Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model

CVPR 2022

报告人：徐静远

目录

目录

作者介绍

Yu Du[1] Fangyun Wei[2†] Zihe Zhang[1] Miaojing Shi[3†] Yue Gao[2] Guoqi Li[1]

[1]Tsinghua University [2]Microsoft Research Asia [3]King's College London

{duyu20, zh-zhang17}@mails.tsinghua.edu.cn liguoqi@mail.tsinghua.edu.cn

{fawe, yuegao}@microsoft.com miaojing.shi@kcl.ac.uk

Fangyun Wei

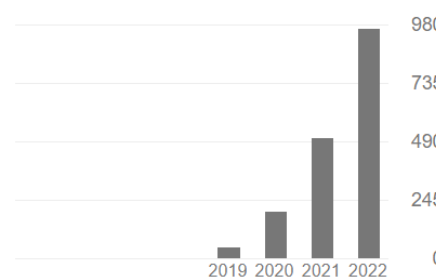
关注

Microsoft Research Asia
在 microsoft.com 的电子邮件经过验证
Computer Vision Deep Learning Machine Learning

标题	引用次数	年份
GCNet: Non-local networks meet squeeze-excitation networks and beyond Y Cao, J Xu, S Lin, F Wei, H Hu CVF International Conference on Computer Vision Workshop (ICCVW), 1971-1980	1090	2019
End-to-End Semi-Supervised Object Detection with Soft Teacher M Xu, Z Zhang, H Hu, J Wang, L Wang, F Wei, X Bai, Z Liu ICCV 2021	142	2021
Point-set anchors for object detection, instance segmentation and pose estimation F Wei, X Sun, H Li, J Wang, S Lin ECCV 2020	75	2020
Aligning Pretraining for Detection via Object-Level Contrastive Learning F Wei, Y Gao, Z Wu, H Hu, S Lin NeurIPS 2021 (Spotlight)	42	2021
RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder C Chi, F Wei, H Hu NeurIPS 2020 (Spotlight)	41	2020

引用次数

	总计	2017 年至今
引用	1739	1736
h 指数	16	16
i10 指数	20	20



开放获取的出版物数量 查看全部

1 篇文章 5 篇文章

无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

目录

研究背景一：目标检测

- 常用的数据集范式
  - 以COCO数据集为例 [1]

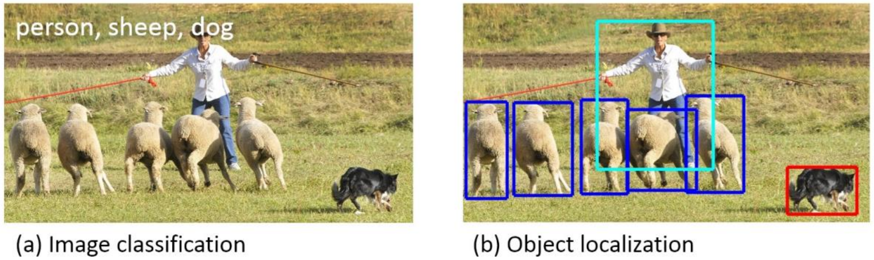


图1：COCO数据集对应的目标检测任务

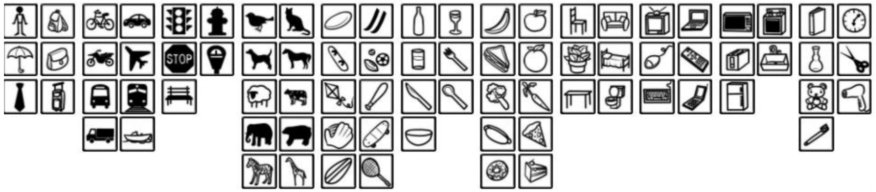


图2：COCO数据集包含的80类

- 常用目标检测方法
  - 以Faster-rcnn 为例 [2]

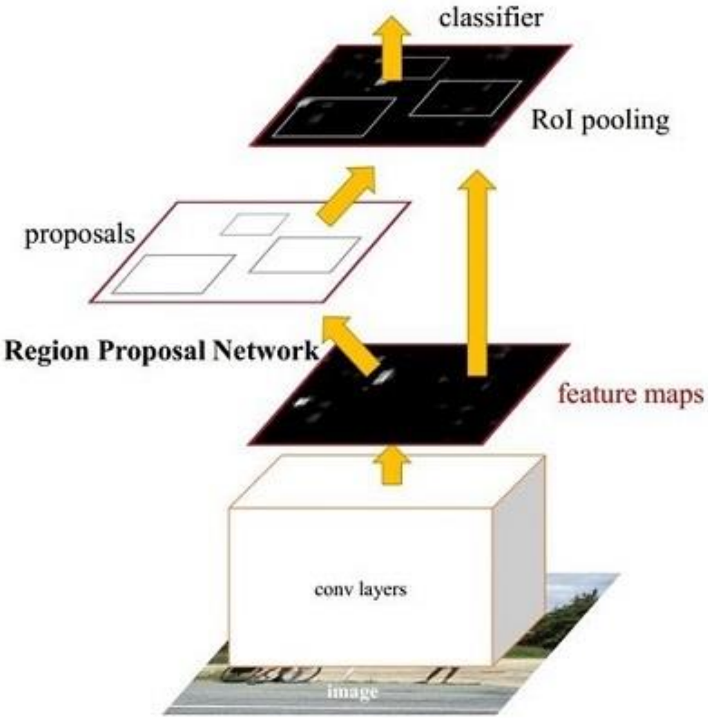


图3：Faster-rcnn方法示意图

[1]. Lin et al. Microsoft COCO: Common Objects in Context. ECCV2014
[2]. Ren et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS2015

研究背景二: 视觉语言模型

- 可用于开放类的视觉语言模型
  - 以CLIP 为例[1]

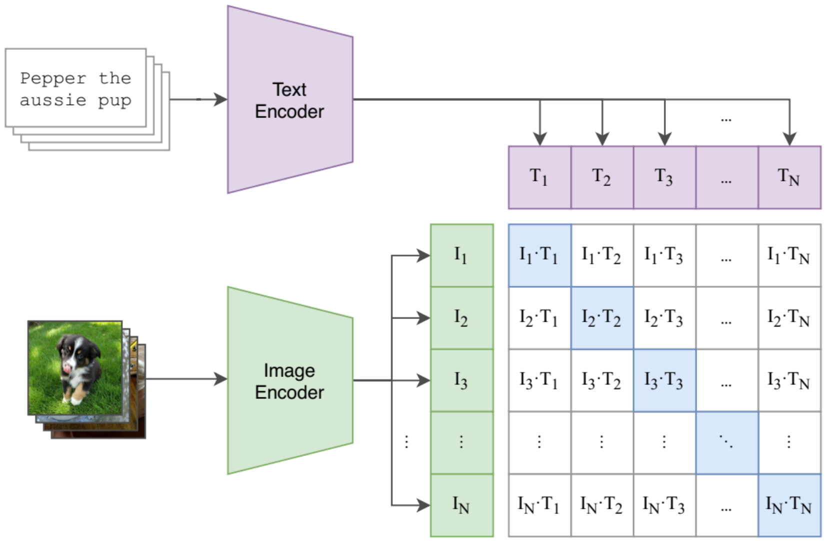


图1：CLIP模型的训练方式，4亿训练对

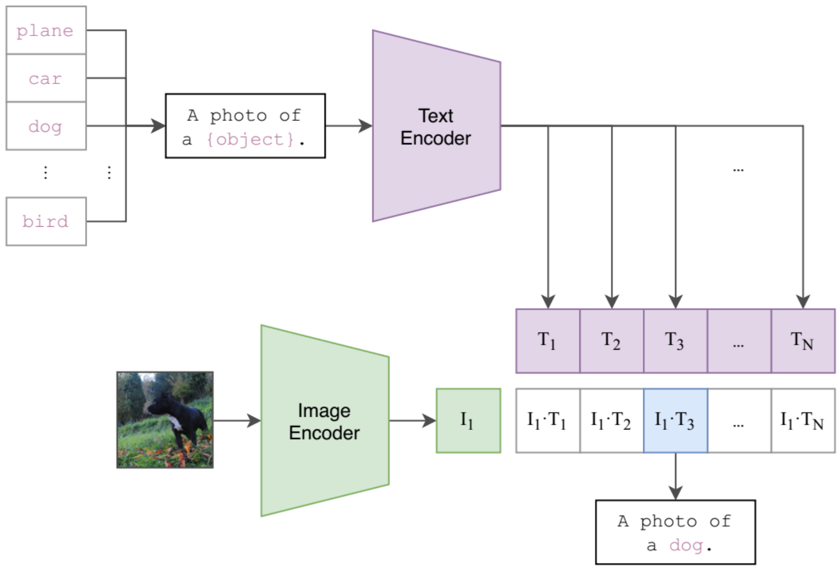


图2：CLIP模型的推理方式，可用于开放类

---

[1]. Radford et al. Learning transferable visual models from natural language supervision. ICML2021

研究背景三：开放类目标检测

- 全监督场景下，如何检测更多类别？
  - 增加新类的标注和数据
- 缺点：
  - 类别标签固定
  - 增加收集数据，成本高昂

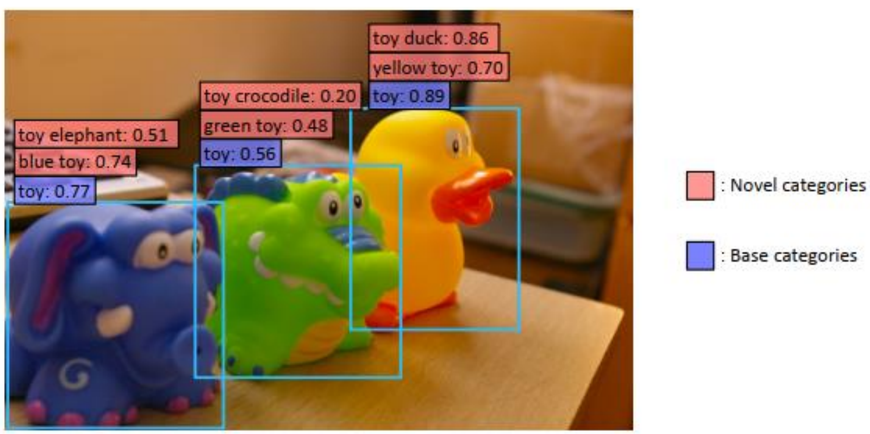


图1：不固定标签新类的目标检测

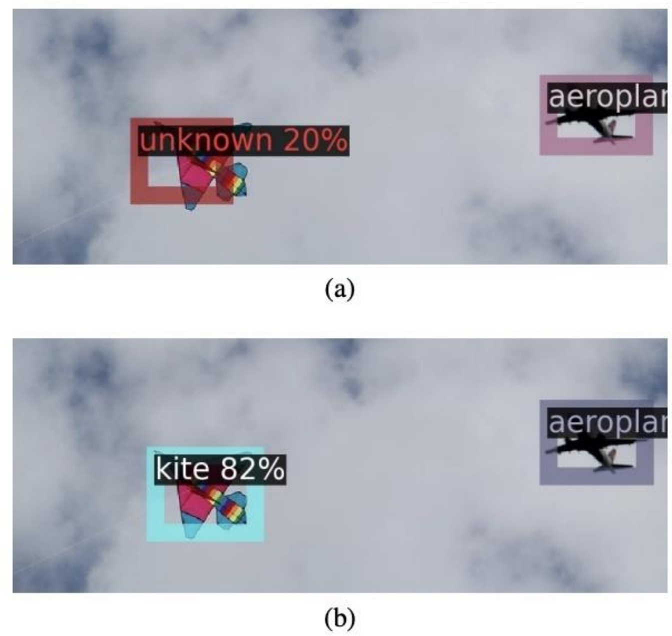


图2：新类检测需要重新收集数据示意

本研究baseline

- 直观方法 ViLD
  - 将新类和已知类名称交给文本编码器获得嵌入
  - 将候选框内的图像交给视觉编码器获得嵌入
  - 计算文本和图像嵌入的距离进行识别

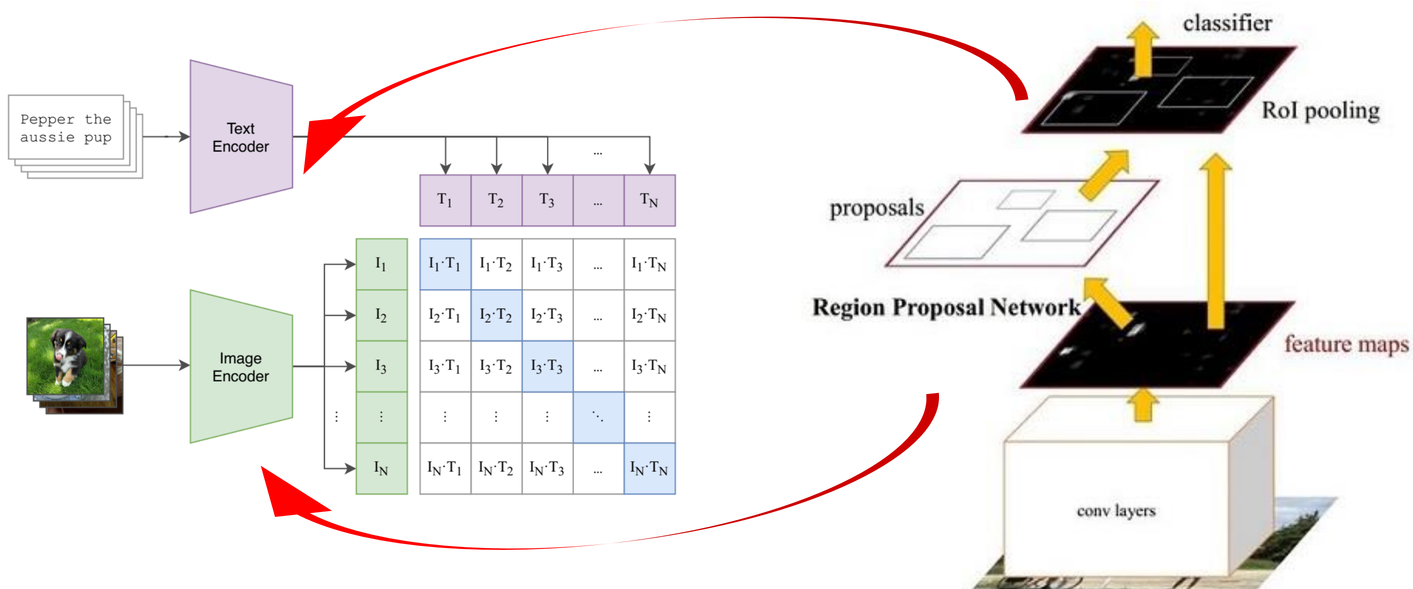


图1：开放类检测方法直观修方案示意图

本研究baseline

- ViLD架构

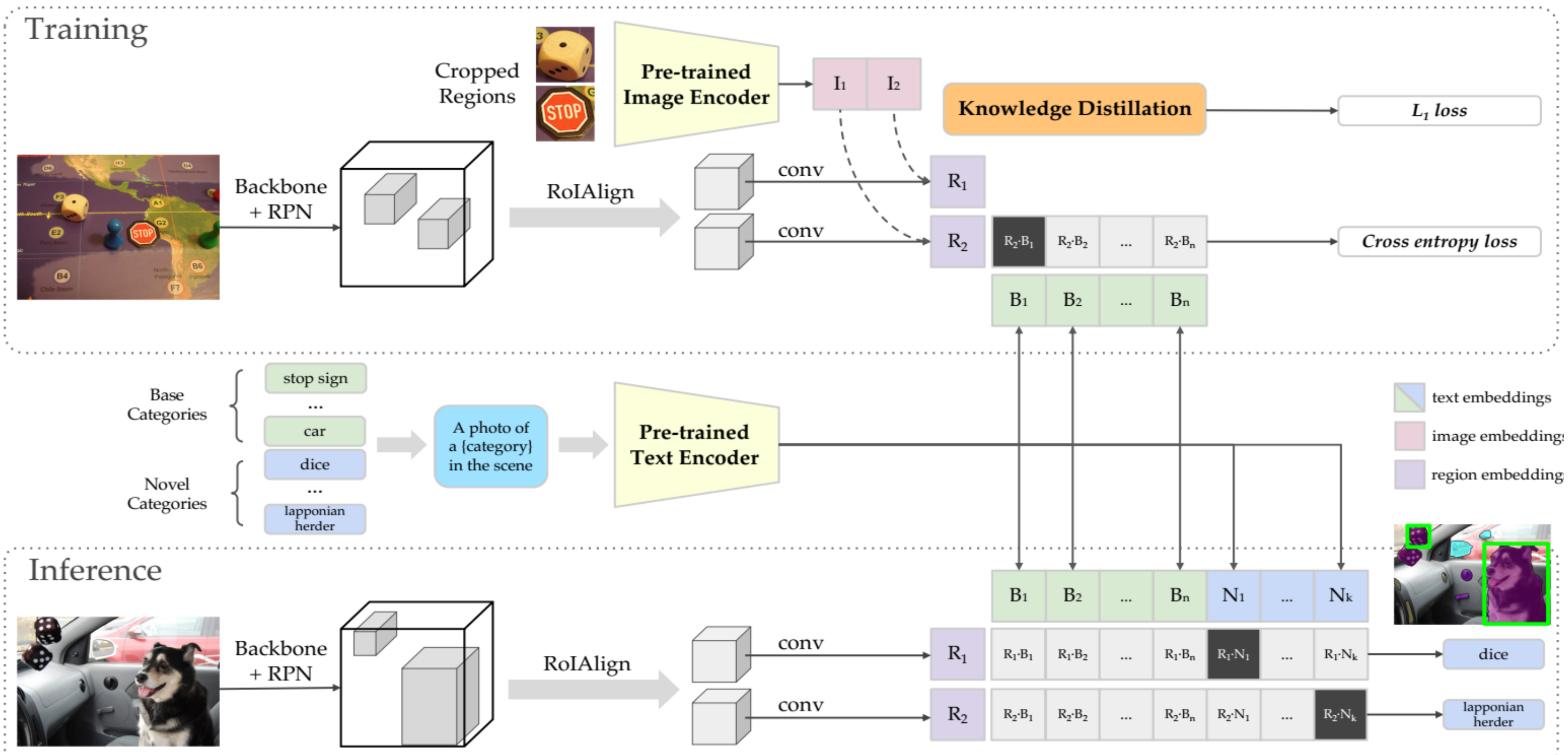


图1：ViLD方法整体框架，上半部分是训练阶段，下半部分是推理阶段，黄色预训练模型表示是固定的。

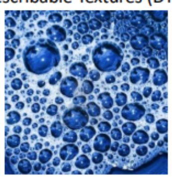
CoOp方案

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**

(a)

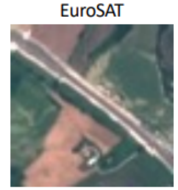
Flowers102 Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)



Describable Textures (DTD) Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)



EuroSAT Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

图1：CoOp的效果图

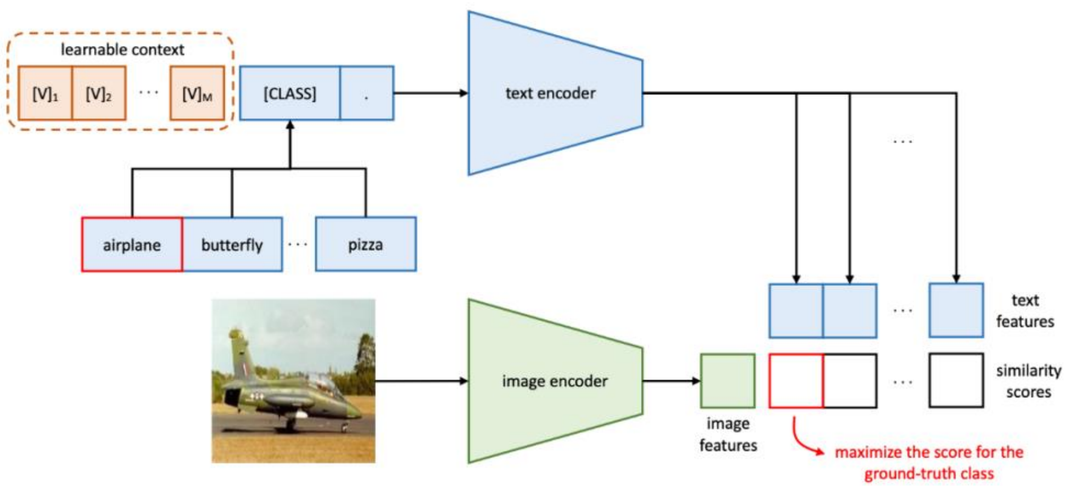


图2：CoOp的框架图

目录

研究方法

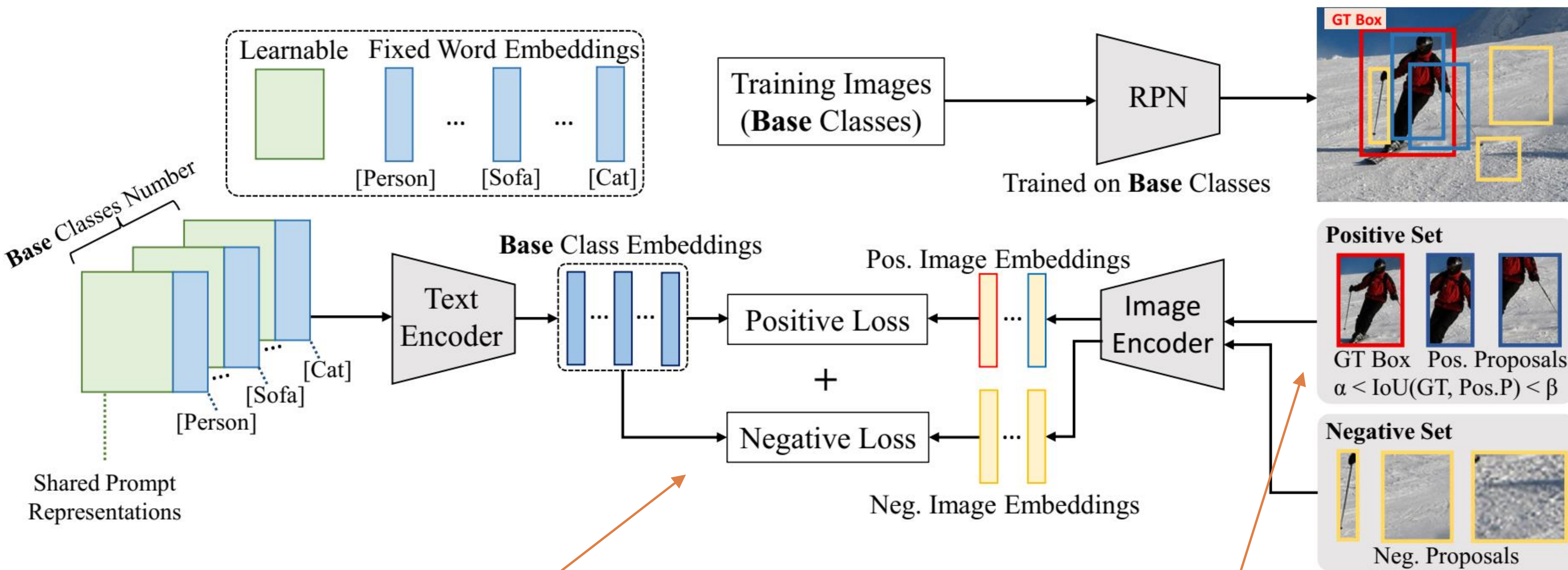
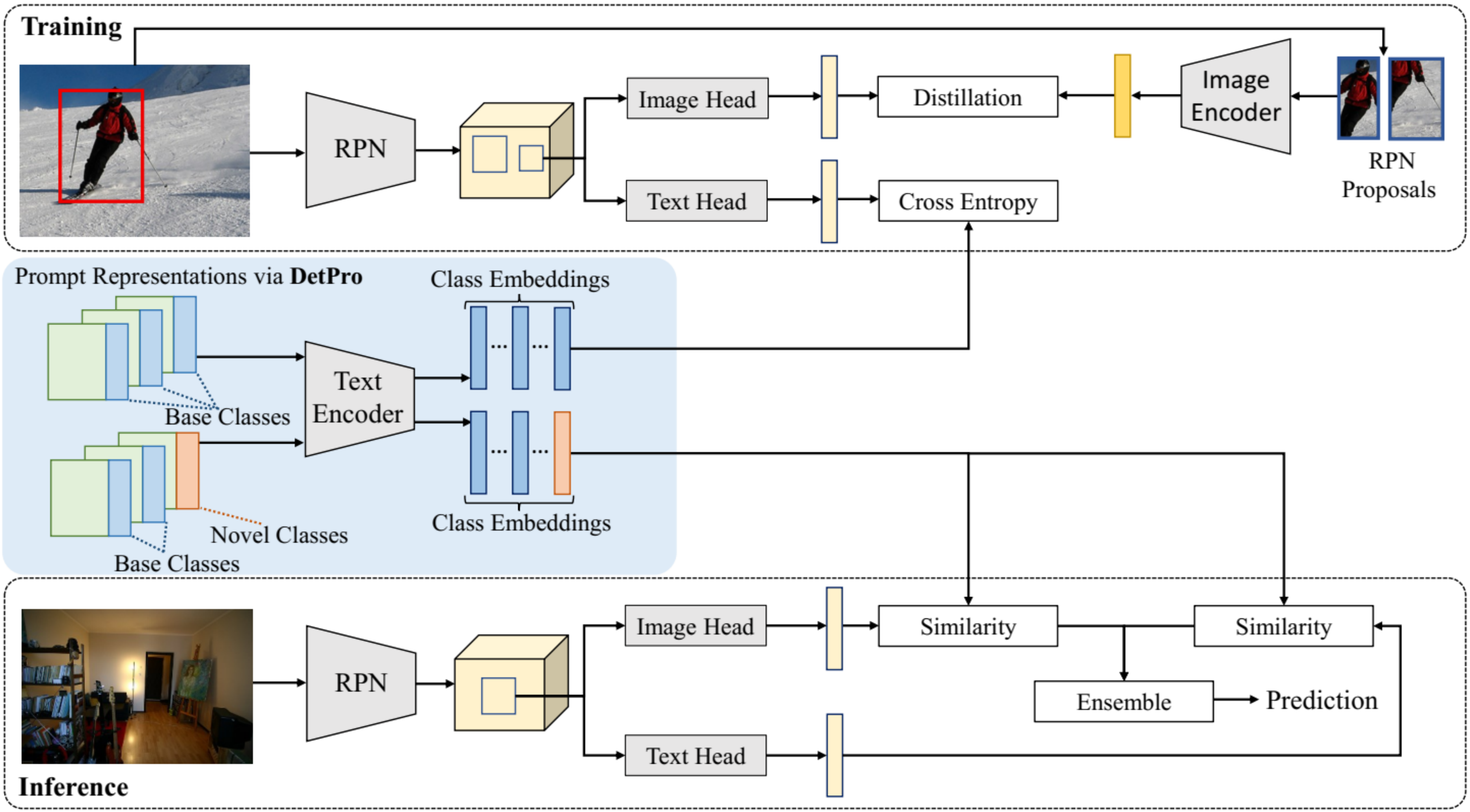


图1：DetPro的框架图

背景解释方案

上下文打分方案

研究方法



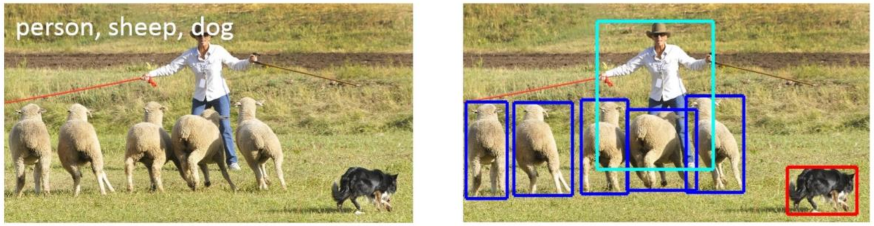
Training
RPN
Image Head
Distillation
Image Encoder
RPN Proposals
Text Head
Cross Entropy
Prompt Representations via DetPro
Class Embeddings
Base Classes
Text Encoder
Novel Classes
Base Classes
Class Embeddings
RPN
Image Head
Similarity
Similarity
Ensemble
Prediction
Text Head
Inference

目录

实验效果

- LVIS v1 [1]
  - 866个基类（frequent & common），337个新类（rare）

- COCO [2]
  - 48个基类，17个新类，移除不包含在WordNet的15类



(a) Image classification (b) Object localization

[1]. Gupta et al. Lvis: A dataset for large vocabulary instance segmentation. CVPR2019
[2]. Lin et al. Microsoft COCO: Common Objects in Context. ECCV2014

实验效果

- LVIS数据集

Method	Epoch	Detection AP_r	AP_c	AP_f	AP	Instance segmentation AP_r	AP_c	AP_f	AP
Supervised (base)	20	0.0	26.1	34.0	24.7	0.0	24.7	29.8	22.4
Supervised (base+novel)	20	15.5	25.5	33.6	27.0	16.4	24.6	30.6	25.5
ViLD (base) [7]	460	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
ViLD* (base) [7]	20	17.4	27.5	31.9	27.5	16.8	25.6	28.5	25.2
DetPro (base)	20	**20.8**	27.8	32.4	28.4	**19.8**	25.6	28.9	25.9

实验效果

- VOC, COCO, Object365 数据集

Method	Pascal VOC		COCO						Objects365					
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised	78.5	49.0	46.5	67.6	50.9	27.1	67.6	77.7	25.6	38.6	28.0	16.0	28.1	36.7
ViLD* [7]	73.9	**57.9**	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	**74.6**	**57.9**	**34.9**	**53.8**	**37.4**	**22.5**	**39.6**	**46.3**	**12.1**	**18.8**	**12.9**	**4.5**	**11.5**	**18.6**

Table 2. We evaluate the LVIS-trained model on Pascal VOC test set, COCO validation set and Object365 validation set.

实验效果

- Ablation Study

Background proposals	AP_r	AP_c	AP_f	AP
10%	**19.1**	25.4	28.2	**25.4**
30%	18.3	**25.6**	**28.4**	**25.4**
50%	17.8	**25.6**	**28.4**	**25.4**
100%	17.6	25.1	28.2	25.0

Table 4. Ablation on number of background proposals involved in DetPro training.

GT	FG	BG	AP_r	AP_c	AP_f	AP
✓			15.3	**25.4**	27.9	24.6
✓	✓		16.9	25.1	27.7	24.7
✓		✓	17.7	25.3	**28.2**	25.1
✓	✓	✓	**19.1**	**25.4**	**28.2**	**25.4**

Table 5. Ablation study on the involvement of different training data. ’GT’: ground-truth; ‘FG’: foreground; ‘BG’: background.

实验效果

Ablation Study

Length	AP_r	AP_c	AP_f	AP
4	18.7	24.9	28.2	25.1
8	**19.1**	**25.6**	**28.3**	25.2
16	17.7	**25.6**	**28.3**	**25.3**

Table 7. Ablation study on context lengths.

Position	AP_r	AP_c	AP_f	AP
Front	16.4	24.5	**28.3**	24.6
Middle	18.0	25.1	**28.3**	25.1
End	**19.1**	**25.4**	28.2	**25.4**

Table 8. Ablation study of inserting class token into different positions of prompt representation.

实验效果

图1：在LVIS，COCO，Object365数据集上的可视化实验。

目录

总结

- 总结

  - 本文对于prompt learning出发，针对两个细节改进：
    - 基于ViLT方法处理背景类使用统一表征的问题
    - 基于ViLT方法处理前景框过于粗糙的问题

  - 提升RPN对开放词汇目标的效果会是一个可能的改进方向

Table 1: **Training with only base categories achieves comparable average recall (AR) for novel categories on LVIS.** We compare RPN trained with base only *vs*. base+novel categories and report the bounding box AR.

Supervision	AR_r@100	AR_r@300	AR_r@1000
base	39.3	48.3	55.6
base + novel	41.1	50.9	57.0