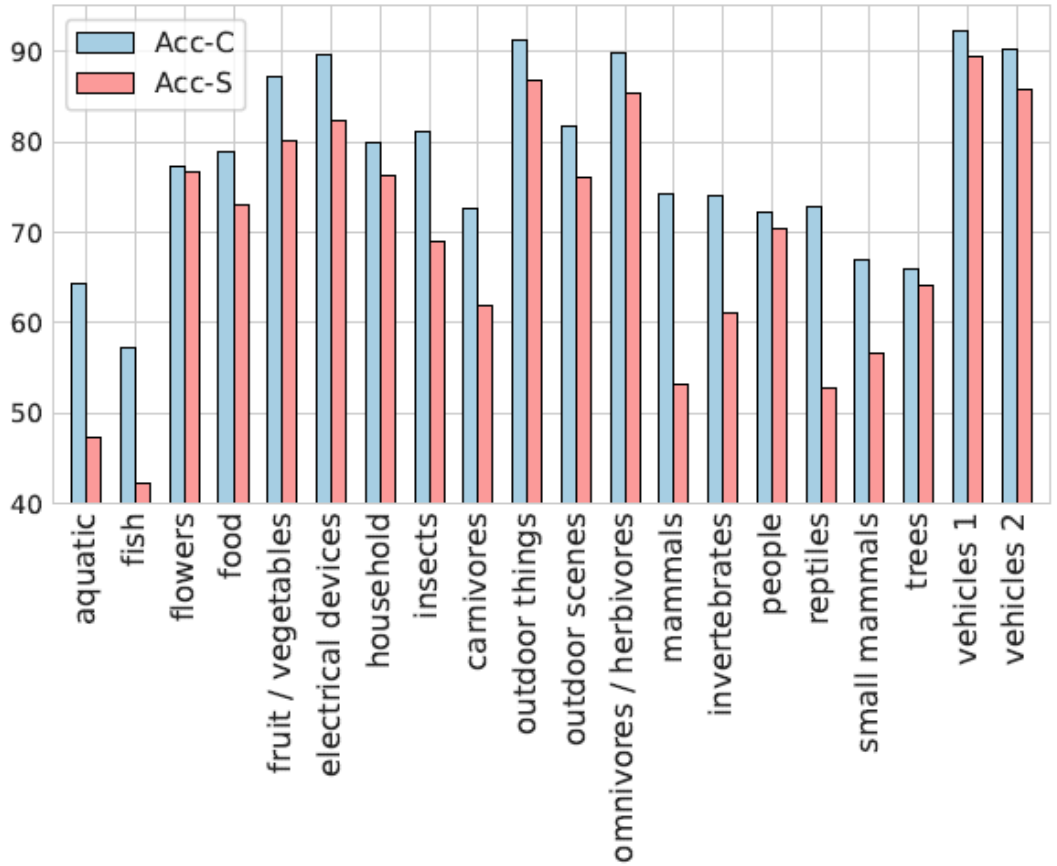
Table 1: Extensibility and stability of CLIP and its variants on CIFAR100 and ImageNet datasets (BREEDS benchmark). Δ refers to the decline of Acc-E/Acc-S (%) compared to Acc-C (%). PE denotes Prompt Ensemble. CoOp is prompt-tuned on all classes with 16 shots.

Model	CIFAR100					ImageNet (Entity13)					ImageNet (Living17)				
	Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability	
		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ
CLIP (RN101)	68.3	55.4	-12.9	54.9	-13.4	80.4	77.4	-3.0	77.3	-3.1	77.6	74.5	-2.9	74.4	-3.0
CLIP (ViT-B/32)	78.0	69.6	-8.4	68.9	-9.1	80.8	78.0	-2.8	77.8	-3.0	78.0	74.4	-3.6	75.0	-3.0
CLIP (ViT-B/16)	**79.7**	**72.6**	**-7.1**	**72.0**	**-7.7**	**83.5**	**81.1**	**-2.4**	**81.0**	**-2.5**	**79.5**	**77.9**	**-1.6**	**77.6**	**-1.9**
SLIP (ViT-B/16)	63.9	51.1	-12.8	50.4	-13.5	65.7	62.3	-3.4	62.0	-3.7	65.7	62.6	-3.1	62.5	-3.2
DeCLIP (ViT-B/32)	**78.7**	**70.8**	**-7.9**	**70.4**	**-8.3**	**81.9**	**79.2**	**-2.7**	**79.1**	**-2.8**	**82.1**	**80.2**	**-1.9**	**80.0**	**-2.1**
PE (ViT-B/32)	78.3	70.3	-8.0	69.9	-8.4	81.9	79.4	-2.5	79.2	-2.7	78.7	76.0	-2.7	75.8	-2.9
PE (ViT-B/16)	**79.6**	**72.6**	**-7.0**	**72.0**	**-7.6**	**85.3**	**83.2**	**-2.1**	**83.1**	**-2.2**	**79.6**	**78.2**	**-1.4**	**78.0**	**-1.6**
CoOp (ViT-B/16)	83.6	76.9	-6.7	76.7	-6.9	87.5	85.3	-2.2	85.5	-2.0	82.7	82.6	-0.1	81.3	-1.4

探究：引入非目标词表量化 CLIP 面对扩展时的稳定性

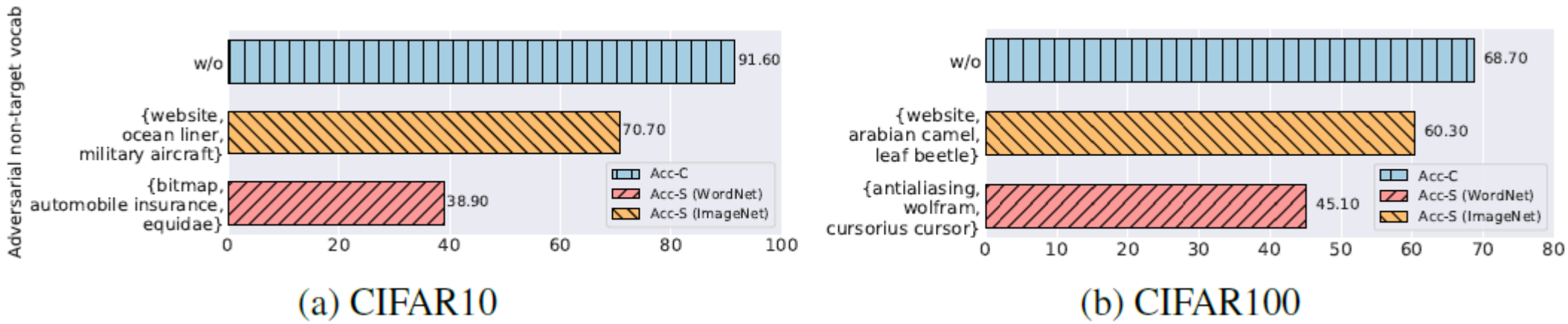


每次下降都说明：
有insects的图片被错分到新引入的类别



有的超类稳定
有的超类不稳定

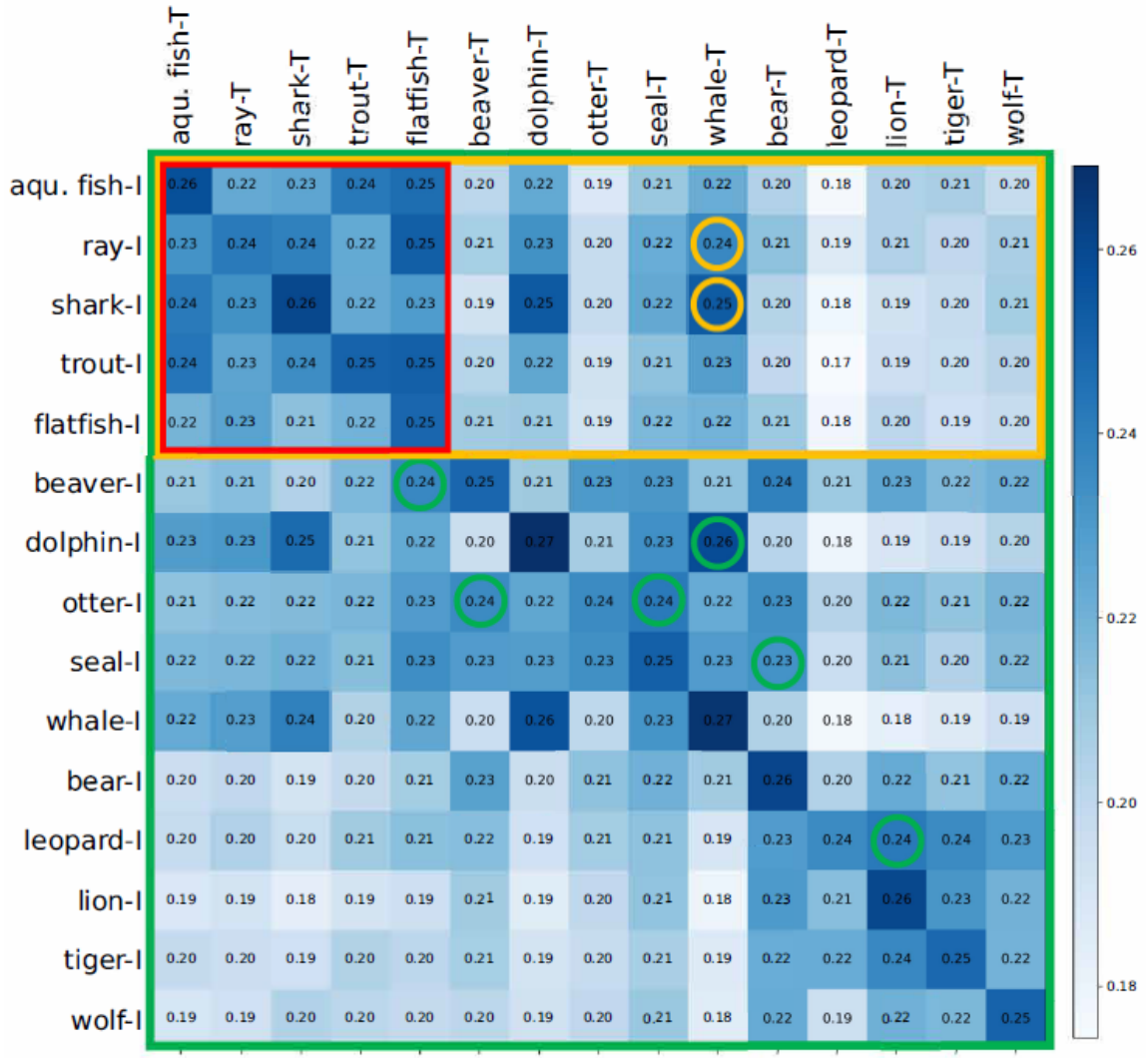
探究：CLIP 稳定性的下界



(a) CIFAR10　　(b) CIFAR100

对抗非目标类：
通过恶意引入少数（3个）非目标类，使得准确率大幅下降

分析：什么限制了 CLIP 的扩展性和稳定性?

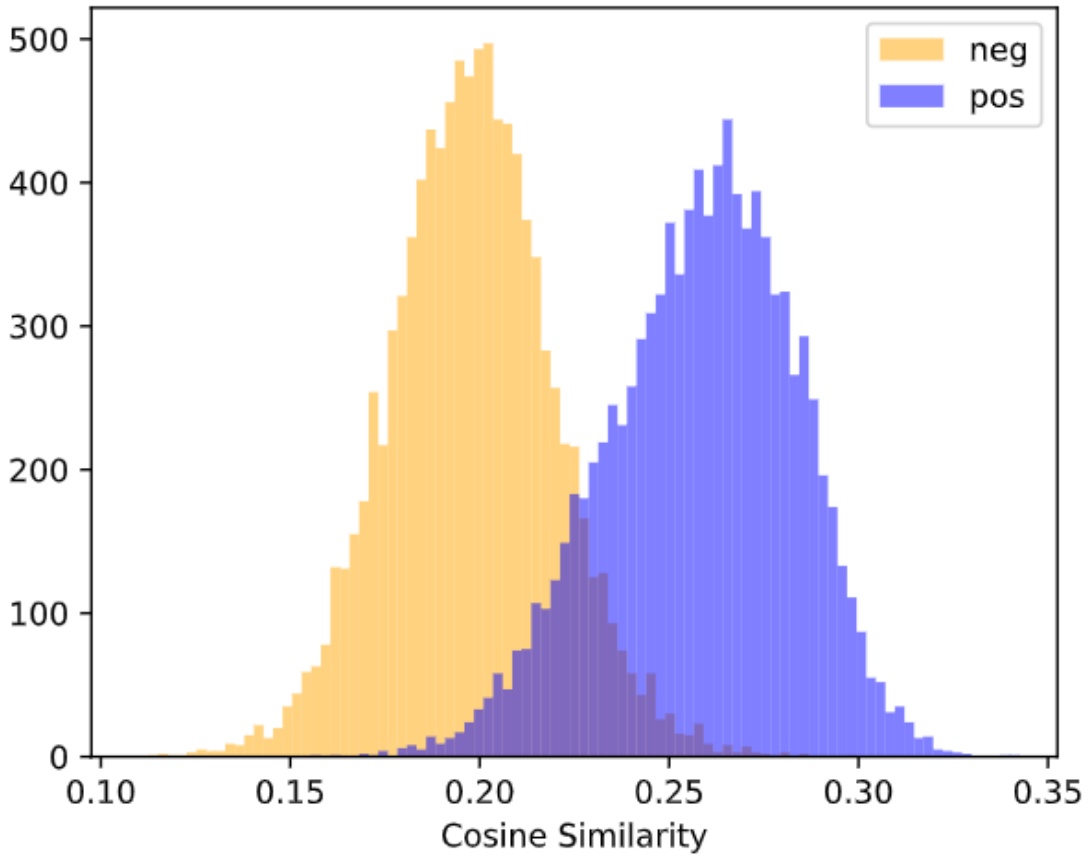


图文匹配的余弦相似度矩阵

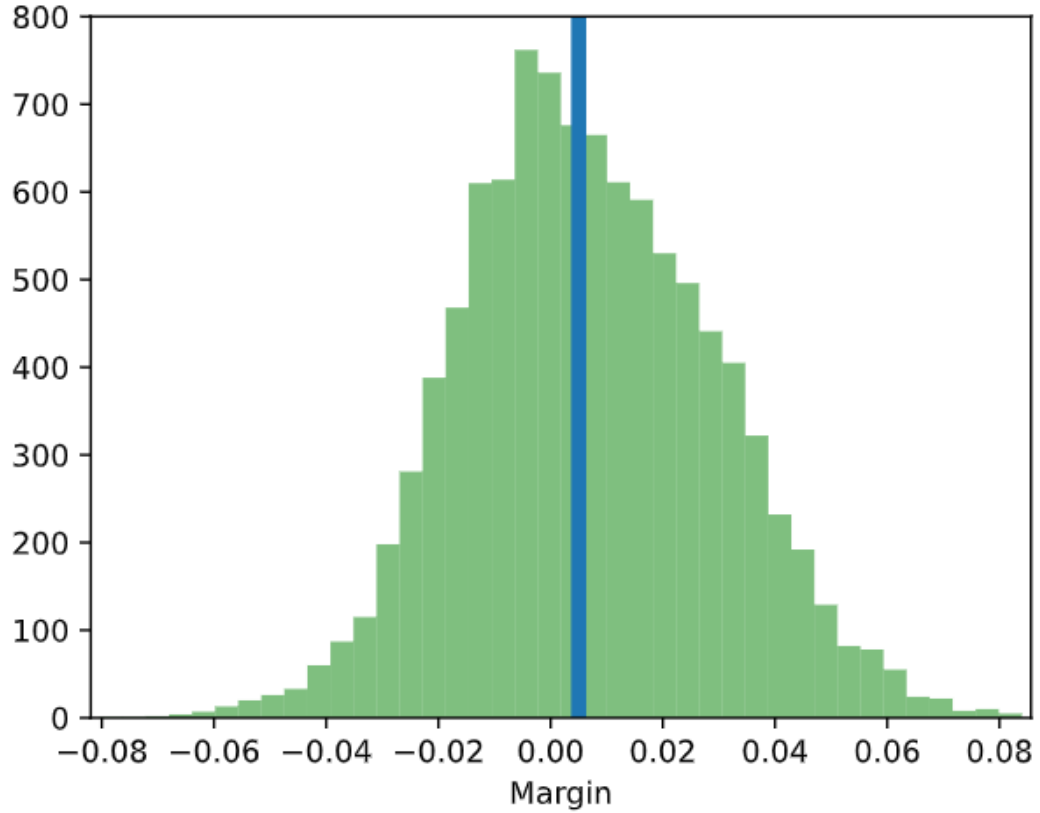
- 纵轴：图片特征；横轴：文本特征
- 文本特征对应分类器中的类别向量（class vector）
- 每一行的相似度对应分类器中的logits
- 对角线：正例pair；其余：负例pair
- 红框到橙框：稳定性对应的词表扩展
  红框到绿框：扩展性对应的词表扩展

无论词表怎么扩展，正例相似度要高于负例相似度，这样才能保证做对

分析：什么限制了 CLIP 的扩展性和稳定性?



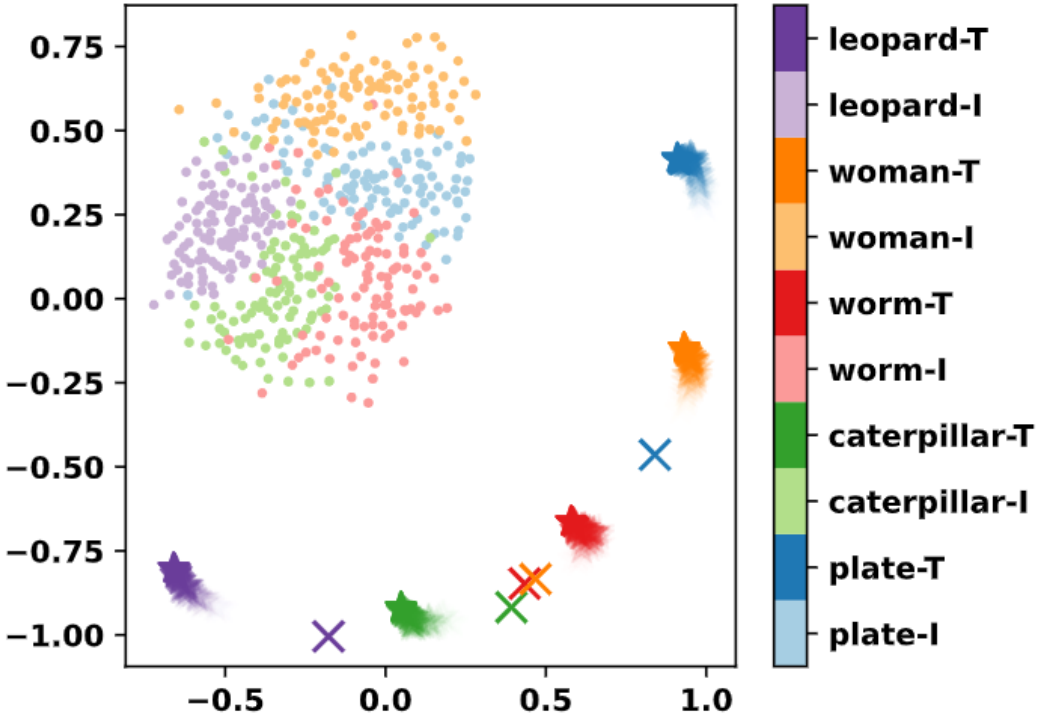
正例平均相似度很低：0.26
正负例overlap很大



Margin：正例相似度-最大负例相似度（越大越好）
但实际的margin（中位数）很小，仅为0.005
使得新负类出现时模型的预测容易出现漂移

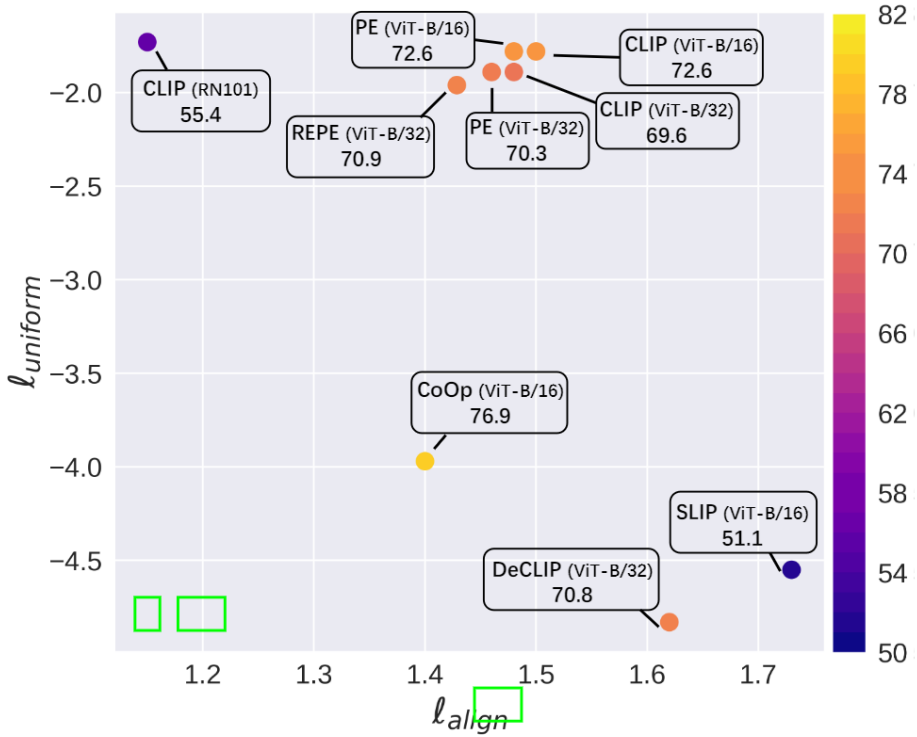
分析：理想的特征空间应该长啥样？

- Inter-modal alignment：图片和其对应的类别描述要聚在一起
- Intra-modal uniformity：单模态内部，不同类别的特征要尽可能分散（均匀分布），提高类别之间的可分性



CoOp（prompt tuning）的优化路径

- 类别描述趋向分散
- 跑向对应的文本特征的质心



alignment和uniform loss越低越好
点的颜色代表扩展性（越深越好）

方法：基于检索增强的提示工程算法（REPE）

之前的类别描述：
A photo of a [CLASS].

Table 6: Instances of the captions retrieved by our REPE on CIFAR100.

Class	Retrieved captions
apple	"Apple slices stacked on top of each other" "Apples growing on a tree" "Still life with apples in a basket"
woman	"Portrait of a young woman" "Woman standing at the window" "Confident woman in a red dress and gold crown"
bridge	"The golden bridge in Bangkok" "Bridge on the River Kwai ~Video Clip" "Wooden bridge over a mountain river"
ray	"Stingray in the Grand Cayman, Cayman Islands stock photography" "Common Stingray swimming close to the sea floor." "Sun Rays Tours: Go Pro captured the rays under water"

预训练数据库中与类别相关的caption

提供了更丰富的类别语义、特征、上下文

方法：基于检索增强的提示工程算法（REPE）

北京大学 PEKING UNIVERSITY

检索增强的推理方法

- 在推理时，利用每个类别对应的文本描述，去**预训练图文对**中检索出最相关的**K个图片**
- 将这些图片的**caption**特征平均后加到文本特征上，作为最终的文本特征
- 使得文本特征更有表达性、可分性
- 无参、无需fine-tune、灵活通用可扩展

方法：基于检索增强的提示工程算法（REPE）

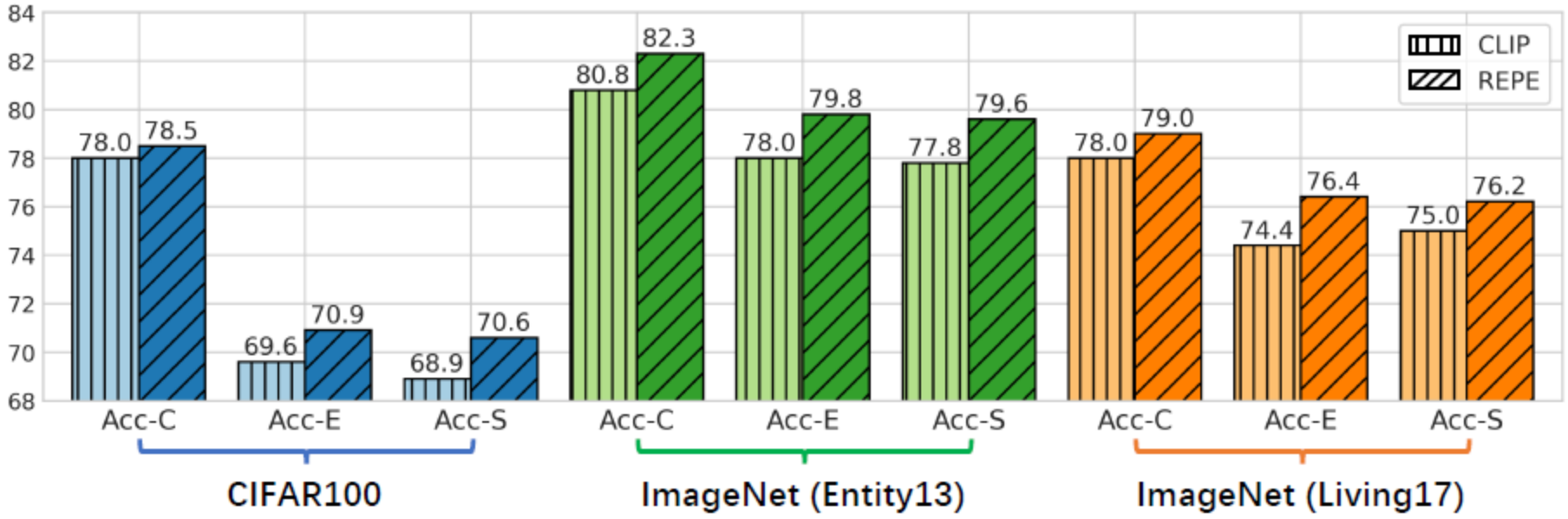
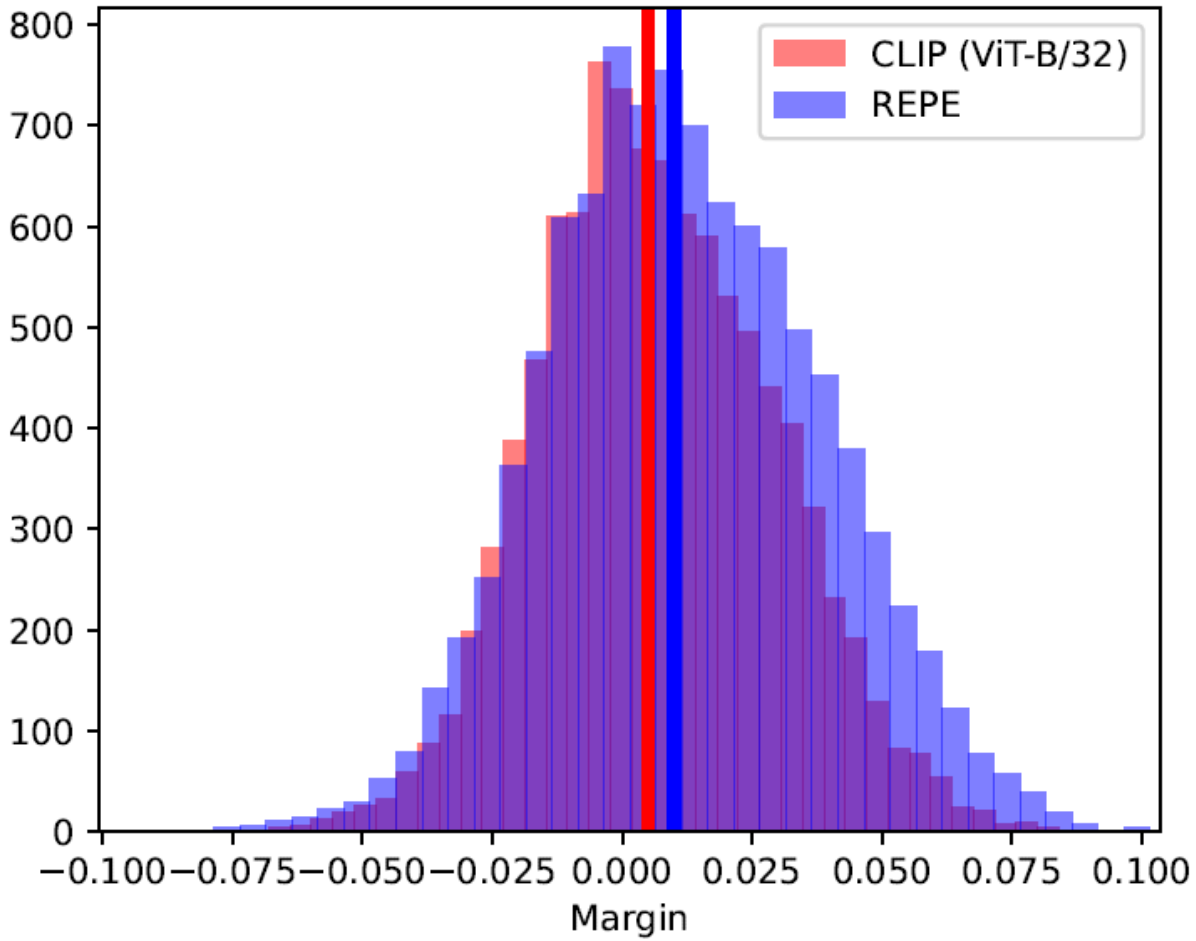


Figure 9: Extensibility and stability of our retrieval-enhanced prompt engineering (REPE) on CIFAR100 and ImageNet datasets. The vision backbone of both CLIP and REPE is ViT-B/32.

方法：基于检索增强的提示工程算法（REPE）



将margin从0.005推大到了0.01

Text uniformity loss从-0.85降到了-1.01

Inter-alignment loss从1.50降到了1.43

总结

北京大学 PEKING UNIVERSITY

➢ 贡献

- 第一个探究了CLIP在真实世界的开放任务上的表现
- 为此设计了一套分析方法和评估指标
- 在特征空间对CLIP进行分析，从margin、alignment和uniformity角度指出其在开放任务上表现不佳的原因
- 提出了检索增强的推理方法，在无需fine-tune的情况下提高CLIP在开放任务上的表现

➢ 结论

- CLIP-like models 在真实世界的开放任务上表现不佳，扩展性和稳定性都较差
- 更好的模型架构（ViT），更广泛的预训练监督信号（DeCLIP），更多的预训练数据有助于提高模型的扩展性
- CLIP面对开放任务表现不佳的原因是正负例类别描述的可分性差，margin过小
- Prompt tuning通过提升类别描述的区分度和图文对齐改善表现
- 使用检索增强的推理方法能够帮助CLIP提高面对开放任务的zero-shot表现

谢谢

任抒怀

renshuhuai007@gmail.com