

基于位置引导的文本提示在视觉语言预训练中的应用

摘要

图像字幕生成 (Image Captioning) 是一项多模态任务, 旨在为图像生成语义丰富且自然的语言描述。随着深度学习技术的进步, 基于注意力机制的模型在这一领域取得了显著成果。本文探讨了注意力机制在图像字幕生成中的关键作用, 特别是其在视觉和语言信息融合中的表现。本文了解了 Transformer 模型和 Vision Transformer (ViT) 在计算机视觉领域的应用, 并复现了相关研究工作, PTP 架构, 并利用 MIA 模型进行实验以验证 PTP 架构的作用。此外, 通过对比实验分析, 本文进一步探究了加入位置信息后的注意力模型如何优化字幕生成的性能。

关键词: 图像字幕生成; PTP; 计算机视觉

1 引言

随着深度学习和人工智能技术的迅猛发展, 多模态任务逐渐成为计算机视觉与自然语言处理领域的重要研究方向。作为其中的一个经典且有效的任务, Image Captioning (图像字幕生成) 不仅要求模型从视觉信息中提取特征, 还需要生成与之匹配的自然语言描述, 涉及到图像理解与语言生成的多层次交互。这一任务的研究有助于推动计算机视觉技术在自动化标注、智能搜索等领域的应用, 为人类与计算机之间的互动方式带来新的可能性。

注意力机制作为深度学习的重要工具, 能够动态关注输入数据中的关键特征, 极大提升了模型的表达能力。Vaswani 等人提出的完全基于注意力机制的 Transformer 模型架构, 为自然语言处理任务提供了高效的特征建模方法 [1]。在计算机视觉领域, Vision Transformer (ViT) 模型首次将 Transformer 架构成功应用于视觉任务 [2]。2023CVPR 发表的 PTP (Pre-training and Transfer Learning) 框架因其出色的表现被广泛关注 [3]。PTP 框架通过预训练和迁移学习的结合, 成功地将多模态信息融合, 为多任务学习提供了强大的支持。在诸多应用场景中, PTP 框架展现了优异的图像理解和语言生成能力, 尤其在 image captioning 任务中, 其结构设计和技术实现让该框架在处理多模态数据时具备了显著优势。

在本次研究中, 复现并应用 PTP 框架进行 image captioning 任务, 结合 NeurIPS2019 的工作 MIA 模型 [4], 本文探讨了注意力机制和位置特征在 image captioning 种的影响。这不仅帮助我深入理解了该框架的技术细节与应用效果, 更加深了我对多模态视觉分析的理解。通过实践这一任务, 我得以掌握相关的多模态学习方法, 进一步提升了自己在计算机视觉与自然语言处理交叉领域的技能, 为未来相关领域的研究奠定了坚实的基础。

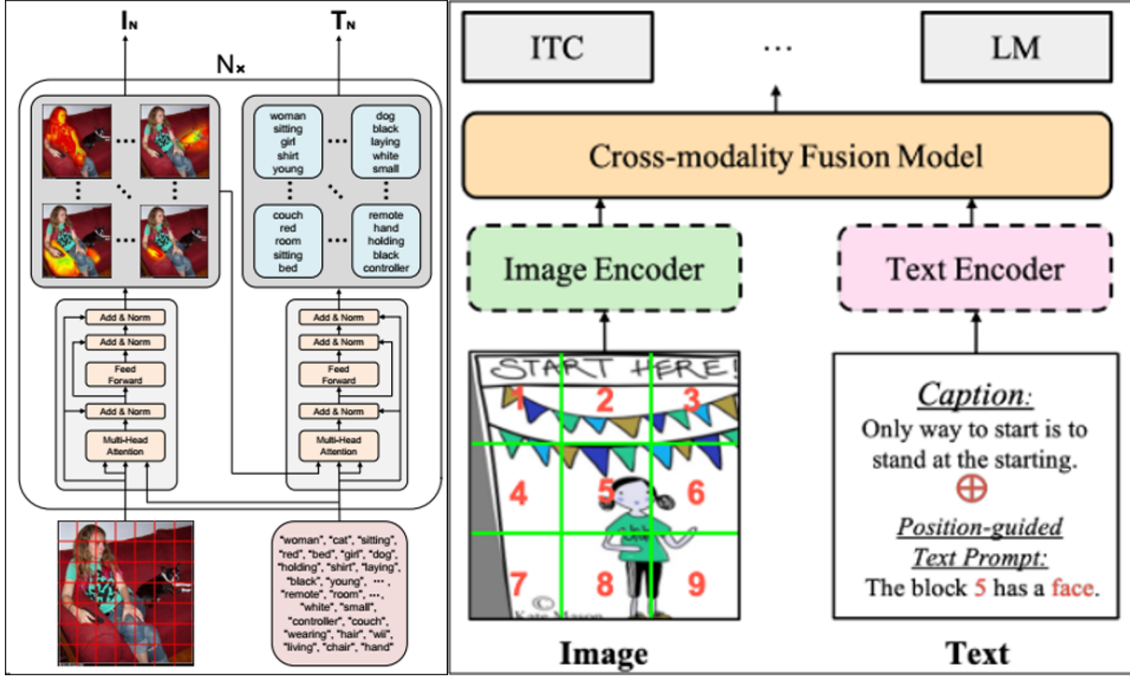


图 1. 左：MIA 模型架构，输入 patch 级别的图像和词向量级别的特征，输出新的图像和文本特征；右：PTP 结构示例，PTP 不改变训练模型的结构，可以被视为数据预处理方式。

2 相关工作

2.1 MIA 模型

NeurIPS 2019 的论文《Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Representations》探索了视觉区域和文本概念之间的语义对齐问题，通过将注意力机制与图像区域特征相结合，提出了相互迭代注意模型（MIA），显著提升了视觉-语言表示学习的能力。这种方法有效捕获了图像区域与生成文本的对应关系，提高了生成描述的准确性和语义一致性。

具体而言，如图 1 左所示，MIA 框架的特点有两个。1. 相互：模型通过让文本特征和图像特征相互询问的方式让两种模态之间对齐，即：

$$T = F(Q_I, K_T, V_T)$$

$$I = F(Q_T, K_I, V_I)$$

其中 T 和 I 分别表示文本特征和图像特征。2. 迭代：取上一轮模块的输出作为下一轮模块的输入，即：

$$I' = FCN(MultiHeadAtt(T, I)), T' = FCN(MultiHeadAtt(I', T))$$

2.2 PTP 架构

此外，Zhou 提出了通过位置引导的文本提示学习，并给出了具体的位置引导文本提示（PTP）框架，进一步强化了视觉和语言之间的联动，使得预训练模型能够更好地适应多模态

Path	Caption	BBox	Class
coco2014/COCO2014/train2014/COCO_train2014_000000019227.jpg	two black bears laying on the ground in the grass	[[56, 166, 450, 332], [302, 175, 585, 315]]	['animal bear', 'animal bear']

图 2. PTP 作者在 Github 中给出的语料库的储存格式样例

任务。

如图 2 所示，PTP 结构步骤如下：1. 识别图像中的主要物体：通过使用预训练的目标检查模型（如 VinVL [5] 模型中的 Fast-Rcnn）或使用 CLIP [6] 模型识别图像中的 k 个主体对象；2. 将图像均分为多个块，如 3×3 分块；3. 按照主体对象的中心（Bbox 的中心）与块中心的距离最小将主体划归到对应块；4. 按照以下格式生成提示（Prompt）：

$$”The block [P] has a [O]”$$

例如图中的提示应当为 “The block 8 has a girl.”；5. 将新生成的提示 p 拼接到原文本 w 后即

$$w' = [w, q]$$

从上述的描述我们可以观察到，PTP 是一种数据预处理方式，是一个不改变原本模型架构的，即插即用的模块。

3 方法

3.1 数据与模型基础设置

本文数据采用了被广泛应用于图像字幕生成任务的 CoCo2014 数据（大约有 330K 图像和 680K 文本）。图像编码器采用 ResNet-152 [7]，没有采用文本编码器，直接使用词字典映射为特征。此处的文本编码并没有引入位置特征，主要用于实验验证 PTP 的位置信息作用。解码器采用了常用的 LSTM [8] 模型。

对于使用 MIA 模型的部分，设置迭代次数为 2（论文中推荐）。对于使用 PTP 模块的部分实验中体现为多使用一个 Caption 数据集（文本）。

3.2 本文实验设置

本文设置了两组实验。第一组为简单的复现实验，使用 PTP 论文提供的开源预训练模型 PTP-BLIP 进行了 zero-shot 检索任务，用以检验模型复现成功；然后又设置了一组较为丰富的对照使用，取出 PTP 架构套用到 MIA 模型上，用以检验 PTP 架构的效果。

3.2.1 PTP 架构对比实验

本文训练了以下多组模型：

- 基线 (B)：直接使用 CoCo 数据集，将经过 ResNet-152 编译得到的图像特征和词向量作为解码器 LSTM 的输入；

- 使用 MIA (M): 直接使用 CoCo 数据集, 图像特征和词向量经过两轮迭代的相互注意后再作为解码器的输入;
- 使用 PTP (P): 在基线模型的基础上, 使用 PTP 提供的语料库 + CoCo 数据集进行模型训练;
- 使用 MIA + PTP (MP): 使用 PTP 提供的语料库的同时, 使用 MIA 模块进行训练。

4 复现细节

4.1 与已有开源代码对比

本文并没有直接使用 PTP 开源提供的 PTP-BLIP 模型进行对照实验, 仅进行了其中的 zero-shot 检索任务, 用以检验复现工作的完成。然后又取出 PTP 架构的核心部分 (一个即插即用的数据预处理方式), 嫁接到 MIA 模型上, 并设置多组对照实验, 用以检验 PTP 的效果, 探究注意力机制和位置特征在图像字幕生成工作的作用。

主要工作有: 1. 复现 MIA 和 PTP (PTP-BLLIP) 模型; 2. 复现 PTP 数据预处理模块; 3. 修改 MIA 代码嫁接上 PTP 模块。

4.2 实验环境搭建

实验环境搭建详细请查阅提交的 “README.md 文件”。

4.2.1 基线模型与 MIA

MIA 模型的代码可见于 [Github](#), 按链接内容完成复现即可。基线模型在复现完成 MIA 时被同时完成, 因为 MIA 提供了 `use_MIA` 参数用于指定是否使用 MIA 模型 (即该参数为 `False` 时, 使用基线模型)。

4.2.2 PTP 架构的使用

PTP 可以被视为一种数据预处理方式, 因此应用 PTP 模块仅需要生成一个与数据对应的语料库, 并在装载数据时进行处理即可。PTP 作者在 [Github](#) 上提供了 “CoCo、VG、SBU 和 CC3M” 的语料库 (样例如图 2 所示), 下载并进行处理即可。本文处理步骤如下:

- 识别 Path 的第一节 (按 ‘/’ 分割), 保留 `coco2014`;
- 调用 PTP 作者给出的函数, 生成附带位置信息的提示, 按以下格式保留为 Json 文件

$\{ImageID, Caption\}$

- 在 MIA 项目中添加读取新制作的 Json 文件的代码, 并进行文本拼接。

5 实验结果分析

在这一节, 本文将通过多组对照实验, 验证注意力机制为图像字幕生成任务带来的影响。

```

{
  "test_txt_r1": 72.28,
  "test_txt_r5": 91.8,
  "test_txt_r10": 95.68,
  "test_txt_r_mean": 86.58666666666666,
  "test_img_r1": 49.364254298280684,
  "test_img_r5": 75.76569372251099,
  "test_img_r10": 84.34626149540183,
  "test_img_r_mean": 69.8254031720645,
  "test_r_mean": 78.20603491936558
}

```

Method	#Images	Parameters	MSCOCO (5K test set)							
			Image → Text			Text → Image			Avg	
			R@1	R@5	R@10	R@1	R@5	R@10		
Unicoder-VL [18]	4M	170M	—	—	—	—	—	—	—	—
ImageBERT [31]	4M	170M	44.0	71.2	80.4	32.3	59.0	70.2	59.5	
ViLT [16]	4M	87M	41.3	79.9	87.9	37.3	67.4	79.0	65.5	
PTP-ViLT (ours)	4M	87M	55.1	82.3	89.1	43.5	70.2	81.2	70.2 _{+4.7}	
BLIP [†] [19]	4M	220M	57.4	81.1	88.7	41.4	66.0	75.3	68.3	
PTP-BLIP (ours)	4M	220M	69.7	90.0	95.7	49.5	75.9	84.2	77.3 _{+9.0}	
PTP-BLIP (ours)	14M	220M	71.4	91.3	95.5	51.2	77.4	87.1	78.6	

图 3. 复现 PTP-BLIP 预训练模型并进行检索工作，结果与原论文相似

5.1 PTP-BLIP 的 zero-shot 检索任务复现结果

如图 3 所示，复现结果（左）在 TR@1 和 TR@5 明显高于论文的效果，IR@10 略高于论文效果，TR@10、IR@1 和 IR@5 略低于论文效果。总体均分略高。复现结果正常。

5.2 基线模型与 MIA 模型：在字幕生成任务中使用注意力机制需要引入位置信息

如图 4 所示，基线模型训练正常而引入 MIA 模块后训练崩溃，考虑到 MIA 模块基于注意力机制进行运作，而本文的设置（包括源代码的设置）中并没引入位置编码，这可能导致了 MIA 模块中的注意力机制仅仅学习了文本与图像的相关性，而失去了文本自身的连贯属性。即注意力机制需要提供位置信息，才能保证较好的完成字幕生成任务。图 5 中的实验例子可以作为该观点的进一步佐证。可以观察到，MIA 模型在训练 4 个 Epoch 时生成的文本具有一定的准确性；而当训练 30 个 Epoch 时，生成的文本基本是多个性性质基本一致的词的组合（颜色形容词），表明了缺乏位置信息时，使用注意力机制进行字幕生成任务会使得模型缺乏序列特征。

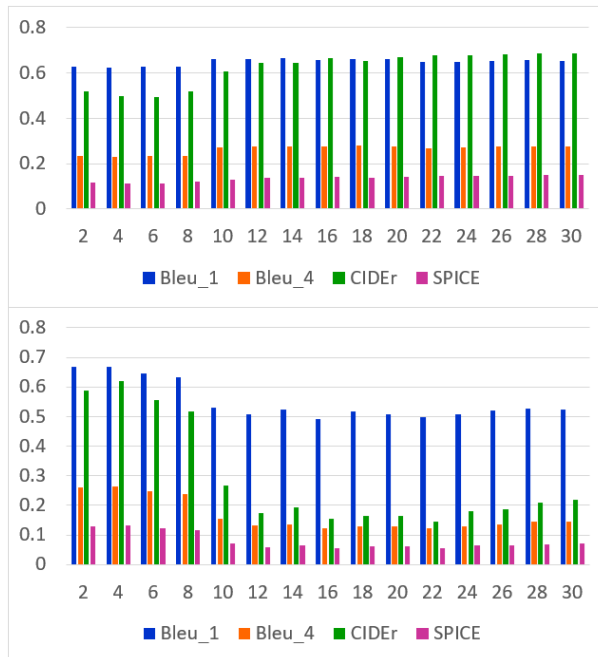


图 4. 上：基线模型训练结果展示，可以观察到各项指标均呈现上涨趋势，并在 Epoch=12 时逐渐区域平缓；下：MIA 模型训练结果展示：可以观察到 4 项指标均在 Epoch=4 时达到最高，后续指标迅速下跌，这表明添加 MIA 模块后模型训练失效了



图 5. MIA 生成的文本展示，在少次训练时，模型表现正常，但是之后逐渐崩溃

5.3 引入 PTP: PTP 模块可以提供位置信息，并使基于注意力机制 MIA 模块正常运作

如图 6 所示，MIA + PTP (MP) 组明显优势于 B 组和 M 组，而单独的 MIA 则表现出极为明显的下跌。即 PTP 架构带来的位置信息被 MIA 模块所利用，并重新表现出正常的训练情况。

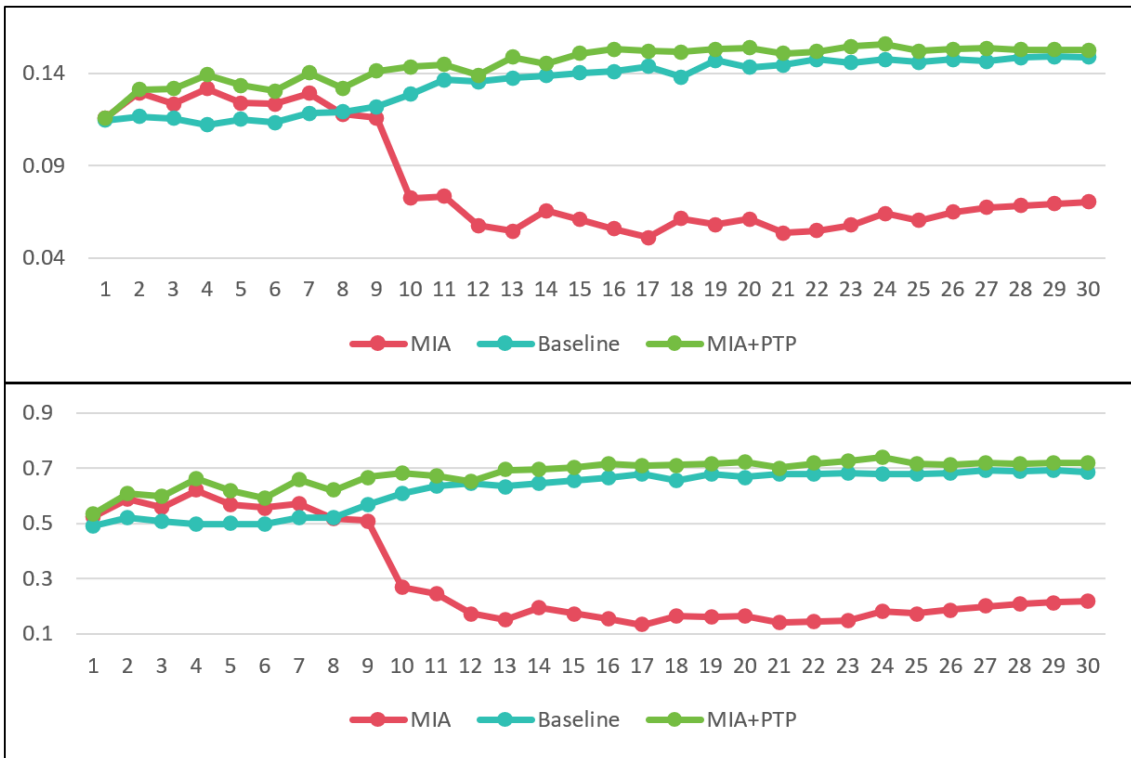


图 6. 上：纵轴指标为 SPICE，该指标强调生成文本与 GroundTrue 的语义相似性；下：纵轴指标为 CIDEr，该指标强调了生成文本与 GroundTrue 的内容相似性。

5.4 验证 MIA 模块的贡献

为了证明在 MP 组中 MIA 发挥了作用而不是仅仅由于 PTP 的能力过于强大才展现出上升趋势，进行了对照实验仅使用 PTP (P 组)，设置 Epoch 为 10 (设置较小的 Epoch 是因为此实验组仅作为验证即可)。如图 7 所示，P 组略微高于 B 组，而 MP 组明显优于两者，这证

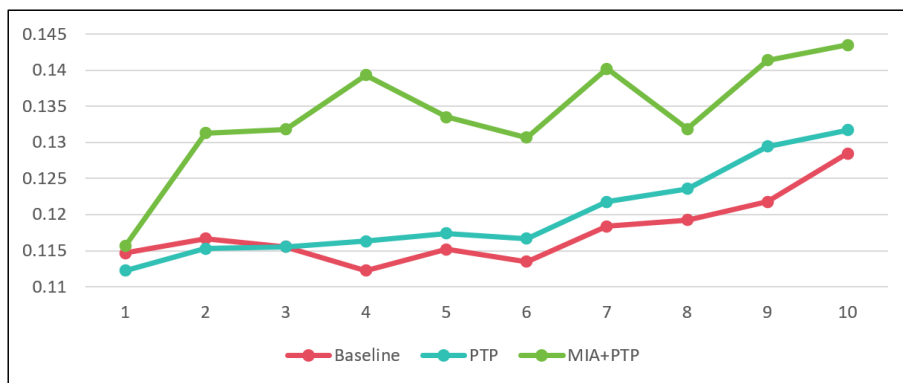


图 7. 从图中可以显著观察到 MP 组效果优于其他组别

明了在 MP 组实验 MIA 确实发挥了作用。

5.5 各实验组结果展示

图 8 为本次实验中多个模型在 coco 数据集中某张图像的结果展示，可以明显的观察到 MP 组生成的文本在可读性是最佳的，也不含有 `<unk>` 标签。值得注意的事，在样例中 P 组仅训练了 10 个 Epoch 而其余组别都训了 30 个 Epoch。


GT:	a airplane that is flying through the sky A passenger jetliner flying through a gray sky a large air plane flying in the air A British Airways airplane taking off into the sky A large passenger airplane taking off into the sky	
Baseline:	a large jetliner flying through a cloudy blue <code><unk></code>	
MIA:	a large white and black and white <code><unk></code>	
PTP:	a large commercial airplane flying through the <code><unk></code>	
MIA+PTP:	an airplane flying in the sky with a sky background	

图 8. 结果展示：可以观察到 MP 组的明显优势（没有 `<unk>` 标签）

6 总结与展望

本文探讨了注意力机制在图像字幕生成任务中的关键作用，通过复现与对比实验，验证了注意力机制在视觉和语言模态融合中的卓越表现。研究表明，基于 Transformer 架构的模型，例如 Vision Transformer (ViT) 和 MIA 模型，可以有效捕获图像特征与文本特征的语义关联，从而生成更加精准和语义一致的图像描述。此外，通过引入 PTP 模块提供位置信息，进一步提升了注意力机制在跨模态任务中的性能，解决了训练过程中的序列信息丢失问题。未来的研究可以进一步探索注意力机制与其他特征提取方法的结合，以及更高效的跨模态表示学习方法。这些改进将为图像字幕生成和其他多模态任务提供更强的模型能力和应用潜力。同时为了进一步验证 MIA 失效的原因，在实验中为文本引入位置特征或者使用一个强大的文本编码器（如 BERT [9]）都是一种有效的办法。

参考文献

- [1] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [2] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [3] Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23242–23251, 2023.
- [4] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [5] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [8] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.
- [9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.