开放词汇目标检测 CVPR2023

(Open-Vocabulary Object Detection)

报告人：徐静远

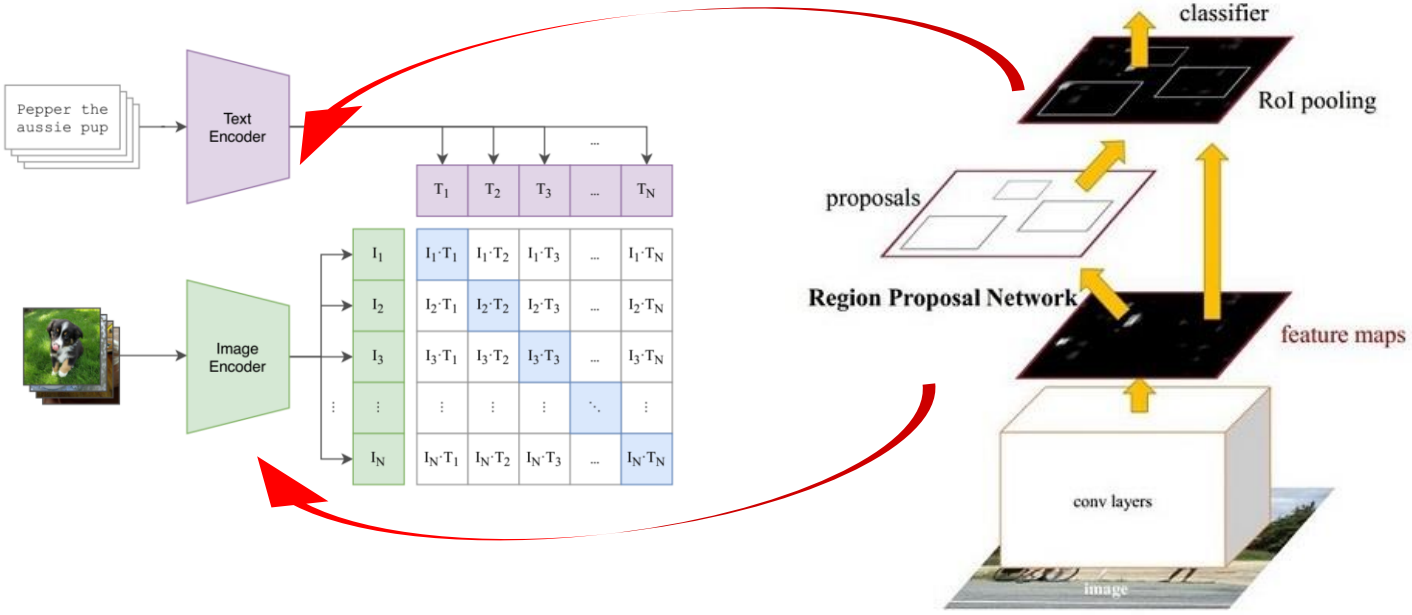
Contents

Preliminary

开放词汇检测任务

视觉语言模型VLM
（CLIP, ALGN）

检测模型Detector
(RCNN, DETR)



作者单位

Region-Aware Pretraining for Open-Vocabulary Object Detection with Vision Transformers:RO-VIT

Dahun Kim Anelia Angelova Weicheng Kuo
Google Research, Brain Team
{mcahny, anelia, weicheng}@google.com

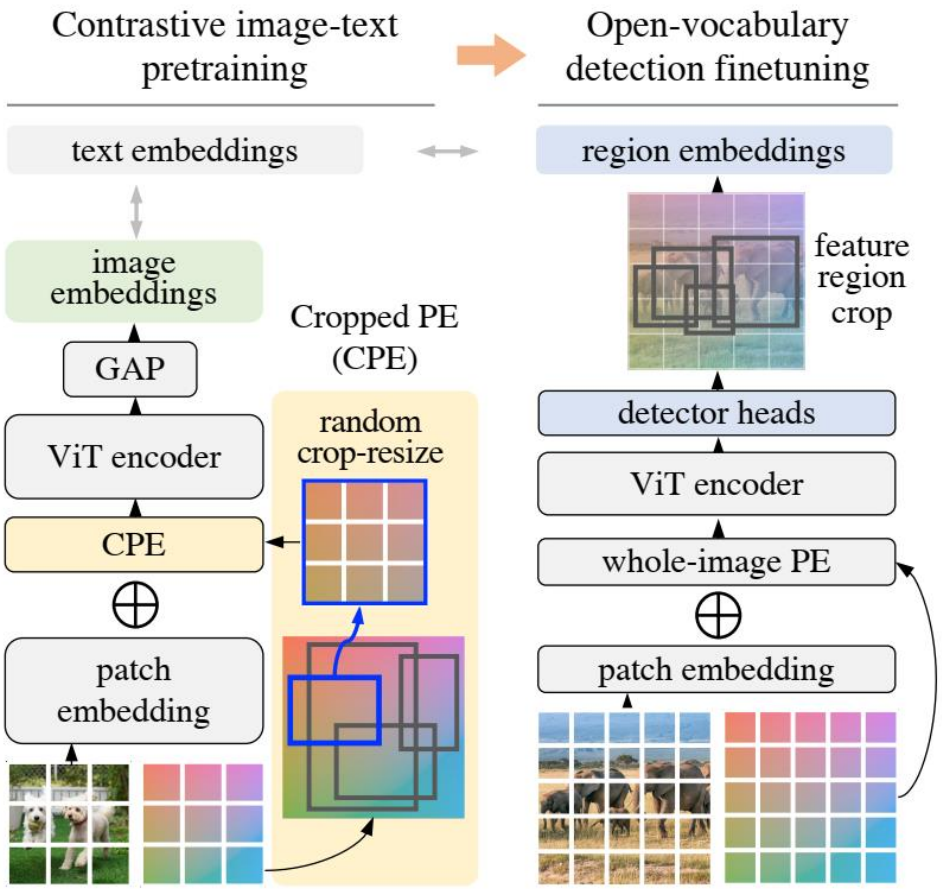
Aligning Bag of Regions for Open-Vocabulary Object Detection: BARON

Size Wu[1] Wenwei Zhang[1] Sheng Jin[2,3] Wentao Liu[3,4] Chen Change Loy[1*]
[1]S-Lab, Nanyang Technological University [2] The University of Hong Kong
[3] SenseTime Research and Tetras.AI [4] Shanghai AI Laboratory
{size001, wenwei001, ccloy}@ntu.edu.sg {jinsheng, liuwentao}@sensetime.com

Contents

Region-Aware Pretraining for Open-Vocabulary Object Detection with Vision Transformers (RO-ViT)

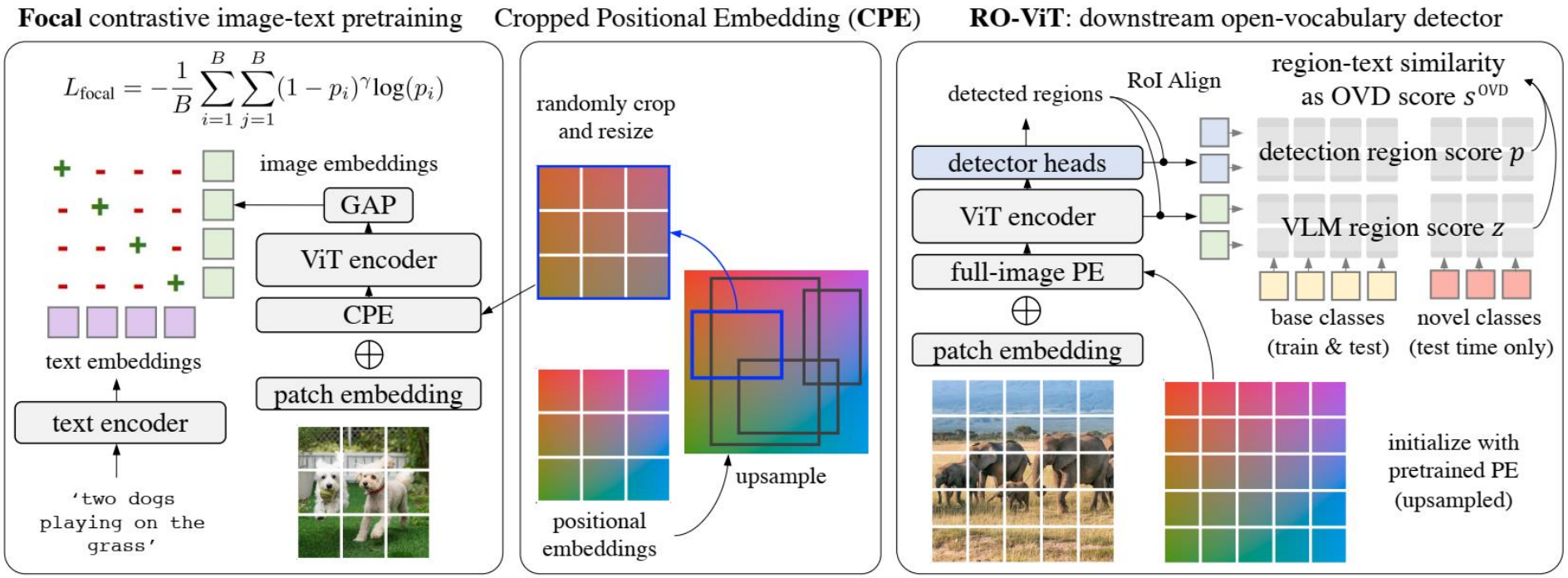
motivation



以往的视觉语言模型是为图像级别的任务设计（分类，检索），本文重新设计预训练范式

Region-Aware Pretraining for Open-Vocabulary Object Detection with Vision Transformers (RO-ViT)

method



在ALIGN模型上做了两点改变：
1.使用区域级的位置编码；
2.使用focal loss取代CEloss

Region-Aware Pretraining for Open-Vocabulary Object Detection with Vision Transformers (RO-ViT)

Results

method	pretrained model	detector backbone	**AP_r**	AP
ConvNet based:				
DetPro-Cascade [13]	ViT-B/32	R-50	20.0	27.0
Detic-CN2 [63]	ViT-B/32	R-50	24.6	32.4
RegionCLIP [60]	R-50x4	R-50x4	22.0	32.3
ViLD-Ens [19]	ViT-B/32	R-152	18.7	26.0
ViLD-Ens [19]	ViT-L/14	EffNet-B7	21.7	29.6
ViLD-Ens [19]	EffNet-B7	EffNet-B7	26.3	29.3
VL-PLM [57]	ViT-B/32	R-50	17.2	27.0
OV-DETR [53]	ViT-B/32	R-50	17.4	26.6
Rasheed *et al.* [41]	ViT-B/32	R-50	21.1	25.9
PromptDet [14]	ViT-B/32	R-50	21.4	25.3
ViT based:				
OWL-ViT [35]	ViT-H/14	ViT-H/14	23.3	35.3
OWL-ViT [35]	ViT-L/14	ViT-L/14	25.6	34.7
RO-ViT (ours)	ViT-B/16	ViT-B/16	28.0	30.2
RO-ViT (ours)	ViT-L/14	ViT-L/14†	31.4	34.0
RO-ViT (ours)	ViT-L/16	ViT-L/16	**32.1**	34.0

Table 1. **LVIS open-vocabulary object detection (mask APs).**

method	pretrained model	detector backbone	**novel AP**	AP
ConvNet based:				
ViLD [19]	ViT-B/32	R-50	27.6	51.3
OV-DETR [53]	ViT-B/32	R-50	29.4	52.7
w/ pseudo box labels:				
XPM *et al.* [25]	R-50	R-50	27.0	41.2
RegionCLIP [60] †	R-50x4	R-50x4	39.3	55.7
PromptDet [14]	ViT-B/32	R-50	26.6	50.6
VL-PLM [57]	ViT-B/32	R-50	34.4	53.5
Rasheed *et al.* [41] ‡	ViT-B/32	R-50	36.9	51.5
w/ weak supervision:				
Detic-CN2 [63]	ViT-B/32	R-50	24.6	32.4
ViT based:*				
RO-ViT (ours)	ViT-B/16	ViT-B/16	30.2	41.5
RO-ViT (ours)	ViT-L/16	ViT-L/16	33.0	47.7

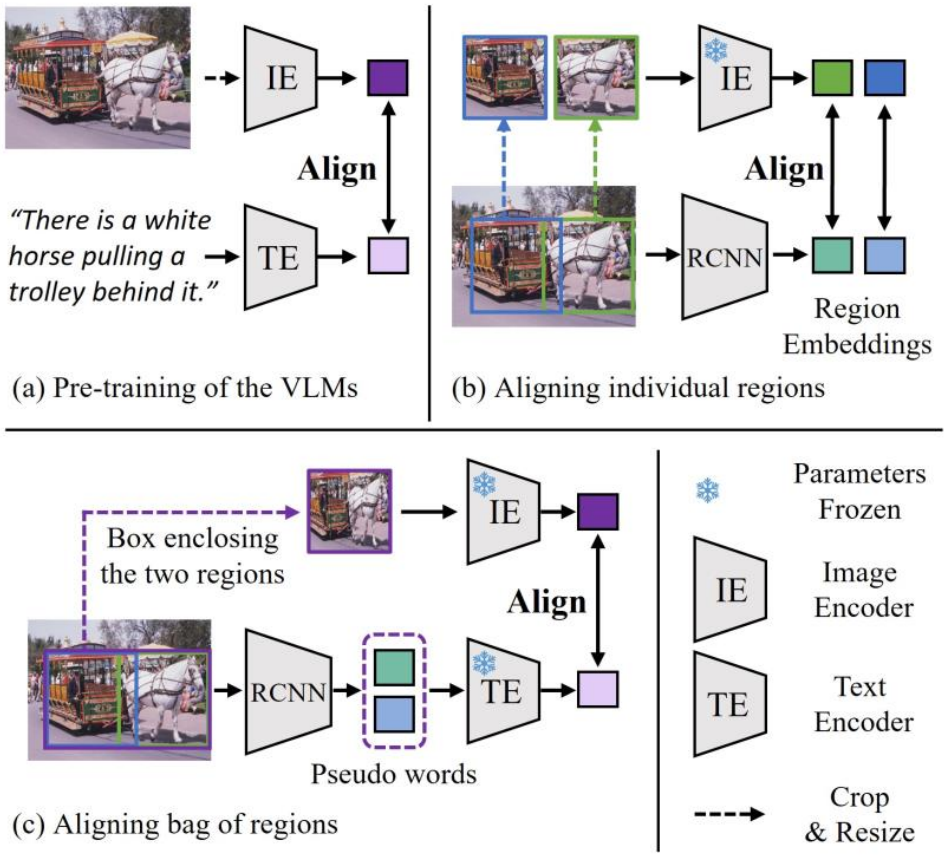
Table 2. **COCO open-vocabulary object detection (box AP50).**

预训练bs=16384, iter=500k，ALIGN：1024TPU

Contents

Aligning Bag of Regions for Open-Vocabulary Object Detection (BARON)

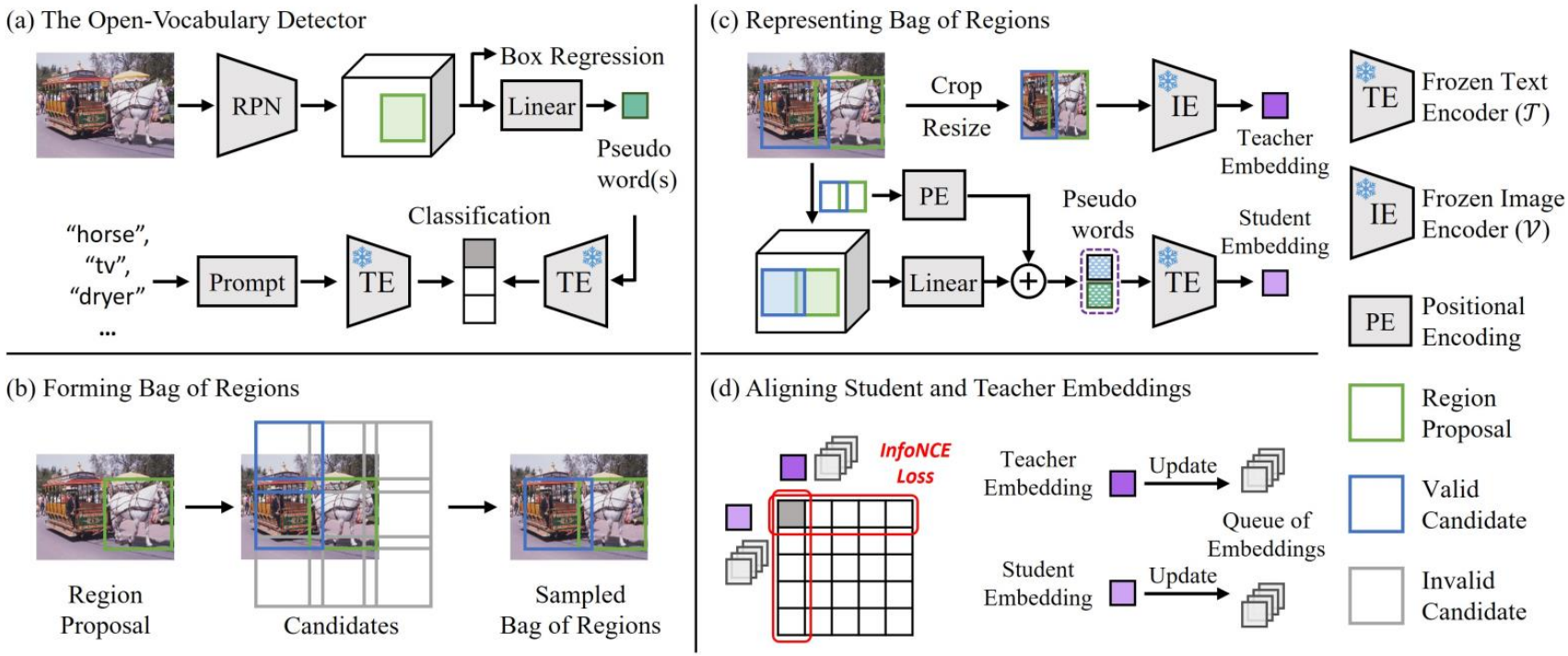
motivation



以往的ovod方法试图对齐区域图像特征与文本特征。
本文希望对齐区域图像特征与伪文本特征。

Aligning Bag of Regions for Open-Vocabulary Object Detection (BARON)

Method



(a) BARON整体基于faster-rcnn的框架图; (b) 训练过程中对齐伪文本
(c) 形成区域袋；(d) InfoNCE对比训练

Aligning Bag of Regions for Open-Vocabulary Object Detection (BARON)

Experiment

Table 2. Comparison with state-of-the-art methods on OV-LVIS. * denotes the re-implemented ViLD [15] reported in DetPro [10].

Method	Ensemble	Learned Prompt	Object Detection				Instance segmentation			
			AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [15]	-	-	16.3	21.2	31.6	24.4	16.1	20.0	28.3	22.5
OV-DETR [52]	-	-	-	-	-	-	17.4	25.0	32.5	26.6
BARON (Ours)	-	-	**17.3**	25.6	31.0	26.3	**18.0**	24.4	28.9	25.1
ViLD [15]	✓	-	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
ViLD* [15]	✓	-	17.4	27.5	31.9	27.5	16.8	25.6	28.5	25.2
BARON (Ours)	✓	-	**20.1**	28.4	32.2	28.4	**19.2**	26.8	29.4	26.5
DetPro [10]	✓	✓	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
BARON (Ours)	✓	✓	**23.2**	29.3	32.5	29.5	**22.6**	27.6	29.8	27.6

Table 1. Comparison with state-of-the-art methods on OV-COCO benchmark. We separately compare our approach with methods distilling knowledge from CLIP and approaches using COCO caption. † means using proposals produced by MAVL [34].

Method	Supervision	Backbone	Detector	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
ViLD [15]	CLIP	ResNet50-FPN	FasterRCNN	27.6	59.5	51.2
OV-DETR [52]	CLIP	ResNet50	DeformableDETR	29.4	61.0	52.7
BARON (Ours)	CLIP	ResNet50-FPN	FasterRCNN	**34.0**	60.4	53.5
OVR-CNN [53]	Caption	ResNet50-C4	FasterRCNN	22.8	46.0	39.9
RegionCLIP [56]	Caption	ResNet50-C4	FasterRCNN	26.8	54.8	47.5
Detic [58]	Caption	ResNet50-C4	FasterRCNN	27.8	51.1	45.0
PB-OVD [13]	Caption	ResNet50-C4	FasterRCNN	30.8	46.1	42.1
VLDet [28]	Caption	ResNet50-C4	FasterRCNN	32.0	50.6	45.8
BARON (Ours)	Caption	ResNet50-C4	FasterRCNN	**33.1**	54.8	49.1
Rasheed *et al.* [41]†	CLIP + Caption	ResNet50-C4	FasterRCNN	36.6	54.0	49.4
BARON (Ours)†	CLIP + Caption	ResNet50-C4	FasterRCNN	**42.7**	54.9	51.7

实验要求：类似于DetPro 350卡时

Contents

总结

省时省力
泛化能力弱

ViLD, DetPro, F-VLM, BARON

基于已有模型**微调**的

基于**预训练**的 （cvpr22->23 :9篇->32篇）

泛化能力强
耗时耗力

RegionCLIP, GLIP, RO-ViT

思考

- 提供两种不同任务的路线，理解不同的 motivation，抛砖引玉。
- 基于预训练路线的，天花板高一点，但是训练难度大、要求高
- 基于已有模型路线的，依赖于CLIP模型能力，在该模型基础上寻找motivation做改进
- 固定VLM成为固定范式，改进集中在设计head, loss,实现区域级别的对齐