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

1.1 人物交互(Human-Object Interaction，HOI)检测

人物交互(Human-Object Interaction，HOI)检测是计算机视觉中的一项任务，涉及识别和理解图像或视频中人与物体之间的交互。该领域对于开发机器人、监视、人机交互和增强现实等领域的先进系统至关重要。HOI 检测通常涉及识别场景中人与一个或多个物体之间的关系。这可以包括人类对物体所做的动作，比如“拿着手机”、“从杯子里喝水”或“看书”。目标是使机器能够理解视觉数据中发生的复杂交互。

HOI 检测任务主要涉及三个子问题，包括物体检测、人机配对和交互识别。研究人员和从业人员使用深度学习技术，特别是卷积神经网络(CNN)和循环神经网络(RNN)来解决 HOI 检测挑战。而目前基于 Transformer 的神经网络模型逐步占据主流地位。以往的 HOI 检测方法可分为两阶段和一阶段两种。两阶段范式方法使用独立的检测器来获得对象的位置和类别，然后是专门设计的用于人-对象关联和交互识别的模块。一种典型的策略是使用基于图的方法提取关系信息，以支持交互理解。取而代之的是，单阶段范式直接检测具有交互的人-物对，而不需要分阶段处理。最近基于 Transformer 的探测器启发的 HOI 方法取得了很好的性能。这些模型在带有人物体交互标签的大型数据集上进行训练，以学习和推广模式。精确的 HOI 检测模型的开发有许多应用，包括人类行为分析、视频监控、人机交互等。

图 1. HICO-DET 基准测试中的 4 个已注释的示例: 每张图像对应一组边界框对和一个标签，每个边界框对定位一个人和一个对象，并预测一个 HOl 类标签。

1.2 原论文的背景和动机

(1) 传统方法的限制

由于端到端对象检测器的发展, 最近的研究在定位交互中的人类对象实例方面取得了显著进展。尽管如此，识别人-物对之间的交互类的问题仍然特别具有挑战性。传统策略只需学习多标签分类器，通常需要大规模注释数据进行训练。因此，它们经常受到长尾类分布的影响，并且缺乏对看不见的交互的泛化能力。而论文提出的方法就在于解决传统方法所面临的挑战。

(2) 视觉语言预训练

最近，对比视觉语言预训练已被探索来解决此类开放词汇和零样本学习问题，因为其学习的视觉和语言表征在各种下游任务中表现出强大的迁移能力。特别是，最近关于开放词汇检测的工作利用知识蒸馏将 CLIP 的对象表示转移到对象检测器。这种策略已被用于 HOI 检测工作，包括 GEN-VLKT 和 EoID，它们利用 CLIP 的知识来解决 HOI 任务中的长尾和零样本学习。

(3) 仍然叩待解决的问题

虽然视觉语言预训练模型的运用给 HOI 检测提供了巨大的便利, 但如何有效地将 CLIP 知识转移到 HOI 识别任务中仍然是一个需要解决的问题。它涉及由视觉对象和交互组成的组合概念。而普遍采用的师生蒸馏目标与提高学生模型的泛化能力不一致。且学习 HOI（如 GEN-VLKT）中的知识蒸馏通常需要大量的训练数据，这表明其数据效率较低。此外，知识蒸馏通常会在零样本泛化中遇到性能下降的问题，因为它缺乏对看不见的类的训练信号，而这对于从教师模型继承知识至关重要。

(4) 论文的主要贡献

a) HOICLIP 是第一项利用基于查询的知识检索从预先训练的 CLIP 模型到 HOI 检测任务的有效知识转移的工作。

b) 论文开发了一种细粒度的转移策略，通过交叉注意力利用 HOI 的区域视觉特征，并通过视觉语义算法利用动词表示来获得更具表达力的 HOI 表示。

c) 论文通过利用零样本 CLIP 知识而无需额外培训，进一步提高了 HOICLIP 的性能。

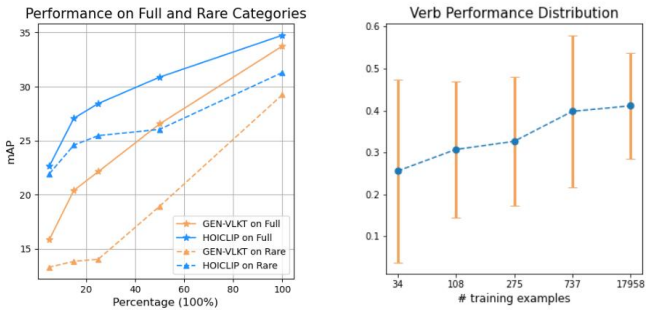
1.3 GEN-VLKT 方法的缺陷

GEN-VLKT 是论文主要参考的方法之一。改文章提出的 HOICLIP 方法基本可以看成是 GEN-VLKT 方法的改进。GEN-VLKT 设计的亮点在于设计了一个两分支 pipeline，以提供并行的正向过程，并使用人和对象的分离查询，而不是统一查询。具体关于 GEN-VLKT 架构的设计细节会在第 2 章“方法”中介绍。GEN-VLKT 方法存在的主要缺陷如下：

(1) 普遍采用的师生蒸馏目标与提高学生模型的泛化不一致；

(2) 数据效率低（如图 2 实验结果所示）；

(3) 知识蒸馏在零次泛化过程中，由于缺乏对隐性类的训练信号而导致性能下降，而隐性类是继承教师模型知识的关键。



(a) Data Efficiency Comparison　　(b) Verb Long-tail Problem

图 2. 数据效率比较和动词分布分析: 在(a)图中，将训练数据从 5%增加到 100%，并显示了 HOICLIP 和 GEN-VLKT 的结果。在(b)图中，数据点表示平均 mAP，垂直线的长度表示按样本数分组的动词的 mAP 的方差。

2、方法

HOICLIP 的总体架构如图 3 所示。该方法首先采用基于 Transformer 的端到端实体检测器 DETR 来定位人和对象。在给定人和物体特征的情况下，HOICLIP 引入了一种新的交互解码器来进行交互识别，其中 HOICLIP 利用来自先前提取的特征图 Vd 和 CLIP 生成的空间特征图 Vs 的信息，并通过交叉注意力模块进行知识集成。随后，动词适配器提取动作信息以增强交互表示和识别。线性分类器利用交互解码器的输出来预测 HOI 类别，该类别通过使用 CLIP 的语言特征的无训练分类器得到进一步增强。GEN-VLKT 的总体架构如图 4 所示。可以看出，两个模型的模型架构大体相似。HOICLIP 主要在 GEN-VLKT 的细节部分做了改进，并提出了零样本 HOI 增强。

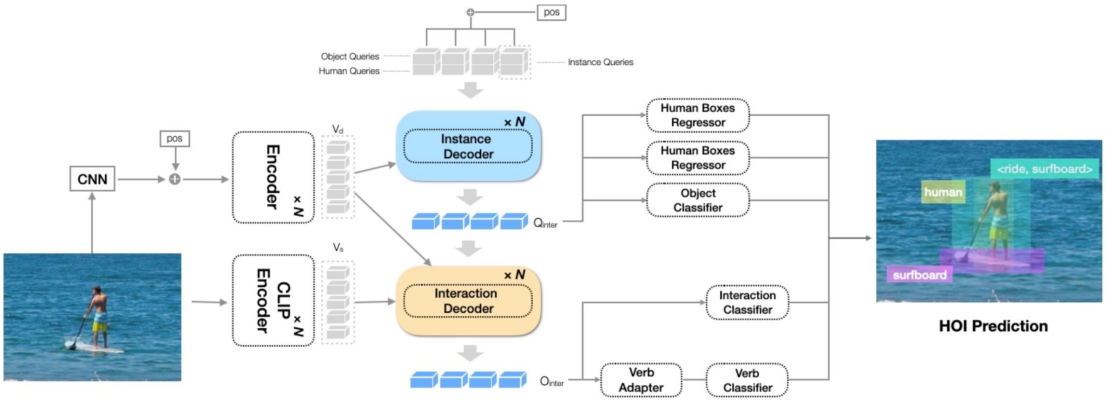


图 3. HOICLIP 的体系结构：给定一个图像，HOICLIP 用检测编码器和 CLIP 编码器对其进行编码。实例解码器使用来自检测编码器的特征来定位人和对象对。交互解码器利用来自编码器和提取交互表示两者的特征。动词适配器基于交互表示提取动词表示。

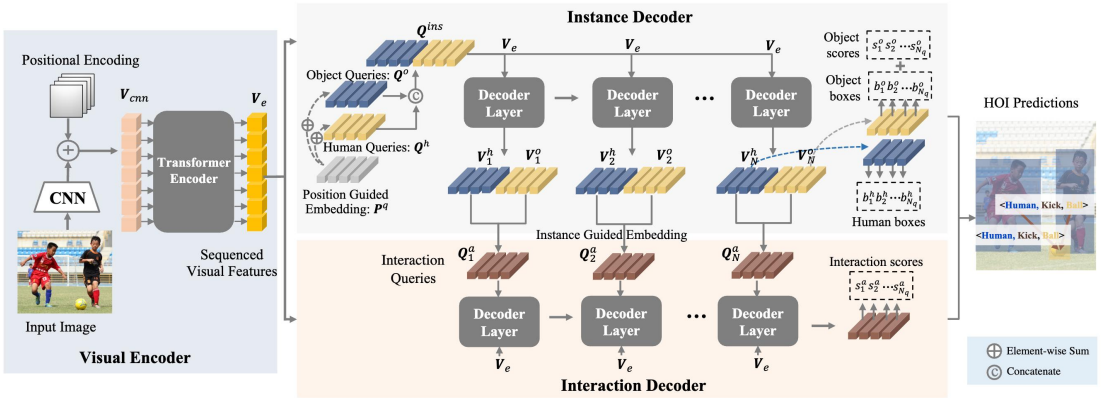


图 4. HOICLIP 的 GEN 的框架：GEN-VLKT 被组织为一个配备了两个分支解码器架构的视觉编码器。给定图像，首先应用视觉编码器来提取视觉特征。然后，使用两个分支，即实例解码器和交互解码器，分别基于可学习查询来定位人类对象对和分类 HOI 三元组。此外，

HOICLIP 设计了一个位置引导嵌入（p-GE）来关联交互式人和对象，并设计了一种实例引导嵌入（i-GE）来使交互查询在特定的人和对象查询的指导下预测相应的 HOI 类别。

2.1 实体检测

HOICLIP 首先采用基于 Transformer 的端到端实体检测器 DETR 来检测实体(人和对象)。

DETR 是 2020 年 ECCV 上提出的方法，在当时引起了较大的反响。DETR 将目标检测任务视为一个图像到集合（image-to-set）的问题，即给定一张图像，模型的预测结果是一个包含了所有目标的无序集合。而传统的目标检测（以 faster-rcnn 为代表）的流程：backbone 提取特征——>利用 RPN 枚举所有的框并筛选 region proposal——>在 region proposal 上得到每个框的类别和置信度。

具体而言 DETR 解决了传统方法存在的如下问题：(a) 枚举了每个特征图上的像素点，在每个像素点上枚举预定义的 anchor, 造成大多数的候选框是坏的, 无效的, 缓慢的; (b) RPN 输出了太多冗余的框需要 NMS 来删除；(c) 手工涉及的元素只有很少的超参数可以调节；(d) 模型 tuning 比较复杂。

DETR 方法的框架如下图 5 所示：

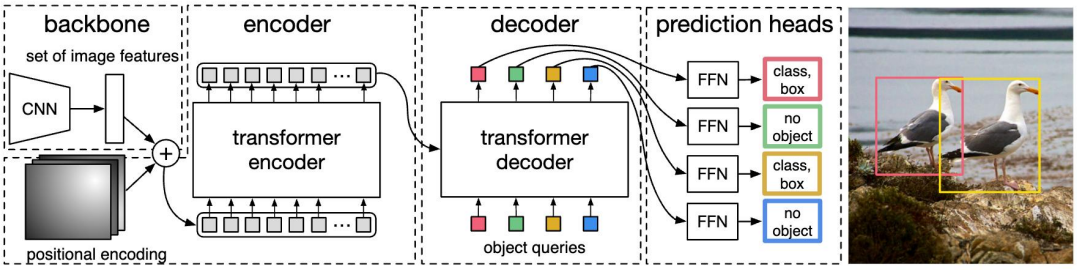


图 5. DETR 框架：DETR 使用传统的 CNN 主干来学习输入图像的 2D 表示。该模型对其进行平坦化，并在将其传递到转换器编码器之前用位置编码进行补充。然后，转换器解码器将少量固定数量的学习位置嵌入作为输入, 这个过程称之为对象查询, 并额外处理编码器输出。DETR 将解码器的每个输出嵌入传递到共享前馈网络（FFN），该网络预测检测（类和边界框）或“无对象”类。

2.2 关系检测

为了预测一对人和对象查询的 HOI 类别，HOICLIP 通过将人和对象特征 O_h 和 O_o 提供给投影层来生成一组交互查询 $Q_{inter} \in \boldsymbol{R}^{Nq \times Cs}$。为了充分利用 CLIP 知识，HOICLIP 从 CLIP

中检索与分类器权重中的先验知识更好地一致的交互特征。具体而言 HOICLIP 将 CLIP 空间特征 V_s 和投影检测视觉特征 V_d 保留为与 V_s 相同的维度：

$$Q_{inter} = \text{Pool}(O_o, O_h)W_i + b_i$$
$$V_d' = V_d W_p + b_p$$

其中 W_i, b_i, W_p, b_p 是投影参数，$V_d' \in \boldsymbol{R}^{Hs \times Ws \times Cs}$，Pool 取平均值。为了指导交互查询 $Q_{inter} \in \boldsymbol{R}^{Nq \times Cs}$ 探索 V_s 和 V_d 中的信息区域，HOICLIP 设计了一个用于知识集成的交叉注意力模块，其架构如图 6 所示。Q_{inter} 首先通过自注意进行更新，然后分别输入到具有 V_s 和 V_d' 的交叉注意模块中，获得两个输出特征。公式如下：

$$Q_{inter} = \text{SelfAttn}(Q_{inter}),$$
$$C_{inter} = \text{CrossAttn}(Q_{inter}, V_s),$$
$$D_{inter} = \text{CrossAttn}(Q_{inter}, V_d'),$$
$$Q_{inter} = \text{FFN}(C_{inter} + D_{inter})$$

其中 V_s、V_d' 分别是关键字和值，Q_{inter} 是共享交叉注意力中的查询。为了提取最终的交互表示 $O_{inter} \in \boldsymbol{R}^{Nq \times D}$，HOICLIP 采用与 CLIP 相同的投影操作，将交叉注意力的输出转换到 CLIP 特征空间，如下所示：

$$O_{inter} = \text{Proj}(Q_{inter})$$

通过这种方式，HOICLIP 利用来自实例解码器的对象和人类信息，从 CLIP 的空间特征图中检索交互表示，并从检测器中检索视觉特征。

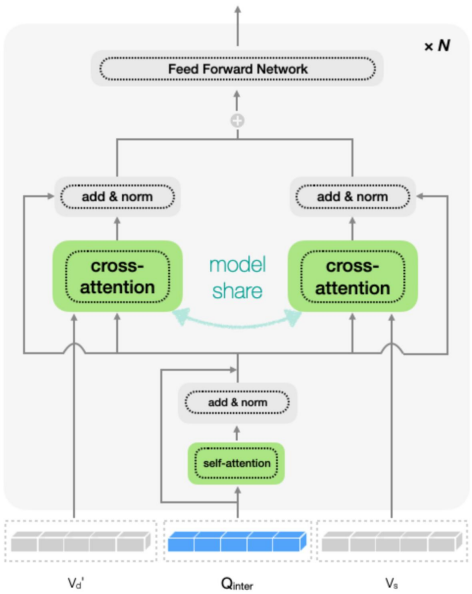


图 6. 基于 Cross Attention 的知识整合结构：交互查询首先经过一个自关注层。然后，它被馈送到具有 V_s 和 V_d 的两个共享的交叉注意力层中。将输出相加并输入前馈网络。

2.3 动词分类

HOICLIP 引入了一种新的提取全局动词类表示的管道和一种基于 CLIP 特征构建的动词分类器，以应对标签不平衡。视觉语义算法为了更好地从自然不平衡的 HOI 注释中捕获细粒度的动词关系，HOICLIP 通过视觉语义算法构建了一个动词分类器，该分类器表示训练数据集的全局动词分布。HOICLIP 假设动词类表示可以从 HOI 的全局视觉特征和其对象的全局视觉特性的差异中得出。如图 7 所示。

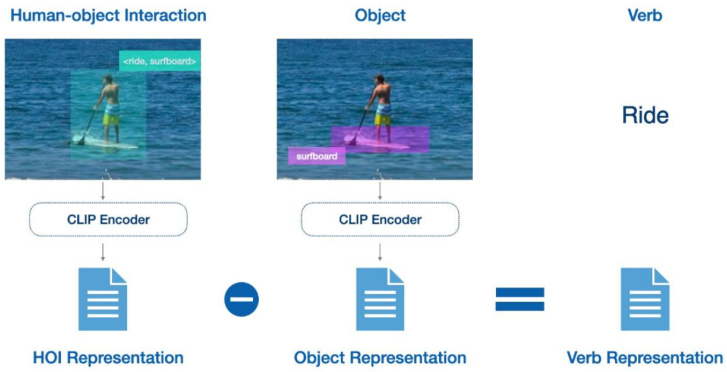


图 7. 视觉语义算术：通过对对象和 HOI 的裁剪区域进行编码来提取对象和 HOI 表示。然后，通过 HOI 表示减去宾语表示得到动词表示。

具体来说，HOICLIP 使用覆盖对象的最小区域和人类边界框来表示 HOI 三元组。然后，HOICLIP 将 $\mathbf{OBJ}_j$ 定义为包含对象类 j 的所有实例的集合。此外，这里使用元组(i,j)来指代一个 HOI 类别。其中 i 和 j 分别代表动词和对象的类别。类似地，HOICLIP 将 $\mathbf{HOI}_{(i,j)}$定义为包含 HOI 类别(i,j)的所有实例的集合。对于 HOI 和对象区域，HOICLIP 使用 CLIP 图像编码器来获得它们的视觉特征，然后采用投影仪将特征映射到全局特征空间中。形式上，给定一个区域 R，计算其特征如下：

$$f(R) = \text{Proj}(\text{VisEnc}(R))$$

动词类 k 的表示是通过取平均 HOI 和对象区域特征的差来计算的：

$$E_h^{k,j} = \text{L2Norm}(\sum_{R_m \in \mathbb{HOI}_{k,j}} f(R_m))$$

$$E_o^j = \text{L2Norm}(\sum_{R_n \in \mathbb{OBJ}_j} f(R_n))$$

$$E_v^k = \text{L2Norm}(\sum_{n \in (k,\cdot)} (E_h^{k,n} - E_o^n))$$

为了使用动词类表示进行分类，HOICLIP 设计了一个轻量级适配器模块来提取基于交互特征 O_{inter} 的动词特征 $O_{verb} \in \boldsymbol{R}^{Nq \times D}$。具体来说，HOICLIP 使用 MLP 将交互特征映射为动

词特征 $O_{verb} \in \boldsymbol{R}^{Nq \times D}$，其中动词 *logits* S_v 被计算为动词特征 O_{verb} 和动词类表示 E_v 之间的**余弦相似性**。计算动词类得分如下，

$$O_{verb} = \mathrm{MLP}(O_{inter}),$$
$$S_v = O_{verb} E_v^{\mathrm{T}}$$

在训练过程中 HOICLIP 通过结合 HOI 预测 S_{inter} 和动词预测 S_v 来获得训练 HOI *logits* S_t:

$$S_t = S_{inter} + \alpha \cdot S_v$$

2.4 零样本 HOI 增强

最后，HOICLIP 引入了一种由 CLIP 文本编码器中的先验知识生成的 HOI 分类器，它为 HOI 分类提供了一种无训练的增强。具体而言，HOICLIP 通过利用 CLIP 学习的可视化语言对齐来构建零样本 HOI 分类器，其中使用 CLIP 文本编码器 TextEnc 嵌入的标签描述作为分类器权重。HOICLIP 使用手工制作的模板“一个人的照片[动词]和[对象]”将每个 HOI 类别转换为一个句子。将模板输入到 CLIP 文本编码器 TextEnc 中，以获得 HOI 分类器 $E_{inter} \in R^{Kh \times D}$，其中 K_h 是 HOI 类别的数量。为了利用零样本 CLIP 知识，HOICLIP 根据图像 V_g 的全局视觉特征和 HOI 分类器 E_{inter} 计算一组额外的 HOI *logits*。为了滤除低置信度预测，HOICLIP 只保留顶部 K $\in$ [0，Kh]的分数。公式如下：

$$S_{zs} = \mathrm{TopK}(V_g E_{inter}^{\mathrm{T}})$$

其中 TopK 是选择具有最高 K 分的 HOI 逻辑的操作，并且 S_{zs}^{i} 指示第 i 个 HOI 类别的得分。更新后的 S_z 是一种无训练的高置信度 HOI 预测，它利用零样本 CLIP 知识来有益于尾类预测。给定零样本 HOI 分类器 E_{inter}，HOICLIP 还使用基于交互的表示 O_{inter} 生成交互预测得分

$$S_{inter} = O_{inter} E_{inter}^{\mathrm{T}}$$

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，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（部分）
- 撰写报告的“改进方案”与“实验结果”部分

夏惟：

- 实验环境配置
- 改进代码实现，运行实验（部分）
- 撰写报告的论文介绍部分