Conditional HOTR:基于 Transformer 的人物交互检测

张诗凡,叶海波
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要:人物交互检测任务(HOI 任务)旨在检测出图片中所有存在交互关系的人和物,最后得到<人,动作,物>这样形式的三元组。一般的方法包括两阶段和一阶段算法,最近一些工作提出的基于 transformer 的 HOI 检测方法使整个管道变得更加简单。对于已有的检测模型 HOTR,旨在优化其内部 transformer 结构,使其更好地适应 HOI 检测任务。对于其中用于交互检测的交互解码器,根据其交互查询嵌入分别生成了人和物的参考点,并以此设计了交互点生成公式,然后利用交互点的信息设计了条件交互查询,将其作为位置嵌入与内容嵌入相加得到 query,最后与 key 点乘进行注意力计算。这有助于 transformer 显式地定位与交互相关的区域,缩小搜索范围并缓解对内容嵌入的依赖。最终,在基准数据集 V-COCO 和 HICO-DET 上,mAP 分别提升了 2.13 百分点和 8.33 百分点,并且精度在 V-COCO 数据集上达到了目前最优。

关键词:人物交互检测;计算机视觉;Transformer;查询嵌入;交互点
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2023)08-0023-07
doi:10.3969/j.issn.1673-629X.2023.08.004

Conditional Human-object Interaction Detection with Transformer

ZHANG Shi-fan, YE Hai-bo
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Human-object interaction task (HOI) aims to detect all <human, action, object> triplets in the image that exist interaction relationships. General methods contain two-stage algorithm and one-stage algorithm. Some recent work has proposed a transformer-based HOI inspection approach that makes the whole pipeline much simpler. For the existing detection model HOTR, we aim to optimize its internal structure of transformer to better adapt to the HOI detection task. For the interaction decoder used for interaction detection, we generate the reference points of humans and objects according to its interaction query embeddings, and design the interaction points generation formula. Then, we use the information of the interaction points to design the conditional interaction query, which is seen as the position embedding and added to the content embedding to obtain the query, and do dot multiplication with the key finally. It helps to locate the interaction-related region explicitly, which narrows the search range and ease the dependence on content embedding. We perform the experiments based on HOTR, with the mAP gain of 8.33% on HICO-DET and 2.13% on V-COCO, and the mAP is currently SOTA on the V-COCO dataset.

Key words: human-object interaction detection; computer vision; transformer; query embedding; interaction point

0 引 言

人物交互检测(HOI)是一项视觉关系检测任务,旨在将一幅图片中所有具有交互关系的人和物体成功配对,识别出人和物体的位置和类别,以及交互动作的类别,以帮助更好地理解场景。这可以表示为检测一组<人,动作,物>的 HOI 三元组。

目前主要有两类 HOI 检测方法,一类是顺序 HOI 检测,也叫做两阶段方法,另一类是并行检测,即一阶段方法。两阶段方法将 HOI 检测任务解耦为目标检测任务和交互分类任务,可想而知这种方法比较耗时、昂贵。在一阶段方法中,人类通过先验知识预先定义交互检测的规则,有些工作借助交互点[1-2]、交互框[3]来定位交互关系。因为目标检测可以和交互分类并行,所以这类一阶段方法更加高效,但它们仍然需要手工后处理阶段来对匹配规则进行匹配。

最近,因为 NLP 领域 transformer[4]的火热应用,以

收稿日期:2022-10-08 **修回日期**:2023-02-08
基金项目:国家自然科学基金青年基金(61702261)
作者简介:张诗凡(1997-),女,硕士,通讯作者,研究方向为人物交互检测;叶海波(1987-),男,副教授,研究方向为机器学习。

及受到一些将 transformer 运用到 CV 领域的工作——如 DETR 的启发和影响,一些基于 transformer 的端到端 HOI 检测算法被提出。如 DETR[5] 一样,它们将 HOI 检测看作是一个集合预测问题,因此消除了对额外的手工后处理阶段的需要。利用 transformer 强大的建模能力,它们提取图片的全局信息,transformer 的解码器通过交叉注意力模块中的 query 来查询与交互相关的特征,之后解码器的输出结果通过检测头,以端到端的方式得到 HOI 关系。这些方法解决了一阶段算法的问题,即不需要手工后处理而是直接端到端检测,取得了较好的效果,但仍然面临新的挑战。

在目标检测领域,DETR[5] 和 Conditional DETR[6] 模型都利用 transformer 来实现端到端的目标检测,它们发现,在 transformer 的交叉注意力模块中,内容嵌入起主要作用,而位置嵌入对 mAP 的贡献很小。然而,交叉注意力的内容嵌入必须同时匹配 key 的内容嵌入和位置嵌入,并且其所定位的区域对于检测物体的位置和类别非常重要,因此高质量的内容嵌入是非常必要的。因此,笔者认为在基于 transformer 的 HOI 领域,同样需要高质量的内容嵌入来识别和定位,要减少对内容嵌入的依赖。

受 Conditional DETR[6] 的启发,针对交叉注意力层,该文提出了条件交互查询,它作为位置嵌入缩小了搜索范围以帮助显式地定位与交互相关的区域。这样,对于内容嵌入方面的要求便没那么高了,因为它可以更专注于内容方面来进行识别,而定位则交由条件交互查询。虽然 HOI 检测任务与目标检测任务类似,但两者存在一定的差异。HOI 检测的关键是正确匹配人物对,而不仅仅是检测物体实例,因此需要根据 HOI 检测任务的特点来设计条件交互查询。该文实现了基于 HOTR[7](一种基于 transformer 的 HOI 检测模型)的条件交互查询。与 Conditional DETR 不同的是,通过交互 query 来预测生成人和物体的参考点,以此来表示人和物体在交互中的位置,并且让它们参与检测头的最终预测过程。

然后,根据设计的公式由人和物体参考点生成交互参考点,可以理解这些交互参考点定位了与交互相关的区域。对于设计条件交互查询,包含了两部分内容:交互参考点和当前解码器的输出,因为它们都包含了与定位相关的信息,所以要将这些信息都考虑进去。将该模型称为 Conditional HOTR,它改进了 transformer 的注意力机制,以便更好地适应 HOI 检测任务。与基线方法 HOTR 对比,Conditional HOTR 的 mAP 在 V-COCO 上提高了 2.13 百分点,在 HICO-DET 上提高了 8.33 百分点,并且在 V-COCO 数据集上精度达到了目前最优。

1 相关工作

1.1 传统人物交互检测

传统人物交互检测算法可以划分为两阶段和一阶段方法。

1.1.1 两阶段 HOI 检测

在两阶段方法中[8-21],首先会执行目标检测任务,预训练的目标检测器首先检测人和物体的边框及其对应的类别,然后将所有检测出的人和物体两两组合配对,将所有成对的组合传入一个单独的神经网络进行训练和交互分类。一些工作利用人类姿态[9,14,17,20] 来检测 HOI 关系,这有助于细粒度交互类别的检测。还有一些工作提出了以实例为中心[11,18] 的注意力机制、面向动作[15] 的关系推理网络进行 HOI 检测。ACP[13] 的目的是解决 HOI 的长尾分布问题。还有一些方法用图结构表示 HOI 关系[10,16,19]。还有方法[21] 基于关系推理的交互实例推荐网络来进行 HOI 检测。

1.1.2 一阶段 HOI 检测

由于两阶段方法受到其冗余推理结构的限制,一些工作提出了一阶段方法,它意味着交互分类检测可以与目标检测并行,而无需再等待其结果才能进入下一阶段。在这些方法中,设计一个合理的匹配模式是将目标检测与交互检测结果匹配的关键。IPNet、PPDM[1-2] 将 HOI 检测视为交互点检测问题,通过将目标检测器检测到的人和物体与交互点关联来完成匹配。除交互点外,Uniondet[3] 将人与物体的联合框看作是交互区域来定位交互。由于不需要训练全部的人物组合对,一阶段方法的时间复杂度得到了很大程度的降低,但是它们仍然需要为 HOI 检测手工设计匹配策略。

1.2 基于 transformer 的人物交互检测

随着 transformer 在 NLP 领域的广泛应用,它最近也被用来解决计算机视觉领域的一些任务,如目标检测。transformer 擅长捕捉远距离依赖关系,这正是 HOI 检测所需要的,transformer 帮助聚合图片中的全局上下文信息。

作为目标检测领域的一项创新性工作,DETR[5] 利用 transformer 以端到端的方式来检测物体,许多工作[7,22-24] 都受其启发,并将其改进为 HOI 检测器以简化检测管道。这些方法将 HOI 检测视为一个集合预测问题,并通过匈牙利算法执行二部匹配来将预测结果和真实值进行一一对应,然后根据匹配的结果计算损失。QPIC[24] 设计了 query,每个 query 最多可以捕获一个人物对,所以即使不同 HOI 关系的实例距离很近,也可以单独提取每个 HOI 实例的特征。CDN[22] 在分析和总结一阶段和两阶段 HOI 检测方法优缺点的基础上,提出了一种新的方法,即以级联的方式分离目

标检测和交互分类。HOTR[7]为 HOI 检测设计了一个共享编码器和两个并行解码器,分别是实例解码器和交互解码器。

1.3　CV 领域对于 transformer 的改进

随着 transformer 在 CV 领域的广泛应用,许多研究不再仅仅满足于 transformer 的基本结构,因为图像的注意力权重计算量很大,而且一些基于 transformer 的工作的训练收敛速度较慢。在目标检测领域,对于 DETR 进行改进的一些工作[6,25-26]致力于解决上述 transformer 的问题。Deformable DETR[26]提出了一种新的注意力模块,它只关注目标物体参考点周围少量的关键采样点,而不是所有像素点。SMCA[25]通过预测物体区域的高斯映射,加快了检测的收敛速度。Conditional DETR[6]提出了条件位置嵌入,它可以明确地定位物体的边缘,缩小搜索的范围,从而解决检测依赖高质量内容嵌入的问题。

而在 HOI 检测领域,尚未有研究考虑 transformer 结构的缺点,受 Conditional DETR[6]的启发,设计了条件交互查询,以改进 transformer 的结构更好地用于 HOI 检测。

2　文中方法

2.1　概　述

在本节中,详细了介绍基于改进 transformer 结构的 HOI 检测方法,帮助显式定位与交互相关的区域并缩小搜索范围。该方法的整体实现框架如图 1 所示(其中虚线框即为图 2 所示的条件交互查询的生成方法)。

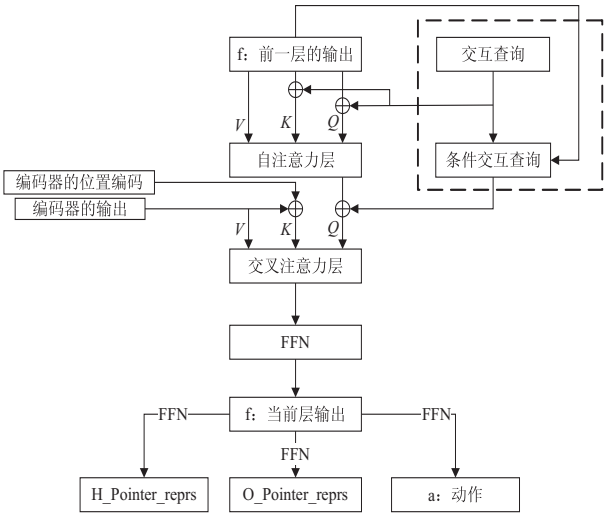

图 1　Conditional HOTR 整体框图

Conditional DETR[6]认为解码器中的自注意力模块的主要功能类似于非极大值抑制(NMS),它不涉及查询交互区域,所以只在交叉注意力模块中设计了条件交互查询。设计条件交互查询如图 2 所示。

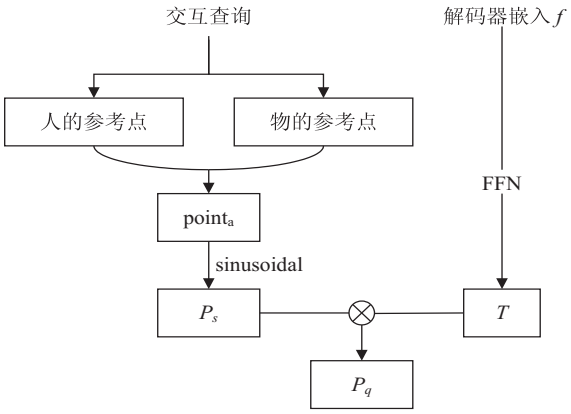

图 2　条件交互查询 p_q 的生成

2.2　相关知识回顾

2.2.1　HOTR

HOTR[7]是基于 transformer 的 HOI 检测模型,它主要由四个部分组成:骨干网络、transformer 结构、检测头和组合配对。首先 CNN 网络提取图片特征,接着将这些特征与映射到正弦编码空间的位置编码相加,结果输入 transformer 结构,做进一步的特征增强。transformer 结构由一个共享的编码器和两个并行解码器构成,其中一个是负责生成实例表示的实例解码器,一个是交互解码器,它负责生成交互表示信息。只在交互解码器上应用条件交互查询,因为在实例解码器上进行目标检测的改进是 Conditional DETR 所做的工作。实例解码器的检测头预测物体的边框和类别,而交互解码器的检测头负责预测人指针表示、物指针表示和交互类别,而不是直接回归人和物体的边界框。在最后的组合配对部分,对于之前得到的 human 指针表示和 object 指针表示,分别寻找与其相似度最高的实例表示(实例表示即实例解码器的输出结果),找到的索引即为相应的 human 指针和 object 指针,通过这种方法将具有 HOI 关系的人物进行匹配,完成 HOI 检测。

人和物的指针表示向量定义为:

$$v_i^h = \mathrm{FFN_h}(f_i) \tag{1}$$

$$v_i^o = \mathrm{FFN_o}(f_i) \tag{2}$$

其中,f_i 表示解码器输出的第 i 个交互表示,给定 N 个 interaction query 则得到 N 个交互表示。

2.2.2　Conditional DETR

为了解决 DETR 训练收敛速度慢的问题,Conditional DETR 提出了条件交叉注意力机制,通过条件位置查询嵌入直接寻找物体的边缘区域,以此来缩小搜索物体的范围。简要介绍下它是如何生成条件位置查询的。

Conditional DETR 是基于解码器嵌入 f(即当前解码器层的输出)以及参考点信息 s(由 object query 生

成,有多少个 query 就生成多少个参考点,代表了每个 query 所定位的区域)这两部分信息来进行边框预测的。因此,Conditional DETR 认为,条件位置查询的生成也应该考虑这两部分,因为它们包含了与位置相关的信息。即,条件位置查询 p_q :

$$(s,f) \rightarrow p_q \tag{3}$$

2.3 检测头

检测头部分内容见图 1。对于交互解码器,最终的检测头不直接回归人和物体的边框,而是像 HOTR 一样,预测 human 指针和 object 指针的表示,回归实例边框的部分由实例解码器完成。为 Conditional HOTR 设计了一种新的预测方法来生成指针表示信息:

$$\text{H\_Pointer\_reprs} = \text{normalize}(\text{FFN}_{\text{h1}}(f) + \text{FFN}_{\text{h2}}(\text{h\_reference\_point})) \tag{4}$$

$$\text{O\_Pointer\_reprs} = \text{normalize}(\text{FFN}_{\text{o1}}(f) + \text{FFN}_{\text{o2}}(\text{o\_reference\_point})) \tag{5}$$

其中,f 表示解码器嵌入,即当前解码器层的输出,h_reference_point 和 o_reference_point 是 human 参考点和 object 参考点,它们是由 interaction query 经过两层 MLP 预测得到的 2D 坐标,并且 N 个 interaction query 分别生成 N 个 human 参考点和 N 个 object 参考点。这些坐标用于表示人和物体的参考位置。$\text{FFN}_{\{\text{h,o}\}1}$ 由三层 MLP 组成,作用于解码器嵌入 f 得到初步的 human 指针表示和 object 指针表示。这正是公式(1)和(2)所表示的。然后,通过 $\text{FFN}_{\{\text{h,o}\}2}$ 将 human 参考点和 object 参考点映射到与 $\text{FFN}_{\{\text{h,o}\}1}(f)$ 相同的维度(设置为 256),并且将两者的结果相加。normalize 意味着对结果进行 L2 标准化操作。

对于动作类别预测,则保持不变。

$$a = \text{FFN}_a(f) \tag{6}$$

2.4 条件交互查询设计

提出的条件交互查询有助于交互解码器的交叉注意力模块定位交互相关区域,因此在设计它时,考虑所有与位置相关的组件。在 2.3 节中,详细描述了 Conditional HOTR 的检测头,它利用解码器嵌入 f 以及人和物的参考点来预测得到人和物体的指针表示。由于指针表示不仅包含实例的类别信息,还包含了位置信息,因此在设计条件交互查询时,将这两部分考虑在内,即 f 以及人和物体的参考点。

那么,如何利用人和物体的参考点呢? 根据 HOI 检测的特点,设计了公式,使用人和物体参考点来计算出交互参考点,定义交互点位于人和物体的参考点中间连线上,这也符合真实世界的逻辑。交互参考点 a 可以表示为:

$$\text{point}_a = \text{ratio} * \text{o\_reference\_point} + (1 - \text{ratio}) * \text{h\_reference\_point} \tag{7}$$

ratio 是一个超参数,它的值应该在[0,1]之间。不同的交互点定位不同的 HOI 三元组。使用交互点生成条件交互查询有助于显式地定位与交互相关的区域。

然后,遵循 Conditional DETR 的步骤,将 point_a 映射到 256 维的正弦编码空间,使得它与 key 的位置嵌入编码方式保持一致:

$$p_s = \text{sinusoidal}(\text{point}_a) \tag{8}$$

对于另一个包含位置相关信息的成分:解码器嵌入 f,还遵循 Conditional DETR 的操作,即 f 通过一个两层的 MLP,形成可学习的转换 T。因此,最终的条件交互查询 p_q 的组成是:

$$p_q = Tp_s \tag{9}$$

最终,p_q(即位置查询嵌入)与自注意力层的输出(即内容查询嵌入)相加作为交叉注意力模块的 query,参与最后的注意力计算,即,query 与 key 进行点乘得到注意力权重。

3 实 验

为了证明 Conditional HOTR 是有效的,在本节中展示了比较全面的实验。

3.1 数据集和评估指标

3.1.1 数据集

在 HICO-DET[8] 和 V-COCO[27] 这两个被 HOI 检测任务广泛使用的数据集上进行了实验,以验证文中方法的有效性。HICO-DET 包含了 47 776 张图片(38 118 张用于训练,9 658 张用于测试),并且包括超过 150 K 对的人物对。它有 117 个动作类别和 80 个物体类别,构成 600 个 HOI 三元组,其中 138 个是稀少类别(即少于 10 个训练实例),其余 462 个类别为非稀少类别。V-COCO 是 MS-COCO[28] 的一个子集,其中包括 10 346 张图片(2 533 张用于训练,2 867 张用于验证,以及 4 946 张用于测试)。它包含 29 个动作类别,每个都是一个二进制标签,还包含了 80 个物体类别。

3.1.2 评估指标

与 HOTR 一样,使用 mAP 作为评估指标。对于检测结果,仅当预测的边框和对应的真实边框的交并比(IOU)大于 0.5,并且物体类别和动作类别都预测正确时,HOI 检测结果被视为正确的正样本。对于 V-COCO,报告了两个场景的 mAP:场景 1 需要报告没有物体的情况,而场景 2 则忽略这种情况。对于 HICO-DET,评估默认情况下的性能,即根据所有测试图像来计算 AP。报告了三种类型的 mAP:所有类别(Full)、稀少类别(Rare)和非稀少类别(Non-Rare)。

3.2　实现细节

因为只修改了 HOTR 中交互解码器的交叉注意模块及其最终的检测头部分,其他的都遵循原始的模型结构,所以训练过程与 HOTR 几乎相同。使用 AdamW[29]对模型进行训练,将主干网络的学习率设置为 1e-5,权重衰减为 1e-4。对于 V-COCO,将 transformer 的初始学习率设置为 1e-4,对于 HICO-DET 设置为 1e-5。与 HOTR 一样,主干特征提取网络、编码器以及实例解码器加载在 MS-COCO 上预训练的模型,这些权重在模型训练期间被冻结。增强机制和损失函数与 HOTR 相同,并且,对模型训练 100 个周期,其中学习率在 80 个周期时衰减一次。

3.3　实验结果

在 V-COCO 和 HICO-DET 上进行实验。表 1 展示了在 V-COCO 数据集上的实验结果,以及基线方法 HOTR 和最近的 SOTA 方法的结果。表 2 是在 HICO-DET 数据集上的结果。将 ResNet-50 作为主干网络。对于 HICO-DET 数据集,目标检测器在 MS-COCO 上进行预训练。与基线方法 HOTR 进行比较时,为了体现出与其结果比较提升的程度,采用提升了多少百分比的形式;而与其他 SOTA 方法比较时,直接利用表格中 mAP 的值进行相减得到差值来直观对比 mAP。

表 1　在 V-COCO 数据集上的结果

	Method	Backbone	AP^{s1}_{role}	AP^{s2}_{role}
两阶段方法	InteractNet[12]	ResNet-50-FPN	40.0	48.0
	GPNN[16]	ResNet-101	44.0	-
	iCAN[11]	ResNet-50	45.3	52.4
	DCA[18]	ResNet-50	47.3	-
	DRG[10]	ResNet-50-FPN	51.0	-
	VSGNet[30]	ResNet-152	51.8	57.0
	ACP[13]	ResNet-152	53.0	-
	IDN[31]	ResNet-50	53.3	60.3
一阶段方法	UnionDet[3]	ResNet-50-FPN	47.5	56.2
	IPNet[1]	Hourglass-104	51.0	-
基于 transformer 的方法	HOI Transformer[23]	ResNet-101	52.9	-
	QPIC[24]	ResNet-50	58.8	61.0
	HOTR[7]	ResNet-50	61.0	65.8
	CDN-B[22]	ResNet-50	62.3	64.4
	Ours(ratio=0.25)	ResNet-50	62.3	67.0

表 2　在 HICO-DET 数据集上的结果

	Method	Backbone	Detector	Default		
				Full	Rare	Non-Rare
两阶段方法	InteractNet[12]	ResNet-50-FPN	COCO	9.94	7.16	10.77
	GPNN[16]	ResNet-101	COCO	13.11	9.41	14.23
	iCAN[11]	ResNet-50	COCO	14.84	10.45	16.15
	DCA[18]	ResNet-50	COCO	16.24	11.16	17.75
	DRG[10]	ResNet-50-FPN	COCO	19.26	17.74	19.71
	VSGNet[30]	ResNet-152	COCO	19.80	16.05	20.91
	ACP[13]	ResNet-152	COCO	20.59	15.92	21.98
	IDN[31]	ResNet-50	COCO	23.36	22.47	23.63
一阶段方法	UnionDet[3]	ResNet-50-FPN	COCO	17.58	11.72	19.33
	PPDM[2]	Hourglass-104	HICO-DET	21.73	13.78	24.10
基于 transformer 的方法	HOTR[7]	ResNet-50	COCO	21.73	18.73	22.63
	HOI Transformer[23]	ResNet-50	COCO	23.46	16.91	25.41
	QPIC[24]	ResNet-50	COCO	24.21	17.51	26.21
	QAHOI[32]	ResNet-50	COCO	24.35	16.18	26.80
	Ours(ratio=0.25)	ResNet-50	COCO	23.54	22.08	23.97

3.3.1　与基线 HOTR 比较

考虑到不同 gpu 设备对实验结果的影响,重新跑了一遍 HOTR 的源码,并将此结果作为文中方法的基线,以此来体现公平。可以看到,在 V-COCO 测试集上,比 HOTR 提高了 2.13 百分点(61.0→62.3),在 HICO-DET 上提高了 8.33 百分点(21.73→23.54)。表明文中方法在两个基准上都得到了明显的提升,尤其是在 HICO-DET 上,这验证了 Conditional HOTR 的有效性。

3.3.2　与 SOTA 方法比较

在 V-COCO 测试集上,Conditional HOTR 优于所有的两阶段方法和普通的一阶段方法。对于基于

transformer 的 HOI 检测方法,它优于大多数方法,例如相比于 HOI Transformer,超过其 9.4 mAP,超过 QPIC 3.5 mAP。与目前的 SOTA 方法 CDN 相比,在同等条件的 ResNet50 为主干网络的情况下,Conditional HOTR 与其具有相同的精度。值得一提的是,文中方法在场景 2 上达到了 SOTA。

对于 HICO-DET 数据集,Conditional HOTR 优于所有两阶段方法和普通一阶段方法。此外,文中方法优于基于 transformer 的 HOI 检测方法——HOI Transformer。在基线 HOTR 的结果和 QPIC 结果之间差距 2.48 mAP 的情况下,文中方法最终仅比 QPIC 低 0.67 mAP。QAHOI 利用多尺度特征进行 HOI 检测,这对检测结果有利,文中方法没有使用多尺度,比它低了 0.81 mAP。

3.4 消融实验

为了验证设计的条件交互查询的有效性,设置成不同的超参来观察其对结果的影响,不同的参数设置会导致交互点处于不同的位置。从表 3 可以看出,不同的 ratio 检测精度不同,但在两个数据集上结果都优于基线 HOTR,因此文中方法是有效的。

表 3 不同的 ratio 值的结果

Method	AP^{s1}_{role}	AP^{s2}_{role}	Default		
			Full	Rare	Non-Rare
HOTR[7]	60.98	65.83	21.73	18.73	22.63
Ours(ratio=0.25)	62.31	66.95	23.54	22.08	23.97
Ours(ratio=0.5)	62.06	66.59	23.19	21.70	23.63
Ours(ratio=0.75)	61.49	66.03	23.44	21.49	24.02

此外,还做了一些额外的实验,通过改变设计的 Conditional HOTR 的结构来深入探讨其有效性。所有实验均在 V-COCO 上进行,并使用 ResNet50 作为主干,ratio 设置为 0.5。表 4 是实验的结果。Conditional HOTR-Q 表示取消了人和物的参考点的设计,并直接生成一个可学习的向量作为图 1 中的 p_s。Conditional HOTR-P 表示在最终的检测头中,公式(4)和(5)中的 $FFN_{h1}(f)$ 和 $FFN_{o1}(f)$ 直接加上交互点的信息,而不是分别与人和物的参考点信息相加。从结果可以推断,参考点和检测头的设计是有效的。

表 4 在 V-COCO 上进行消融实验

Method	AP^{s1}_{role}	AP^{s2}_{role}
Conditional HOTR	62.06	66.59
Conditional HOTR-Q	61.02	65.54
Conditional HOTR-P	61.78	66.65

4 结束语

提出了条件交互查询,旨在优化基于 transformer 的 HOI 检测方法,并在 HOTR 上验证了其有效性,称它为 Conditional HOTR。在解码器中充当交叉注意的位置查询嵌入,显式地定位与交互相关的区域,减少了对高质量内容查询的依赖。使用交互点和当前解码器层的输出来生成条件交互查询,因为它们包含与位置相关的信息。通过人和物体的参考点来生成交互参考点,其表示 HOI 三元组的定位区域。文中方法在两个基准数据集上都比 HOTR 有显著改进,并且超过了大多数的 HOI 检测方法。

参考文献:

[1] WANG T,YANG T,DANELLJAN M,et al. Learning human-object interaction detection using interaction points[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:4116-4125.

[2] LIAO Y,LIU S,WANG F,et al. PPDM:parallel point detection and matching for real-time human-object interaction detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020:482-490.

[3] KIM B,CHOI T,KANG J,et al. Uniondet:union-level detector towards real-time human-object interaction detection[C]//European conference on computer vision. Glasgow: Springer,2020:498-514.

[4] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. Advances in Neural Information Processing Systems,2017,30:5998-6008.

[5] CARION N,MASSA F,SYNNAEVE G,et al. End-to-end object detection with transformers[C]//European conference on computer vision. Glasgow: Springer, 2020: 213-229.

[6] MENG D,CHEN X,FAN Z,et al. Conditional detr for fast training convergence[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE,2021:3651-3660.

[7] KIM B,LEE J,KANG J,et al. Hotr:end-to-end human-object interaction detection with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:74-83.

[8] CHAO Y W,LIU Y,LIU X,et al. Learning to detect human-object interactions[C]//2018 IEEE winter conference on applications of computer vision (wacv). [s. l.]: IEEE, 2018:381-389.

[9] FANG H S,CAO J,TAI Y W,et al. Pairwise body-part attention for recognizing human-object interactions[C]//Proceedings of the European conference on computer vision (ECCV). [s. l.]:Springer,2018:51-67.

[10] GAO C,XU J,ZOU Y,et al. Drg:dual relation graph for human-object interaction detection[C]//European conference on computer vision. Glasgow:Springer,2020:696-712.

[11] GAO C,ZOU Y,HUANG J B. ICAN:instance-centric attention network for human-object interaction detection[J]. arXiv:1808.10437,2018.

[12] GKIOXARI G,GIRSHICK R,DOLLÁR P,et al. Detecting and recognizing human-object interactions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:8359-8367.

[13] KIM D J,SUN X,CHOI J,et al. Detecting human-object interactions with action co-occurrence priors[C]//European conference on computer vision.[s. l.]:Springer,2020:718-736.

[14] LI Y L,ZHOU S,HUANG X,et al. Transferable interactiveness knowledge for human-object interaction detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:3585-3594.

[15] LIN X,ZOU Q,XU X. Action-guided attention mining and relation reasoning network for human-object interaction detection[C]//Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence. Yokohama:[s. n.],2021:1104-1110.

[16] QI S,WANG W,JIA B,et al. Learning human-object interactions by graph parsing neural networks[C]//Proceedings of the European conference on computer vision (ECCV). Munich:Springer,2018:401-417.

[17] WAN B,ZHOU D,LIU Y,et al. Pose-aware multi-level feature network for human object interaction detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:9469-9478.

[18] WANG T,ANWER R M,KHAN M H,et al. Deep contextual attention for human-object interaction detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:5694-5702.

[19] ZHOU P,CHI M. Relation parsing neural network for human-object interaction detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Los Alamitos:IEEE,2019:843-851.

[20] 王苡蓉,吴静静. 基于关键姿态的快递场景人-物交互行为识别方法[J]. 计算机测量与控制,2022,30(6):182-189.

[21] 薛丽霞,尹凯建,汪荣贵,等. 基于交互实例推荐网络的人-物交互检测方法研究[J]. 光电工程,2022,49(7):45-57.

[22] ZHANG A,LIAO Y,LIU S,et al. Mining the benefits f two-stage and one-stage hoi detection[J]. Advances in Neural Information Processing Systems,2021,34:17209-17220.

[23] ZOU C,WANG B,HU Y,et al. End-to-end human object interaction detection with hoi transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:11825-11834.

[24] TAMURA M,OHASHI H,YOSHINAGA T. Qpic:query-based pairwise human-object interaction detection with image-wide contextual information[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:10410-10419.

[25] GAO P,ZHENG M,WANG X,et al. Fast convergence of detr with spatially modulated co-attention[C]//Proceedings of the IEEE/CVF international conference on computer vision.[s. l.]:IEEE,2021:3621-3630.

[26] ZHU X,SU W,LU L,et al. Deformable detr:deformable transformers for end-to-end object detection[J]. arXiv:2010.04159,2020.

[27] GUPTA S,MALIK J. Visual semantic role labeling[J]. arXiv:1505.04474,2015.

[28] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco:common objects in context[C]//European conference on computer vision. Zurich:Springer,2014:740-755.

[29] LOSHCHILOV I,HUTTER F. Decoupled weight decay regularization[J]. arXiv:1711.05101,2017.

[30] ULUTAN O,IFTEKHAR A S M,MANJUNATH B S. Vsgnet:spatial attention network for detecting human object interactions using graph convolutions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:13617-13626.

[31] LI Y L,LIU X,WU X,et al. Hoi analysis:integrating and decomposing human-object interaction[J]. Advances in Neural Information Processing Systems,2020,33:5011-5022.

[32] CHEN J,YANAI K. QAHOI:query-based anchors for human-object interaction detection[J]. arXiv:2112.08647,2021.