组会_20251119

一、上周工作

已将数据集和代码迁移至新硬盘，已重新恢复代码运行环境

二、论文阅读

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

摘要


```
The cost of vision-and-language pre-training has become increasingly
prohibitive due to end-to-end training of large-scale models. This paper
proposes BLIP-2, a generic and efficient pre-training strategy that
bootstraps vision-language pre-training from off-the-shelf frozen pre-
trained image encoders and frozen large language models. BLIP-2 bridges the
modality gap with a lightweight Querying Transformer, which is pre-trained
in two stages. The first stage bootstraps vision-language representation
learning from a frozen image encoder. The second stage bootstraps vision-to-
language generative learning from a frozen language model. BLIP-2 achieves
state-of-the-art performance on various vision-language tasks, despite
having significantly fewer trainable parameters than existing methods. For
example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with
54x fewer trainable parameters. We also demonstrate the model's emerging
capabilities of zero-shot image-to-text generation that can follow natural
language instructions.
```


引用信息（BibTeX格式）

```
@inproceedings{li2023blip,
  title={Blip-2: Bootstrapping language-image pre-training with frozen image
encoders and large language models},
  author={Li, Junnan and Li, Dongxu and Savarese, Silvio and Hoi, Steven},
  booktitle={International conference on machine learning},
  pages={19730--19742},
  year={2023},
  organization={PMLR}
}
```

本论文解决什么问题

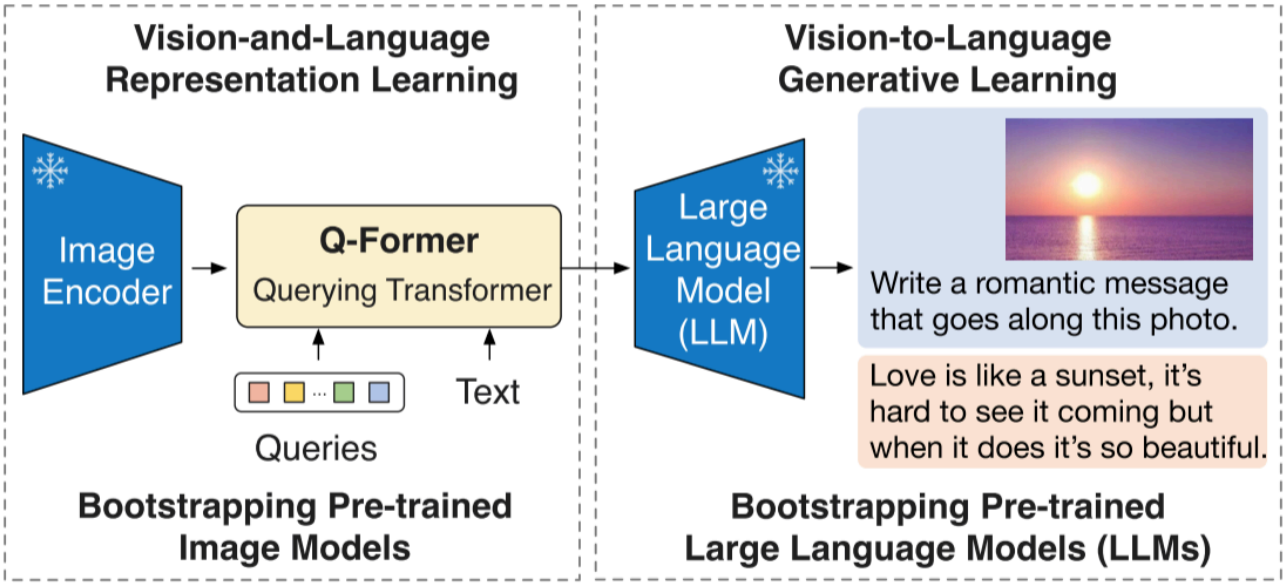
CLIP等多模态大模型要训练视觉和文本编码器，需要巨量数据和计算资源，

BLIP-2旨在解决如何让LLM具备视觉理解和多模态生成能力，而无需训练或修改庞大的LLM。具体问题包括：
1、如何高效地对齐视觉和语言特征，避免大规模端到端多模态训练的高成本。
2、如何将视觉特征输入到冻结的 LLM 中，使其能理解图像内容。
3、如何构建通用、多任务的多模态框架，实现图像描述、视觉问答、跨模态检索和多模态对话能力。

本文采用什么方法及其优缺点

1、模型架构

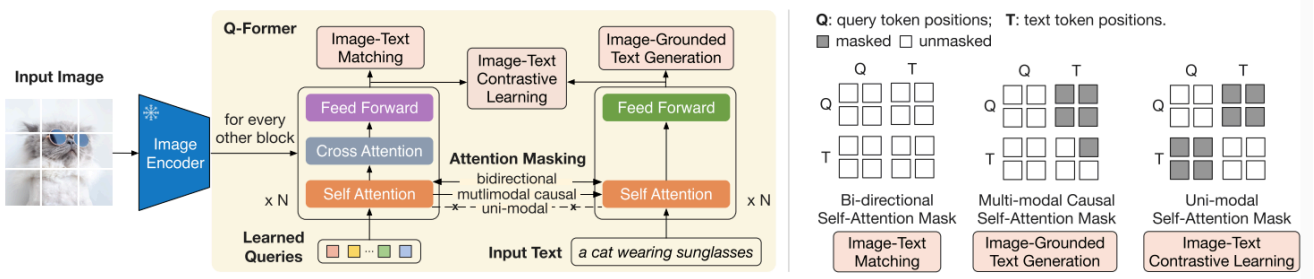
BLIP-2 由预训练的Image Encoder，预训练的Large Language Model，和一个可学习的 Q-Former 组成。Q-Former是BLIP2的核心之处，它负责弥合视觉和语言两种模态的差距，由Image Transformer和Text Transformer两个子模块构成，它们共享相同自注意力层。



2、训练方法

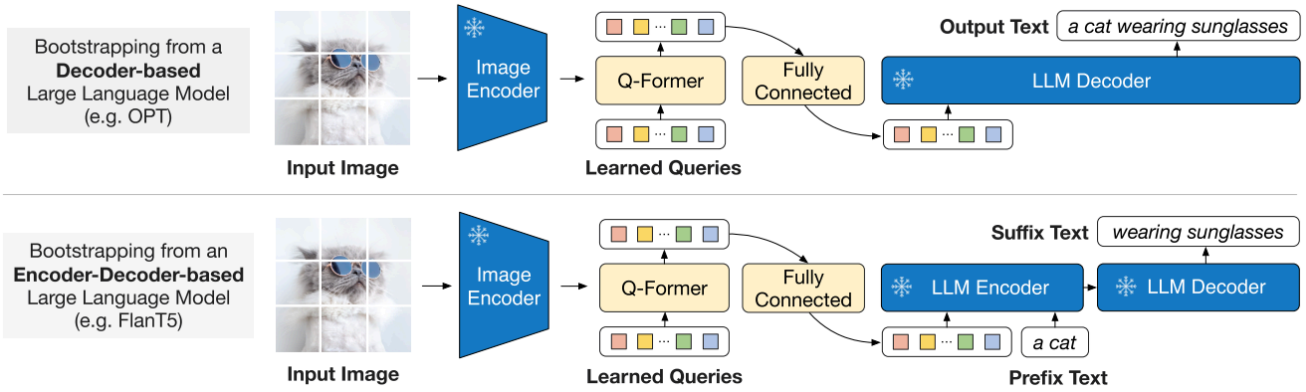
为了减少计算成本并避免灾难性遗忘的问题，BLIP-2 在预训练时冻结预训练图像模型和语言模型，但是，简单地冻结预训练模型参数会导致视觉特征和文本特征难以对齐，为此BLIP-2提出两阶段预训练 Q-Former 来弥补模态差距：表示学习阶段和生成学习阶段。

（1）表示学习阶段
在表示学习阶段，将Q-Former连接到冻结的Image Encoder，训练集为图像-文本对，通过联合优化三个预训练目标，在Query和Text之间分别采用不同的注意力掩码策略，从而控制Image Transformer和Text Transformer的交互方式。



（2）生成学习阶段

在生成学习阶段，将Q-Former连接到冻结的LLM，以利用LLM的语言生成能力。这里使用全连接层将输出的Query嵌入线性投影到与LLM的文本嵌入相同的维度，然后将投影的Query嵌入添加到输入文本嵌入前面。由于Q-Former已经过预训练，可以提取包含语言信息的视觉表示，因此它可以有效地充当信息瓶颈，将最有用的信息提供给LLM，同时删除不相关的视觉信息，减轻了LLM学习视觉语言对齐的负担。



Grounded Language-Image Pre-training

摘要

This paper presents a grounded language-image pre-training (GLIP) model for learning object-level, language-aware, and semantic-rich visual representations. GLIP unifies object detection and phrase grounding for pre-training. The unification brings two benefits: 1) it allows GLIP to learn from both detection and grounding data to improve both tasks and bootstrap a good grounding model; 2) GLIP can leverage massive image-text pairs by generating grounding boxes in a self-training fashion, making the learned representations semantic-rich. In our experiments, we pre-train GLIP on 27M grounding data, including 3M human-annotated and 24M web-crawled image-text pairs. The learned representations demonstrate strong zero-shot and few-shot transferability to various object-level recognition tasks. 1) When directly evaluated on COCO and LVIS (without seeing any images in COCO during pre-training), GLIP achieves 49.8 AP and 26.9 AP, respectively, surpassing many supervised baselines. 2) After fine-tuned on COCO, GLIP achieves 60.8 AP on val and 61.5 AP on test-dev, surpassing prior SoTA. 3) When transferred to 13 downstream object detection tasks, a 1-shot GLIP rivals with a fully-supervised Dynamic Head.

引用信息（BibTeX格式）

```
@inproceedings{li2022grounded,
  title={Grounded language-image pre-training},
  author={Li, Liunian Harold and Zhang, Pengchuan and Zhang, Haotian and Yang, Jianwei and Li, Chunyuan and Zhong, Yiwu and Wang, Lijuan and Yuan, Lu and Zhang, Lei and Hwang, Jenq-Neng and others},
  booktitle={Proceedings of the IEEE/CVF conference on computer vision and
```

```
pattern recognition},
  pages={10965--10975},
  year={2022}
}
```

本论文解决什么问题

视觉领域（如目标检测）与语言理解（如文本描述）之间长期缺乏统一的预训练框架。
传统目标检测模型依赖预定义类别集合，面对标注中没有的类别无能为力；而 CLIP 虽能理解开放词汇，但缺乏像素级或检测级定位能力。
GLIP旨在统一视觉检测和语言理解任务，提出一种“目标检测即文本匹配”的视角，实现开放词汇目标检测（open-vocabulary object detection）

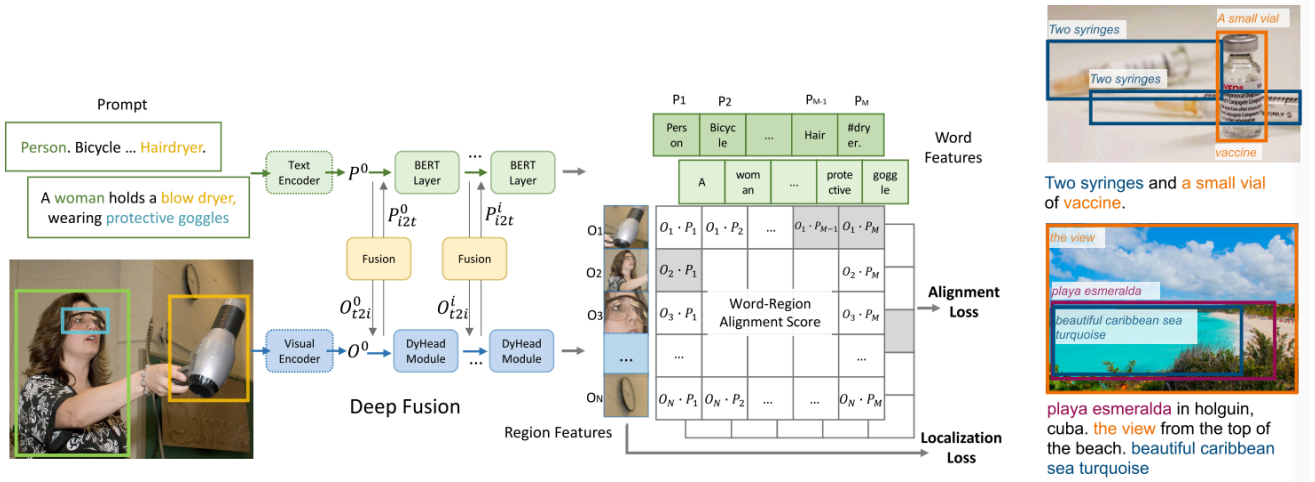
已有方法的优缺点

传统目标检测模型使用大量人工标注bounding box，不能处理标注中没有的类，也无法利用文本增强语义。CLIP文本-图像语义对齐良好，能泛化到开放词汇场景，但无法进行 bounding box 级别的定位，也无法替代目标检测器。且只能进行图像全局匹配，不具备区域级理解

本文采用什么方法及其优缺点

GLIP的核心理念是将目标检测视为：文本与图像区域的对齐任务

模型使用一个 DETR 风格框架，将每个预测框（region query）与输入文本进行对齐匹配，从而实现语言引导的检测。



在CLIP等算法中，image和text特征通常只在最后用于计算对比学习的loss。在本文中，作者在image和text特征之间引入了更深层次的融合（deep fusion），在最后几个encoder layer中进行了image和text的信息融合，具体来说，GLIP采用DyHead作为image encoder、BERT作为text encoder。DyHead本质上是对于尺度、空间、任务三个维度分别进行自注意力机制运算。deep-fused有两个优点：1）提升了phrase grounding的表现；2）使得图像特征的学习与文字特征产生关联，从而让ext prompt可以影响到检测模型的预测。

使用的数据集和性能度量

GLIP训练采用的数据包含了超过2000个类别，并且是bbox+phrase grounding的标注。GLIP分别在COCO、LVIS、Flickr30K数据集上进行了评测。

在COCO数据集上的表现如下，可以看到Zero-Shot的GLIP模型就已经超越了Faster RCNN的表现了，而在经过finetune之后，GLIP-L的mAP达到了略超过DyHead的水平。

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 [9]	Swin-T	-	-	49.7 / -
DyHead-L [9]	Swin-L	-	-	58.4 / 58.7
DyHead-L [9]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [58]	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**

在LVIS数据集上的表现：GLIP Zero-shot的表现超过了supervisied训练的 MDETR模型。

Model	Backbone	MiniVal [19]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [19]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [19]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [13]	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**

在Flickr30K数据集上的表现也达到了SoTA水平：

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**

与我们工作的相关性

1、GLIP 能把文本语义对齐到图像的具体区域，输出bounding box。
2、GLIP 的Deep Fusion通过将文本特征逐层注入视觉特征，有一定的参考价值。
3、BLIP2的对齐方式是在文本语义和图像区域之间建立“语义关联”，而非将图像特征和文本特征对齐到一个向量空间。它可以增强文本描述的理解与生成，但不能直接用于 SAM 的定位输入。