视觉与语言 期末项目报告

黄一凡 2301112019
夏惟 2301112090

复现论文：Ning, S., Qiu, L., Liu, Y., & He, X. (2023). HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23507–23517. https://doi.org/10.1109/CVPR52729.2023.02251

代码地址：https://github.com/AllenYolk/HOICLIP-reproduce

一. 原论文内容简介

1. 研究背景

<HOI detection>
<motivations>
<limitations of the previous work (GEN-VLKT)>

2. 方法

<这里总览一下模型>

实体检测

<DETR (instance decoder)>

关系检测

<interaction decoder>
<Q_{inter}怎么生成>

动词分类

<动词分类任务的目的：缓解HOI detection的label imbalance问题>
<动词特征E_v的生成(visual semantic arithmetic)，以及动词分类分数S_v怎么算(cosine)>
<在这里提一句训练的总体loss S_t = S_{inter} + \alpha S_v>

零样本HOI增强

<无需额外训练，就能利用CLIP知识来增强模型的分类效果>
<training-free enhancement>

二. 改进方案

1. 改进动词特征 E_v

如前一部分所述，原论文在HOI检测任务的基础上额外引入了动词分类任务；通过训练模型同时完成这两个任务（见总训练分数 S_t 的表达式），HOI检测数据中的类别不均衡问题得到了缓解。模型在动词分类任务上的损失是基于interaction decoder所提取出的当前样本的交互特征 O_{inter} 与预先生成的动词特征 E_v 的cosine相似度得到（见动词分类分数 S_v 的表达式），故动词特征 E_v 的质量很大程度上影响了动词分类任务的最终效果，也间接决定了HOI检测效果。因此，我们首先试图对动词特征 E_v 进行优化改进。

原论文中， E_v 通过视觉语义算术得到：作者认为，用HOI类别的CLIP视觉表示向量减去物体类别的CLIP视觉表示向量，就可以得到动词类别的有效表征。原文中的实验也表明，该方法优于基于CLIP文本编码器的方法和对所有含该动词的HOI实例视觉向量去平均的方法（见原文Table 6）。可是，这种只减去物体视觉表征的语义算术没有将“人”这一元素从每个HOI实例的向量中彻底剔除。对于每一个HOI实例的表征向量而言，都应该有 $\sigma^2_{\text{HOI}} = \sigma^2_{\text{verb}} + \sigma^2_{\text{obj}} + \sigma^2_{\text{human}}$ ；但原论文**仅对所有属于该HOI类别的实例的视觉表征向量求平均，希望能够借此粗粒度地消除人的变异性对该HOI类别的向量表征带来的影响**。我们希望在原文视觉语义算术的基础上加以改进，**将人的变异性从每个HOI实例的表示向量中细粒度地排除**（而不是用求均值的方法，从HOI类别表示向量中粗粒度地排除），以求得到更能反映动词本身性质的特征向量组 E_v ，从而提升模型在动词分类任务及最终的HOI检测任务上的表现。

具体来说，对于每一张输入图片 R ，我们提取出其中的HOI、人和物体区域，分别记作 $R_{\text{HOI}}, R_{\text{human}_{\text{HOI}}}, R_{\text{obj}_{\text{HOI}}}$ 。我们将HOI区域CLIP视觉嵌入向量的两倍减去人和物体的CLIP视觉表示，并乘上0.5，得到动词的一个向量表示。对该动词的所有向量表示求和并进行L2归一化，便得到了该动词最终的向量表示。

$$E_v^k = \text{L2Norm}[\sum_{k \in \text{HOI}} \frac{[f(R_{\text{HOI}}) - f(R_{\text{human}_{\text{HOI}}})] + [f(R_{\text{HOI}}) - f(\text{R}_{\text{obj}_{\text{HOI}}})]}{2}]$$

改进后的视觉语义算术充分考虑了每个HOI实例中“人”这一关键成分的变异性，使得动词特征更加合理。

2. 改进 Q_{inter} 的生成方式

Q_{inter} 是interaction decoder的query输入，由instance decoder的输出合并得到，对于HOI检测任务中的HOI三元组分类子任务而言至关重要。在原论文中，对于instance decoder输出的每一对人和物体的特征 O_h, O_o ，先求其平均，再施加一个线性映射，得到 Q_{inter} （见原文公式(4)）。我们认为，这样简单的方法不能很好地将物体和人的特征融合在一起：平均池化操作为人和物体的特征强行分配了相等的权重，需建立在两类特征对等的前提之下；但显然，人和物体的特征在HOI分类任务中并不应该是对等的存在。我们希望能设计更加合理的方法来融合 O_h, O_o ，既能够为两类特征分配合理的相对权重，又能将原特征映射到 Q_{inter} 所在的线性空间中。

受到sparse MLP的启发 (Tang et al., 2022)，我们设计了一种含多阶段线性映射的特征融合方案。首先，我们使用一个 $(2 \to 2)$ 线性映射来调整 O_o, O_h 的相对权重，生成新的特征向量 O'_o, O'_h 。然后，我们使用一个共享参数的线性映射来将 O'_o, O'_h 转换到 Q_{inter} 所在的空间中，记为 O''_o, O''_h 。最后，我们将二者相加，得到 Q_{inter} 。

$$\begin{aligned}
(O'_o, O'_h) &= f_{\text{horizontal}}[(O_o, O_h); W_{2\times 2}] \\
O''_o &= f_{\text{vertical}}(O'_o) \\
O''_h &= f_{\text{vertical}}(O'_h) \\
Q_{inter} &= O''_o + O''_h
\end{aligned}$$

相比原论文中的平均池化，此方法给人和物体特征分配了可学习的权重，让 Q_{inter} 能更好地体现人和物体之间的联系。

三. 实验结果

由于个人服务器的算力有限（两块Nvidia GForce RTX 2080 Ti, 各11G显存），我们无法复现原论文中的所有实验。出于训练开销考虑，我们只在两个数据集上开展复现和实验：V-COCO和5% HICO-Det。

1. V-COCO

V-COCO数据集含81个物体类别，29个动词类别，263个HOI三元组类别。它的训练和测试集中各含5400和4964张图像。由于缺少V-COCO数据集中各HOI标签具体含义的注释文件，且原文作者也并未给出V-COCO数据集上动词表示的生成脚本，故我们无法在V-COCO实验中应用我们的改进。在这一任务上，我们主要进行对原论文实验结果的复现。

表1. V-COCO实验复现结果。

			V-COCO Average Precision	
			scenario 1	scenario 2
原论文结果	baseline		63.50%	64.81%
复现结果	baseline		58.76%	59.31%
	training-free enhancement	k=10	58.86%	59.46%
		k=20	58.87%	59.45%

受到显存限制，我们**将batch size设定成2（原工作batch size为8），其余超参数配置均与原工作完全一致**。训练进行了90个epoch，用时约32小时。从表1可以看出，我们**无法得到与原工作同等水平的V-COCO平均准确率**，在两个测试场景下的平均正确率均比论文中报告的数值低了5%左右。在与原论文作者讨论、查看了其训练日志并反复比对之后，我们一致地将这种性能差距归因为**调小batch size所带来的后果**。

基于训练好的模型，我们进行零样本HOI增强，并重新评估平均准确率。如表1所示，零样本HOI增强使得模型在V-COCO两大测试场景下的平均准确率分别有了0.1%和0.15%左右的提升，而TopK操作的超参数k的具体取值对增强效果的影响并不显著。

2. 5% HICO-Det

HICO-Det数据集含80个物体类别，117个动词类别，600个HOI三元组类别。原始HICO-Det任务的训练集和测试集中各含38118和9658张图像；由于算力和时间的限制，我们**将训练集缩小为原始大小的5%（类别数仍为原始值），而保持测试集的大小不变**。5%训练数据的索引文件由原工作给出，确保了对比的公平性。

表2. 5% HICO-Det实验复现结果。

			HICO-Det Test Average Precision
原论文结果	baseline		22.64%
复现结果	baseline		20.35%
	training-free enhancement	k=10	21.15%

类似于V-COCO实验，受显存影响，我们不得不将batch size设定成2（原工作batch size为8），其余超参数配置均与原工作完全一致。训练进行了90个epoch，用时约27小时。如表2所示，我们得到的HICO-Det**平均准确率低于原工作中同样训练数据规模条件下的结果**，比其报告的数值低了2%左右。有了在V-COCO上的经验，我们认为这种性能差距仍然是调小batch size所带来的后果。我们直接采用原论文中报告的最佳零样本HOI增强超参数设定，令k为10，使得模型平均准确率有了1%左右的显著提升。

表3. 应用改进后，5% HICO-Det上的实验结果。

	HICO-Det Test Average Precision					
	baseline			enhanced, k=10		
	all	rare	non-rare	all	rare	non-rare
原模型	20.35%	**19.33%**	22.78%	21.15%	**20.33%**	23.08%
改进1： E_v	20.09%	18.96%	22.79%	20.87%	19.90%	23.16%
改进2： Q_{inter}	**20.38%**	19.31%	**22.93%**	**21.21%**	20.31%	**23.34%**

我们应用前文提到的两类改进，并将改进后得到的实验结果与我们的原始复现结果进行比较，如表3所示。从整个测试集合上的平均准确率来看，我们对动词表示 E_v 的改进反而使得模型表现有所下降，而对 Q_{inter} 生成方式的改进则略微改善了模型的表现。进一步，我们按照HOI三元组出现的频数是否大于10，将600个HOI类别划分成稀有类别rare（138个）和常见类别（462个），并分别计算在两个类别上的平均准确率。结果表明，我们的**两种改进均使得模型在常见类别上的性能有所提升，但在罕见类别上的性能有所下降。**这一结果揭示了我们的两类改进对模型所产生的效果：**模型的表达能力和对数据的拟合能力有了提升，但是略微牺牲了少样本学习和泛化的能力**；而提升和牺牲的相对程度决定了我们的改进对模型总体表现影响的方向。而应用了零样本增强技术后，得到的结果也相似。

总的来说，表3数据体现了我们所设计的第二类改进的有效性，但也揭露了其牺牲少样本泛化能力的弊病。这表明：能否有效地融合人类和物体的特征表示直接决定了HOI检测任务的最终性能。同时，第一类改进的失败则说明：对动词特征向量过于细粒度的处理可能会导致泛化性能的明显下降。希望我们的项目能对未来的HOI检测工作带来一些启发。

四. 小组成员分工

黄一凡：

- 实验环境配置

- 改进代码实现，运行实验，
- 撰写报告的“改进方案”与“实验结果”部分

夏惟：

- xxx

参考文献

Tang, C., Zhao, Y., Wang, G., Luo, C., Xie, W., & Zeng, W. (2022). Sparse MLP for Image Recognition: Is Self-Attention Really Necessary? *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(2), 2344—2351. https://doi.org/10.1609/aaai.v36i2.20133