Position-guided Text Prompt for Vision-Language Pre-training

💡 Meta Data

Title	Position-guided Text Prompt for Vision-Language Pre-training
Journal	
Authors	Alex Jinpeng Wang; Pan Zhou; Mike Zheng Shou; Shuicheng Yan
Pub. date	2023-06-07
期刊标签	
DOI	10.48550/arXiv.2212.09737
附件	Wang et al_2023_Position-guided Text Prompt for Vision-Language Pre-training.pdf

📜 研究背景 & 基础 & 目的

背景：

缺乏视觉的定位的模块、提升使用VLP训练的跨模态模型的视觉基础能力

提出了位置引导文本提示PTP增强跨模态模型的视觉ground。

PTP放弃了目标检测器直接进行推理

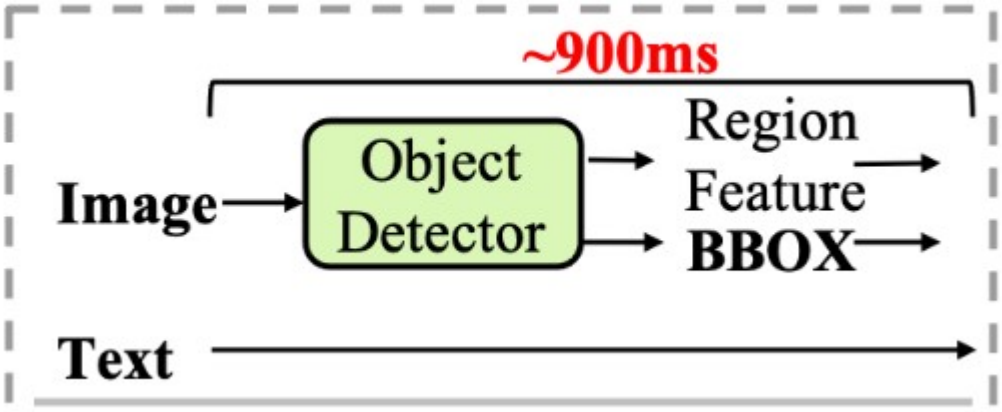
基础：

通用的跨模态模型首先以自监督的方式对大规模图像描述数据进行预训练，以获得足够的数据以获得更好的泛化能力。

然后对下游任务进行微调以进行适应。

“region features” (Wang 等, 2023, p. 1)

为了对位置信息进行建模，传统的 VLP 模型 采用在 1600 个类视觉基因组上预训练的 fast-rcnn 来提取显着特征区域特征和边界框。

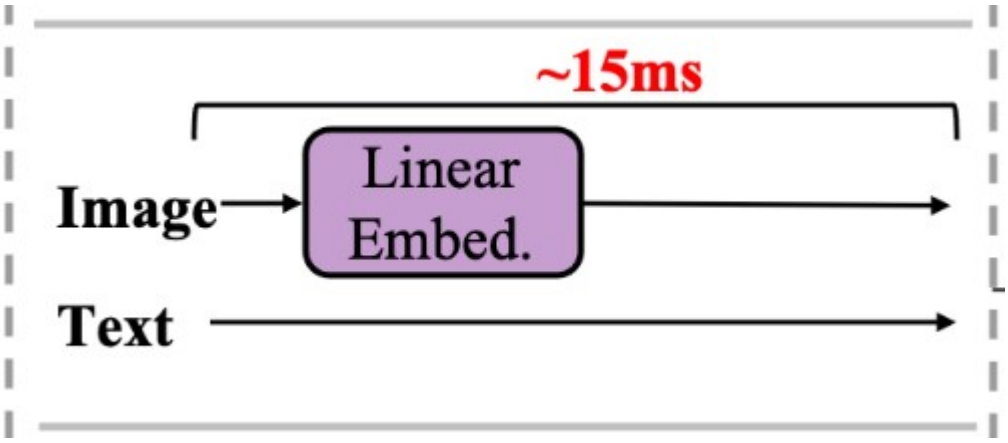


> “To model the position information, traditional VLP models [3,23,46] (the top of Fig. 1 (a)) employ a faster-rcnn [34] pre-trained on the 1600 classes Visual Genome [17] to extract salient region features and bounding boxes.” (Wang 等, 2023, p. 1)

是什么+在哪里

没有上下文信息推理速度较慢

“end-to-end” (Wang 等, 2023, p. 2)



速度很快但无法很好的学习目标及其关系

目的

缓解端到端模型的位置缺失问题，并同时保持下游任务的快速推理时间

> “In this work, we aims to ease the position missing problem for these end-to-end models, and keep fast inference time for downstream tasks at the same time.” (Wang 等, 2023, p. 2)

在图像和文本中添加基于位置的共同标识，将视觉grounding重新表达为一个填空问题。从而简化对目标信息的学习。

在图像数据中实现自然语言的表达

生成标签块

将图像划分为N×N的块，并识别每个block中的目标。

生成文本提示

将查询文本放入基于位置的文本查询模板中。

> “To ground natural language expressions in image data, PTP contains two components: (1) block tag generation to divide image into N × N blocks and to identify object in each block, and (2) text prompt generation that puts the query text into a position-based text query template.” (Wang 等, 2023, p. 2)

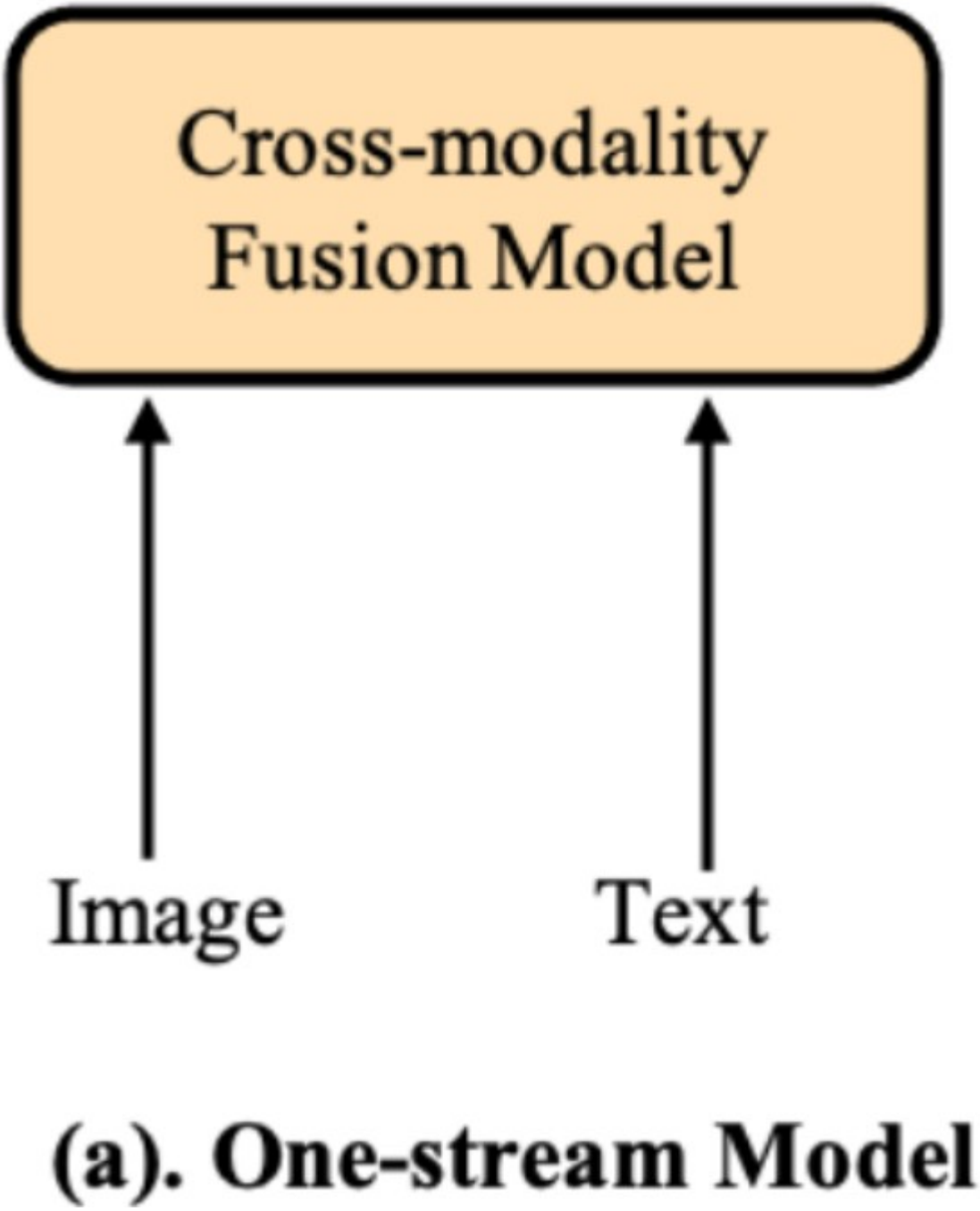
📊 研究内容

Vision-language Pre-training Models

三个模型

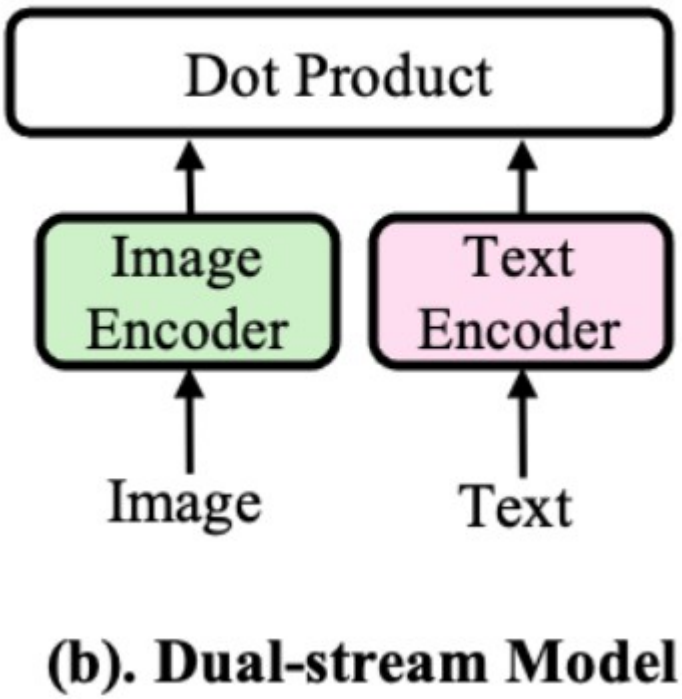
单目标模型、多目标模型、多目标+融合编码模型（自己理解的翻译)。

“One-stream Model” (Wang 等, 2023, p. 2)



对图像和文本输入进行处理。

“Dual-stream Model” (Wang 等, 2023, p. 2)

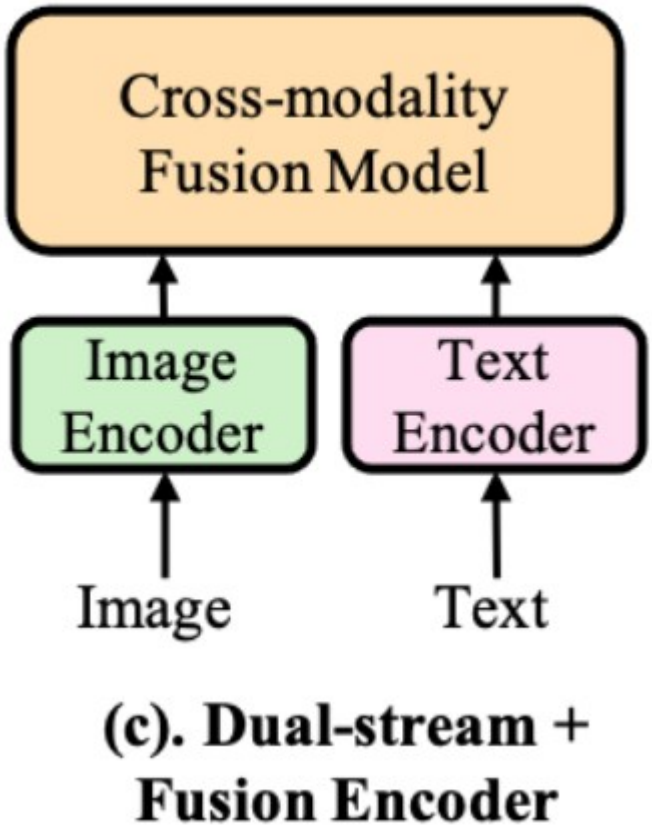


(b). Dual-stream Model

每种模态使用独立但同样expensive transformer编码器。

这两种模式在输入层没有串联，而在浅层，汇集的图像矢量和文本矢量之间存在交互。

“Dual-stream with Fusion Model” (Wang 等, 2023, p. 2)



(c). Dual-stream + Fusion Encoder

以上两种模式的融合

“Prompt Learning for Computer Vision” (Wang 等, 2023, p. 2)

多模态提示，它为 VLPT 模型提供了多模态提示调整，在一些视觉语言任务中取得了可喜的成果。

“Learn Position Information in VLP” (Wang 等, 2023, p. 2)

groundingly对于多模态任务至关重要，为了将此能力引入VLP将区域特征和边界框向量进行连接。但由于在下游任务中目标提取非常耗时。目前的研究为解决这个问题都是根据特定框架设计的，难以扩展。

本文目标提出了一个学习位置信息的通用框架，可以插入现有框架。

“Position-guided Text Prompt” (Wang 等, 2023, p. 3)

以VILT、CLIP、BLIP为例介绍如何将PTP和VLP框架结合来增强其视觉基础能力。

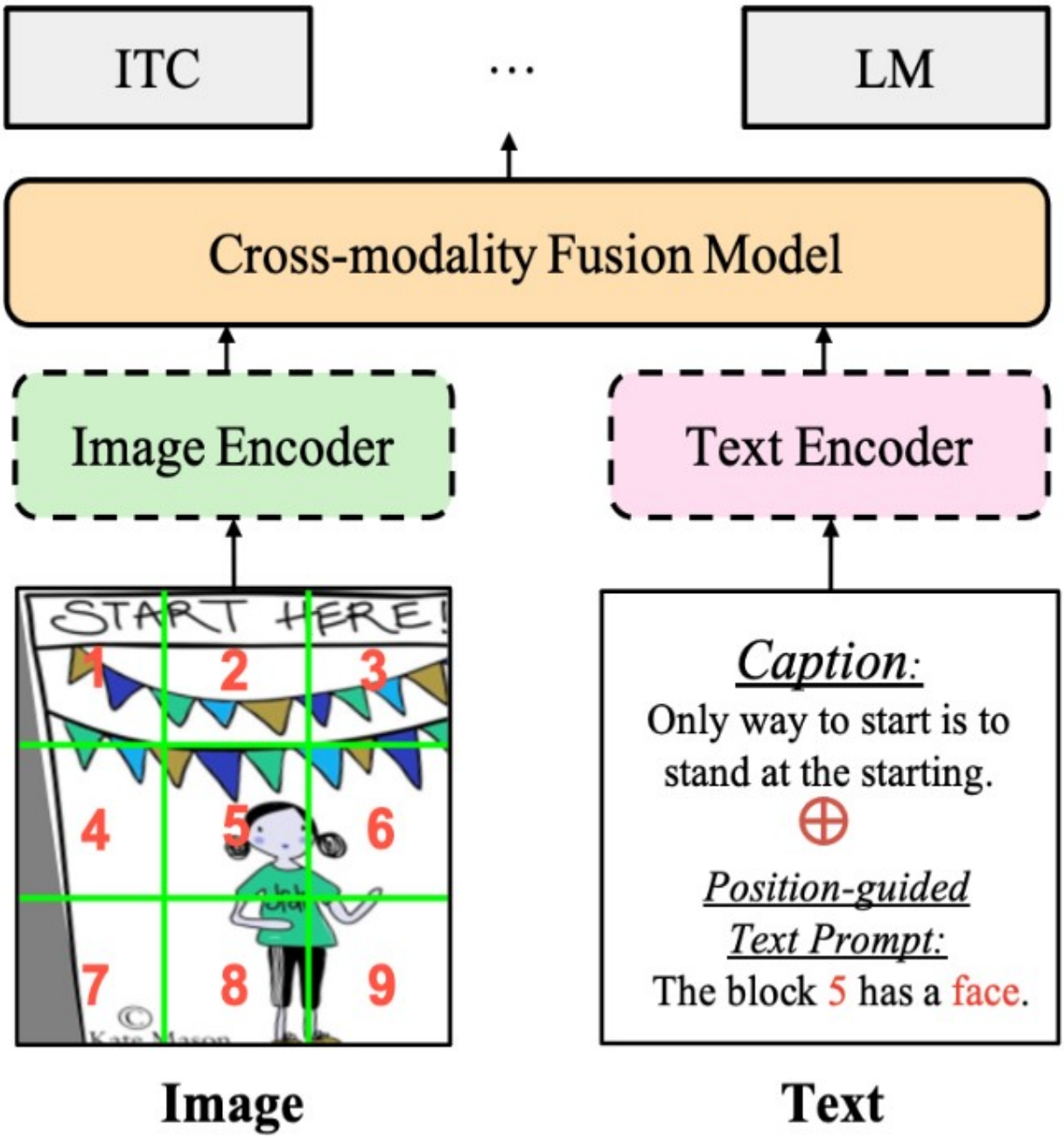
“PTP Paradigm” (Wang 等, 2023, p. 3)

PTP帮助跨模态感知物体，并将物体与相关文本对其。

PTP不同于传统的视觉语言配准方法，PTP将物体特征和边界框作为输入

1.生成区块标签：将输入的图像分成若干块，并识别每个区块中的物体。

2.生成文本提示：根据1中的物体位置信息，将视觉ground任务转换为填空问题。



“Block Tag Generation” (Wang 等, 2023, p. 3)

对每一对图像文本平均分成N×N个区块

1.目标检测：

采用VinVL中的Faster-rcnn提取每幅图像中的所有目标。

选出预测置信度最高的前k个目标，用$\mathcal{O}=\left\{o_i\right\}_{i=1}^K$表示 其中$o_i=(z_i,q_i)$ 表示具有4维区域的位置向量z和目标类别q

2.CLIP模型：

采用区域监测而非haevy目标检测，PTP可以通过CLIP(ViT-B）模型生成分块的目标监督

1.提取短语：

提取整个文本语料库中频率出现最高的M个短语或关键词（默认为3000个)使用NLTK 被视为词汇记为V。

提取文本特征通过CLIP编码器嵌入所有M个短语或关键词。

获取每个区块中的图像嵌入h并计算文本特征的相似度。

相似度得分最高的关键字和短语为特定块的最终目标标签。

以下是每个块的目标标签索引计算公式

$$I = \text{argmax}_{y\in[1,\ldots,M]} \left(\frac{\exp(h^T e_y)}{\sum_{w\in V} \exp(h^T e_w)} \right), \quad (1)$$

argmax（）返回最大值坐标

exp（）以e为底的指数函数

h 是所选块的视觉特征嵌入

3.优点

与目标检测器相比

产生更多样化的目标标签

块标签的生成速度快比Faster-RCNN块40倍

“Text Prompt Generation” (Wang 等, 2023, p. 3)文本提示生成

设置一个简单的文本提示

P是所选块的索引表示块的位置O表示块P的标签。每一个P可能会包含多个O所以每次随机选择一个O

“In this way, each sentence in our PTP incorporates fine-grained object position and language into a model, and thus provides a new way to align the objects and pertinent text.” (Wang 等, 2023, p. 4)

通过这种方式，我们的 PTP 中的每个句子都将细粒度的对象位置和语言合并到模型中，从而提供了一种对齐对象和相关文本的新方法。

“Pre-training with PTP” (Wang 等, 2023, p. 4)使用 PTP 进行预训练

在这项工作中，我们将我们的 PTP 集成到主流 VLP 框架中，产生了 PTP-ViLT 、PTP-CLIP 和 PTP-BLIP。

1.集成到现有任务中:

提示文本和原始标题被简单的填充到一起。$x=[w,q]$其中w是文本，q是本文方法生成的文本提示。

使用传统目标检测的端到端的训练方式训练VLP模型。

PTP-BLIP中采用LM loss、ITM、ITC loss

> TC（图像文本对比学习），ITM（图像文本匹配），MLM（Maked Language Modeling，有时会扩展到MIM），LM（Language Modeling，大部分可以看作是captioning）

PTP-ViLT中采用ITM、MLM loss

PTP-CLIP采用ITC loss

2.新的pretext task：

$$\mathcal{L}_{\mathrm{PTP}}(\theta) = -\mathbb{E}_{\mathbf{y}\sim D}\left[\sum_{t=1}^{T} \log P_\theta\left(\mathbf{y}_t \mid \mathbf{y}_{<t}\right)\right]$$

上式是目标预测的损失函数

D为预测数据

$y_1,...,y_{t-1}$是文本提示q的训练标记序列

t为时间步

$p(t)=p(*|y_1,...,y_{t-1})$是需要通过设计来进行预测的概率分布

θ 是模型的可训练参数。

PTP不需要修改任何基础网络，并且可以应用于任何没有附加功能的VLP模型。PTP旨在从原始像素图像中学习位置信息。只有在预训练阶段才需物体位置信息，在下游的任务中采用正常的端到端的方式评估模型，无需目标的信息，从而摆脱繁重的目标特征提取以提高效率。

“Experiments” (Wang 等, 2023, p. 4)

“Experimental Settings” (Wang 等, 2023, p. 4)

1.“Datasets” (Wang 等, 2023, p. 4)

使用COCO、VG、SBU、CC3M组成4M设置。

Datasets. As in earlier studies [23, 46], we begin by using a 4M setup made up of four popular pre-training datasets (COCO [24], VG [17], SBU [29] and CC3M [35]). Following recent work [19], we also explore 14M setting, which includes additional CC12M [6] (actually only 10M image urls available) dataset besides 4M datasets. We refer readers to supplementary material for more dataset details.

2.“Training Settings” (Wang 等, 2023, p. 4)

PyTorch+8*NVIDIA A100 GPU

图像增强，探索 RandAugment并使用除颜色反转之外的所有原始策略，因为颜色信息很重要。

> RandAugment是**通过网格搜索两个参数(N，M)来寻找最佳DA的方法**，其中N为DA的数量(变换)，M为Augmentation的程度。 这是为图像识别而设计的，所以在视频的情况下，必须对每一帧进行应用。

以图像相同的方式增强边界框以进行仿射变换。

在预训练期间随机采集分辨率为 224 × 224 的图像，并将图像分辨率提高到 384 × 384 进行微调。

3.“Baselines” (Wang 等, 2023, p. 4)

ViTB/16 作为基础视觉编码器并使用相同的数据集。

“Main Results” (Wang 等, 2023, p. 4)

PTP在每个视觉下游任务的部署。

“Image-Text Retrieval” (Wang 等, 2023, p. 4)图文检索

本文在 COCO 和 Flickr30K 基准上评估图像到文本检索 (TR) 和文本到图像检索 (IR) 的 PTP。

> ViLT 是**视觉和语言模型中最简单的架构**，因为它使用Transformer 模块代替单独的深度视觉嵌入器来提取和处理视觉特征。 这种设计本质上显著改善了运行时间和参数效率。 这是我们首次在不使用区域特征或深度卷积视觉嵌入器的情况下，在视觉和语言任务上取得了胜任的性能。
>
> BLIP 是**一种多模态Transformer 模型**，主要针对以往的视觉语言训练(Vision-Language Pre-training, VLP) 框架的两个常见问题： 大多数现有的预训练模型仅在基于理解的任务或者基于生成的任务方面表现出色，很少有可以兼顾的模型。

在MSCOCO上 对于Image->Text Recall@1提高了13.8%

本文只在预训练阶段使用目标检测器。这表明目标检测器并不是成功的秘诀，如何利用位置信息对于 VLP 模型至关重要。

“Image Captioning” (Wang 等, 2023, p. 5)图像字幕

数据集：No-Caps 、COCO 。

均使用在 COCO 上微调并具有 LM 损失的模型进行评估。

> **CIDEr 是 BLEU 和向量空间模型的结合。**它把每个句子看成文档，然后计算 TF-IDF 向量（只不过 term 是 n-gram 而不是单词）的余弦夹角，据此得到候选句子和参考句子的相似度，同样是不同长度的 n-gram 相似度取平均得到最终结果。优点是不同的 n-gram 随着 TF-IDF 的不同而有不同的权重，因为整个语料里更常见的 n-gram 包含了更小的信息量。

“Visual Question Answering” (Wang 等, 2023, p. 5)视觉问答

与 ViLT 相比，PTP提高了1.8%

“Visual Reasoning” (Wang 等, 2023, p. 6)视觉推理

自然语言视觉推理（NLVR2）任务是一个二元分类任务，给定两个图像的三元组和一个自然语言问题。虽然不如SimVLM性能但是此方法也接近 VinVLlarge 模型，采用更大的模型并使用来自强对象检测器的对象特征而不是原始像素图像作为输入。

“Video-Language Tasks” (Wang 等, 2023, p. 6)视频语言任务

在这个实验中分析了我们的方法对视频语言任务的泛化能力。

“Ablation & Design Choices” (Wang 等, 2023, p. 6)

“The Variations of Architecture.” (Wang 等, 2023, p. 6)结构变体

使用三种不同类型的baseline进行实验：ViLT、CLIP 和 BLIP，以探索 PTP 的影响。比较这些baseline实验的结果，我们发现 PTP 极大地提高了 i2t 和 t2i 性能。这表明PTP具有良好的通用性。由于本文在下游任务中不使用目标检测器或提示，因此计算成本与基线模型保持一致，但比基于目标特征的 VinVL 快 20 倍。

“Text Prompt vs. Additional Pretext Task” (Wang 等, 2023, p. 7)文本提示与附加借口任务

提示比借口要好得多，特别是对于 COCO 字幕 CIDER（127.2 vs 123.5）。在这项工作中，由于其效率，本文使用提示作为默认值。

“Other Types of Text Prompt” (Wang 等, 2023, p. 7)

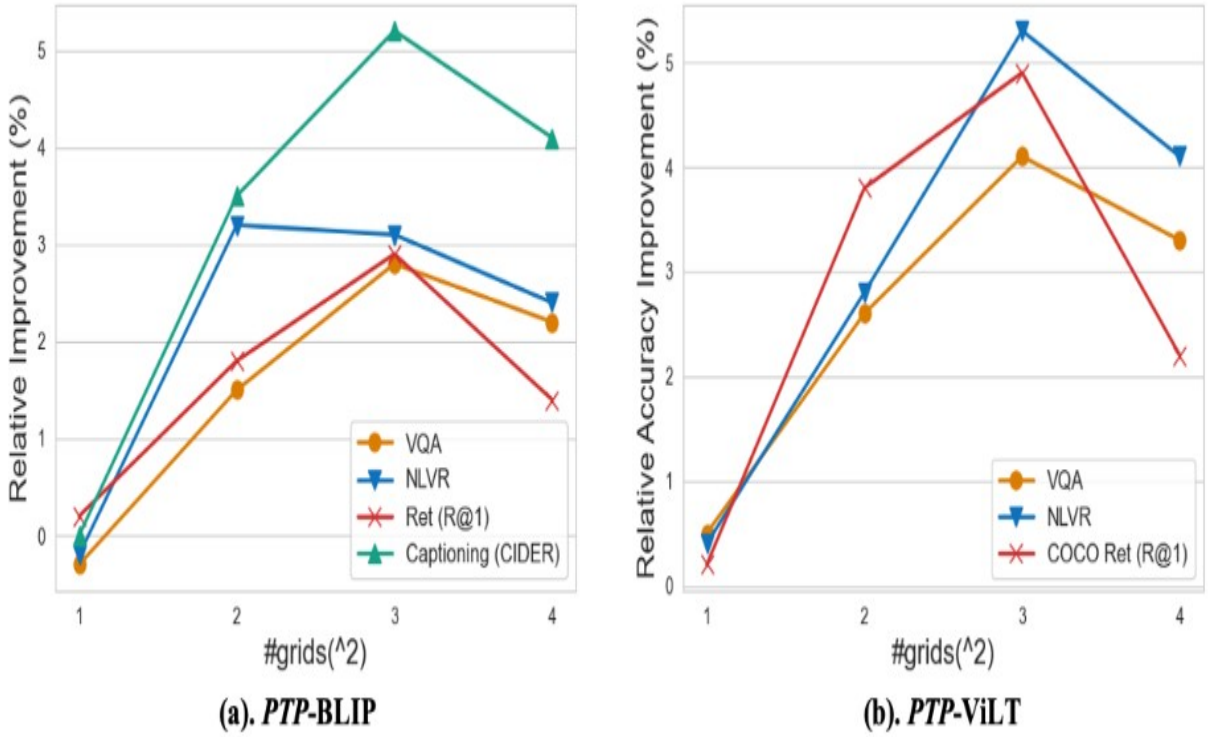
“The Importance of Position in Text Prompt” (Wang 等, 2023, p. 7)

在这个实验中检查了提示 PTP 获取各种粒度信息（例如没有位置信息）的效果。

每一个信息都很重要，如果缺失，下游的性能会变的越来越差。

“Number of Blocks” (Wang 等, 2023, p. 7)

建议使用 3 × 3 块，因为它具有准确性。



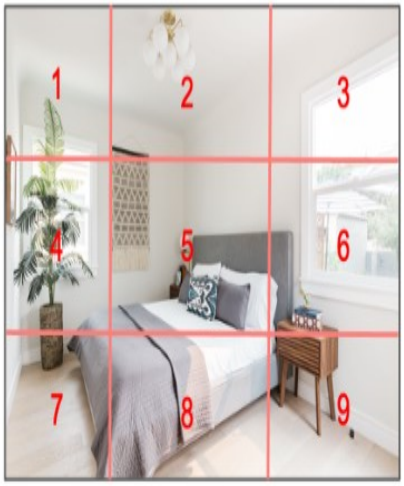
“Is Object Detector Necessary?” (Wang 等, 2023, p. 8)

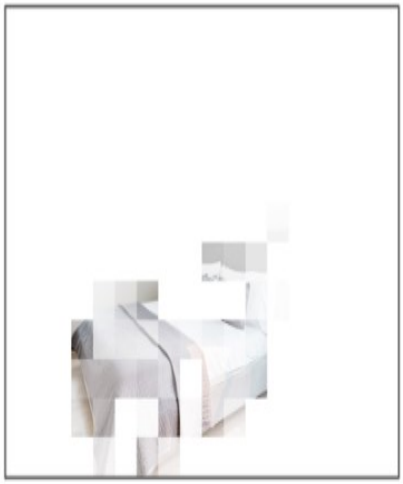
除了基于 ResNext152 的强大目标检测器之外，我们还使用了一个较小的 Faster-rcnn 网络，该网络利用 ResNet101 作为主干。

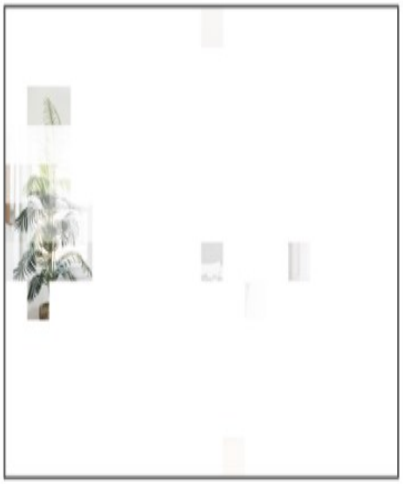
使用更强的检测器会带来更好的结果，但同时也会带来巨大的计算成本。

结论:剪辑模型是 PTP 中目标检测器的一个很好的替代方案。

“Visualization” (Wang 等, 2023, p. 8)







Input

There is a [MASK] in block 8.

There is a plant in block [MASK].

Prediction

Top3: *Bed* (0.84), *Floor* (0.05), *Sheet* (0.02)

Top3: *4* (0.81), *1* (0.09), *5* (0.07)

Input

There is a [MASK] in block 8.

There is a man in block [MASK].

Prediction

Top3: *Dog* (0.87), *Labrador* (0.05), *Pet* (0.02)

Top3: *3* (0.63), *5* (0.19), *4* (0.08)

研究结论

感想 & 疑问

🔬 理论推导