Grounded Language-Image Pre-training

CVPR 2022

报告人：徐静远

目录

目录

作者介绍

Liunian Harold Li[*1†], **Pengchuan Zhang**[*2♠], **Haotian Zhang**[*3†], **Jianwei Yang**[2], **Chunyuan Li**[2], **Yiwu Zhong**[4†],
Lijuan Wang[5], **Lu Yuan**[5], **Lei Zhang**[6], **Jenq-Neng Hwang**[3], **Kai-Wei Chang**[1], **Jianfeng Gao**[2]
[1]UCLA, [2]Microsoft Research, [3]University of Washington,
[4]University of Wisconsin-Madison, [5]Microsoft Cloud and AI, [6]International Digital Economy Academy

Pengchuan Zhang

FOLLOW

Meta AI
Verified email at fb.com - Homepage
Machine learning   Deep learning

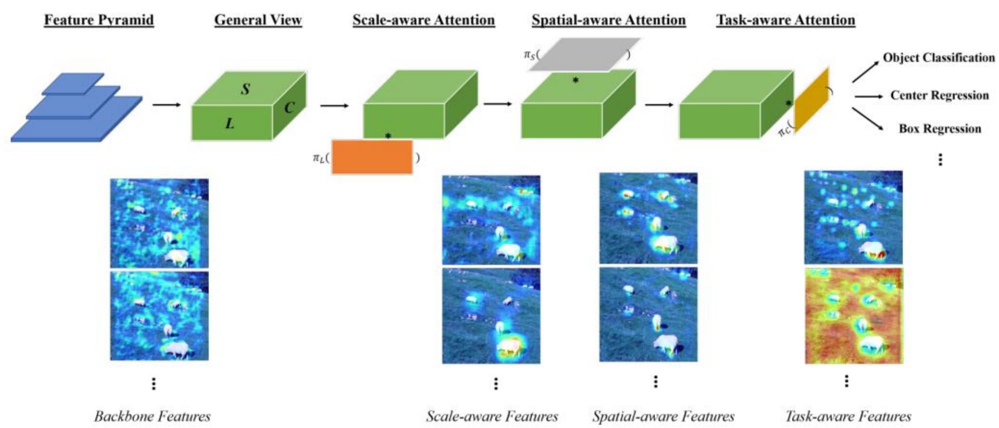
TITLE	CITED BY	YEAR
Attngan: Fine-grained text to image generation with attentional generative adversarial networks T Xu, P Zhang, Q Huang, H Zhang, Z Gan, X Huang, X He Proceedings of the IEEE conference on computer vision and pattern ...	1436	2018
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks X Li, X Yin, C Li, P Zhang, X Hu, L Zhang, L Wang, H Hu, L Dong, F Wei, ... Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23 ...	1263	2020
Scaling vision transformers X Zhai, A Kolesnikov, N Houlsby, L Beyer Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern ...	809 *	2022
Vinvl: Revisiting visual representations in vision-language models P Zhang, X Li, X Hu, J Yang, L Zhang, L Wang, Y Choi, J Gao Proceedings of the IEEE/CVF conference on computer vision and pattern ...	530	2021
Provably robust deep learning via adversarially trained smoothed classifiers H Salman, J Li, I Razenshteyn, P Zhang, H Zhang, S Bubeck, G Yang Advances in Neural Information Processing Systems 32	415	2019
Florence: A new foundation model for computer vision L Yuan, D Chen, YL Chen, N Codella, X Dai, J Gao, H Hu, X Huang, B Li, ... arXiv preprint arXiv:2111.11432	362	2021

目录

研究背景一：目标检测

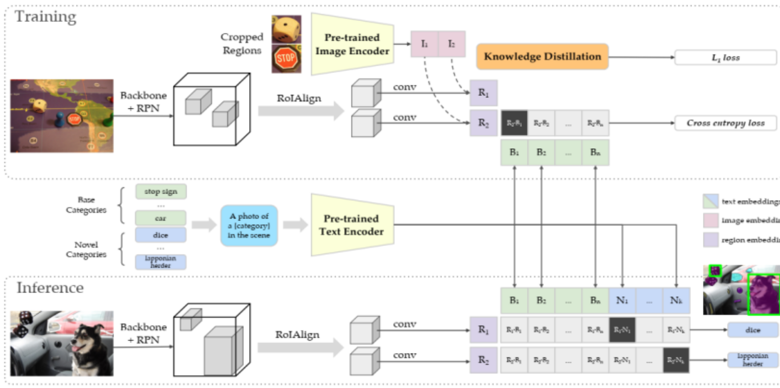
- 数据集
  - COCO：200k图，80类
  - OpenImage：1.9M图，600类
  - Ojbect365：2M图，365类

- 方法：
  - Faster-RCNN系列
  - Swin-Transformer系列
  - Dyhead

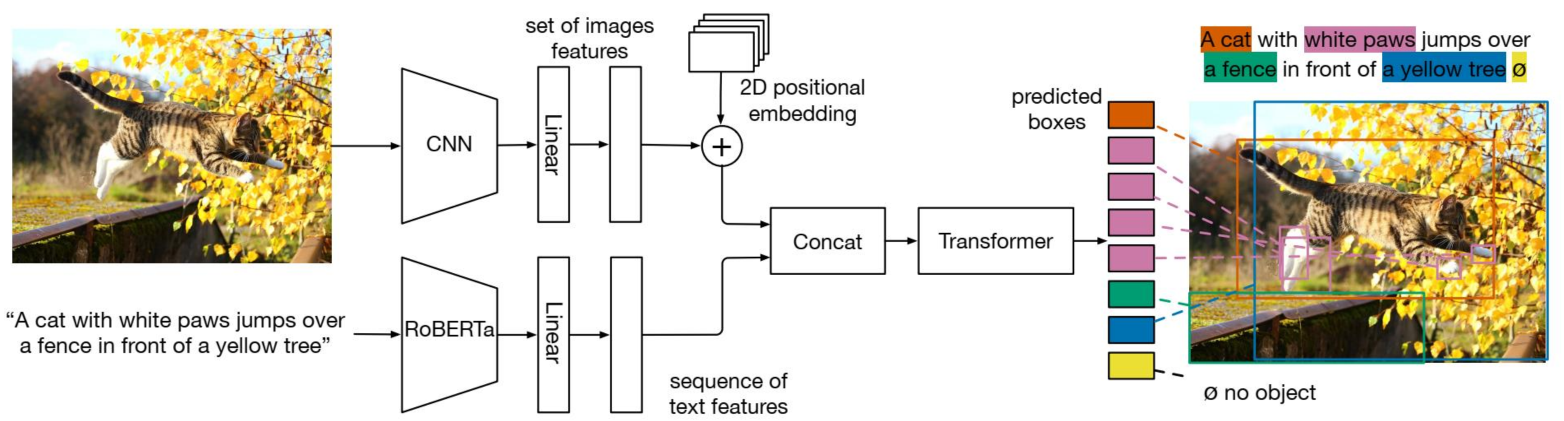


研究背景二: 开放词汇检测

- 路线一：从跨模态模型蒸馏知识
  - ViLd, DetPro, F-VLM, Baron



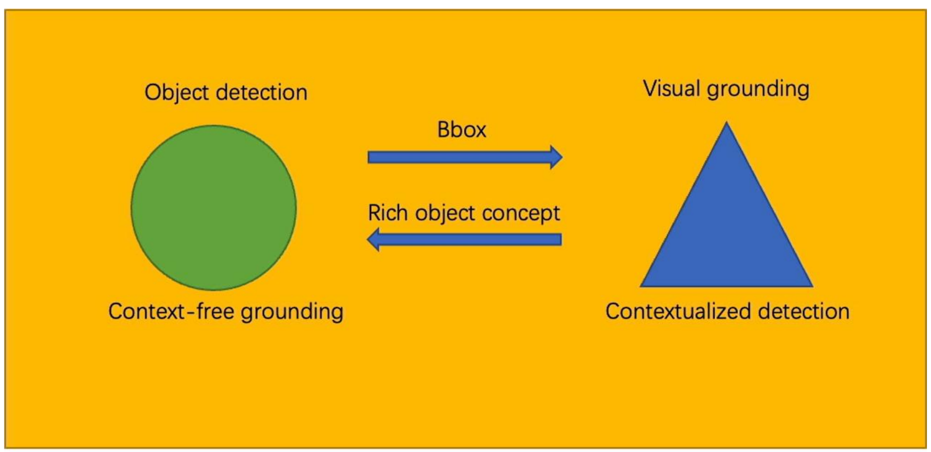
- 路线二：预训练跨模态模型
  - MDETR



目录

研究方法

- Motivation
  - 统一Object Detection和Phrase Grounding的任务
  - 获得更细粒度的类CLIP模型
  - 用27M图文对训练（3M有标注+Cap24M）



Unified framework

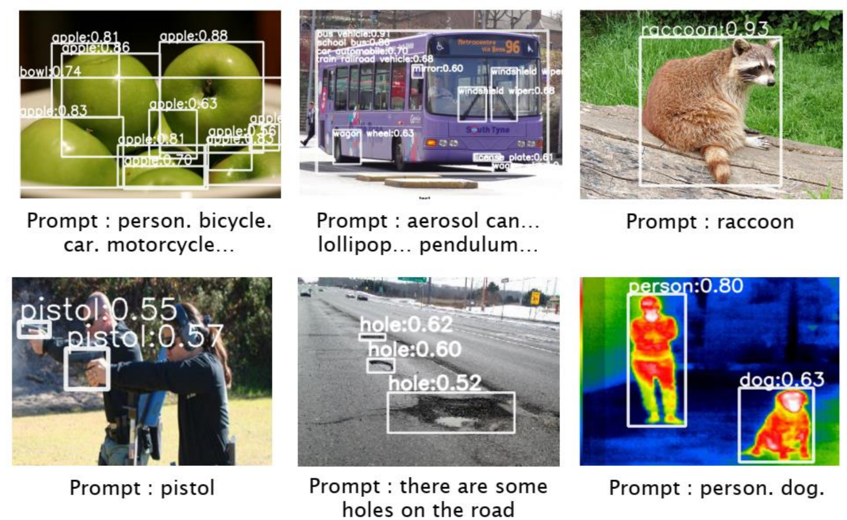
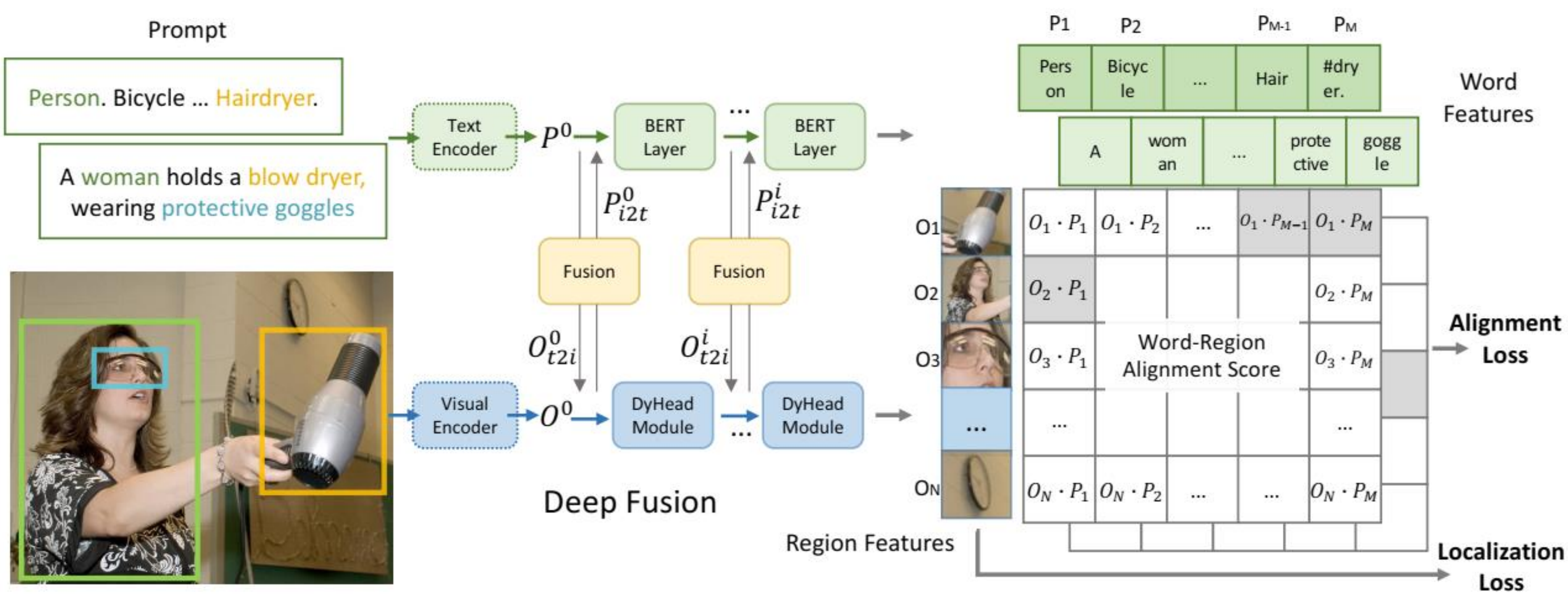


Figure 1. GLIP zero-shot transfers to various detection tasks, by writing the categories of interest into a text prompt.

研究方法

- 训练流程
  - 文本编码器使用Bert，图像编码器采用Swin，Head采用dyhead
  - 设计了文本和视觉特征的融合模块，X-MHA
  - 对齐损失（识别）和定位损失



研究方法

具体方法流程

- 在Dyhead-T基础上增加Bert的语言编码器，将分类器改为文本视觉对齐；得到GLIP-T(A) 掉点0.7
- 增加了视觉文本fusion模块；得到GLIP-T(B)提点2
- 合并Grounding任务，使用Grounding数据集丰富语义内容；得到GLIP-T(C)提点1.8
- 增加Caption数据集，用NLP parser定位名词，以GLIP-T为老师提供bbox伪标签，训练GLIP-L；得到GLIP-L提点2.1。

Model	Backbone	Deep Fusion	Pre-Train Data Detection	Pre-Train Data Grounding	Pre-Train Data Caption
GLIP-T (A)	Swin-T	✗	Objects365	-	-
GLIP-T (B)	Swin-T	✓	Objects365	-	-
GLIP-T (C)	Swin-T	✓	Objects365	GoldG	-
GLIP-T	Swin-T	✓	Objects365	GoldG	Cap4M
GLIP-L	Swin-L	✓	FourODs	GoldG	Cap24M

Table 1. A detailed list of GLIP model variants.

Model	Backbone	Pre-Train Data	Zero-Shot 2017val	Fine-Tune 2017val / test-dev
Faster RCNN	RN50-FPN	-	-	40.2 / -
Faster RCNN	RN101-FPN	-	-	42.0 / -
DyHead-T [10]	Swin-T	-	-	49.7 / -
DyHead-L [10]	Swin-L	-	-	58.4 / 58.7
DyHead-L [10]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [65]	Swin-L	O365,SS-COCO	-	60.7 / 61.3
DyHead-T	Swin-T	O365	43.6	53.3 / -
GLIP-T (A)	Swin-T	O365	42.9	52.9 / -
GLIP-T (B)	Swin-T	O365	44.9	53.8 / -
GLIP-T (C)	Swin-T	O365,GoldG	**46.7**	55.1 / -
GLIP-T	Swin-T	O365,GoldG,Cap4M	46.3	54.9 / -
GLIP-T	Swin-T	O365,GoldG,CC3M,SBU	46.6	**55.2** / -
GLIP-L	Swin-L	FourODs,GoldG,Cap24M	**49.8**	**60.8** / 61.0
GLIP-L	Swin-L	FourODs,GoldG+,COCO	-	- / **61.5**

目录

实验效果

Model	Backbone	MiniVal [23]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [23]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [23]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [15]	RN50	-	-	-	-	12.3	24.3	32.4	25.4
GLIP-T (A)	Swin-T	14.2	13.9	23.4	18.5	6.0	8.0	19.4	12.3
GLIP-T (B)	Swin-T	13.5	12.8	22.2	17.8	4.2	7.6	18.6	11.3
GLIP-T (C)	Swin-T	17.7	19.5	**31.0**	24.9	7.5	11.6	**26.1**	16.5
GLIP-T	Swin-T	**20.8**	**21.4**	**31.0**	**26.0**	**10.1**	**12.5**	25.5	**17.2**
GLIP-L	Swin-L	**28.2**	**34.3**	**41.5**	**37.3**	**17.1**	**23.3**	**35.4**	**26.9**

Table 3. Zero-shot domain transfer to LVIS. While using no LVIS data, GLIP-T/L outperforms strong supervised baselines (shown in gray). Grounding data (both gold and self-supervised) bring large improvements on APr.

Row	Model	Data	Val			Test		
			R@1	R@5	R@10	R@1	R@5	R@10
1	MDETR-RN101	GoldG+	82.5	92.9	94.9	83.4	93.5	95.3
2	MDETR-ENB5	GoldG+	83.6	93.4	95.1	84.3	93.9	95.8
3	GLIP-T	GoldG	84.0	95.1	96.8	84.4	95.3	97.0
4		O365,GoldG	84.8	94.9	96.3	85.5	95.4	96.6
5		O365,GoldG,Cap4M	**85.7**	**95.4**	**96.9**	**85.7**	**95.8**	**97.2**
6	GLIP-L	FourODs,GoldG,Cap24M	**86.7**	**96.4**	**97.9**	**87.1**	**96.9**	**98.1**

Table 4. Phrase grounding performance on Flickr30K entities. GLIP-L outperforms previous SoTA by 2.8 points on test R@1.

实验效果

- COCO数据集

Method	Training source	Novel AP	AP
WSDDN (Bilen & Vedaldi, 2016)	image-level labels in $C_B \cup C_N$	19.7	19.6
Cap2Det (Ye et al., 2019)		20.3	20.1
ZSD (Bansal et al., 2018)	instance-level labels in C_B	0.31	24.9
DELO (Zhu et al., 2020)		3.41	13.0
PL (Rahman et al., 2020)		4.12	27.9
OVR-CNN (Zareian et al., 2021)	image captions in $C_B \cup C_N$ instance-level labels in C_B	22.8	39.9
CLIP-RPN (Gu et al., 2022)	CLIP image-text pairs instance-level labels in C_B	26.3	27.8
ViLD (Gu et al., 2022)		27.6	51.3
Detic* (Zhou et al., 2022c)		27.8	45.0
RegionCLIP‡ (Zhong et al., 2022)		**31.4**	50.4
RegionCLIP† (Zhong et al., 2022)		26.8	47.5
RegionCLIP* (Zhong et al., 2022)		14.2	42.7
F-VLM (Ours)		28.0	39.6

实验效果

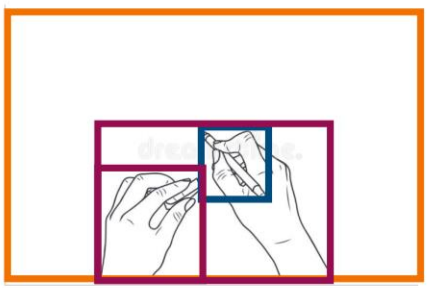
- 训练需求

Model	Fusion	Inference (P100) Speed	Inference (P100) Memory	Train (V100) Speed	Train (V100) Memory
GLIP-T	✗	4.84 FPS	1.0 GB	2.79 FPS	11.5 GB
	✓	2.52 FPS	2.4 GB	1.62 FPS	16.0 GB
GLIP-L	✗	0.54 FPS	4.8 GB	1.27 FPS	19.7 GB
	✓	0.32 FPS	7.7 GB	0.88 FPS	23.4 GB

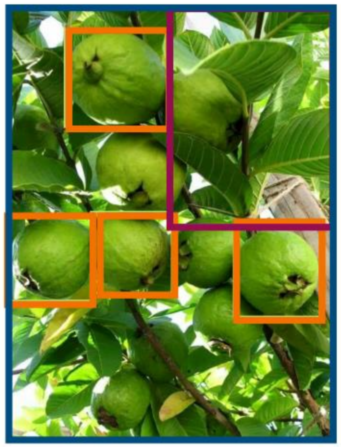
1.62*84K*8卡
=1M数据/天

Table 7. Computational cost of language-aware deep fusion. For speed, we report FPS, which is the number of images processed per second per GPU (higher is better). For memory consumption, we report the GPU memory used in GB (lower is better). Deep fusion brings less than 1x additional computational cost.

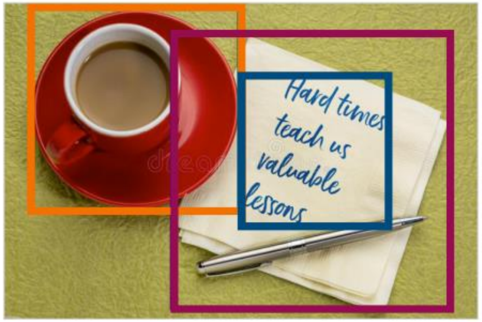
实验效果



sketch illustration - female hands write with a pen. arm, art, background, black, care, concept, counting, design, drawing, finger, fingers, five, gesture royalty free illustration



dwarf fruit tress are perfect for small spaces. here are 10 dwarf fruit trees which you can easily grow on your porch, or in containers or on the terrace. banana plants, fruit plants, fruit garden, garden trees, fruit and veg, fruits and vegetables, fresh fruit, apple plant, guava tree



hard times teach us valuable lessons. handwriting on a napkin with a cup of coffee stock photos

person battles with person in the production sedans

save the straws classic t-shirt

this week i'm going to share 20 ideas with you. 20 different lunchbox ideas. packing school lunch is about nourishment.

目录

总结

- 总结
  - 本文训练大模型思路，从sota的小模型开始，incremental成长
    - 修改V-L融合模块
    - 增加grounding数据集
    - 增加Caption数据集
    - 通过蒸馏形成大模型
  - 在COCO上的zeroshot表现不足，finetune效果较好