基于 EPIC-KITCHENS 数据集的组合动作识别

Yangcen Liu
School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
2020080601018@std.ue s tc.edu.cn

摘要—近年来视频理解问题**(video understanding)**已经取得了长远的发展，各项工作已经能够在各个数据集取得良好的效果。然而，组合动作识别**(compositional action recognition)** 却没有得到很好的解决，其与普通识别任务相比，组合动作识别识别内容抽象，训练集与测试集分布不一致，涉及长尾**(l ong-tail)**问题和元学习**(Zero-Shot)**问题。我们的工作中，提出了 一种 **compaction** 模型，为了解决在之前组合动作识别没有能解决的泛化问题。主流的方法主要有：将组合动作识别视为普通的视频理解问题，或将所有物体映射为无偏的特征进行处理。前者非常容易对训练集进行过拟合，不具备良好的鲁棒性，后者则对于模型的特征信息完全丢失。而我们的 **co mpaction** 架构作为一种数据增强模式，在训练样本中引入了配对**(pair)**以及交换**(mix up)**操作，尝试在保留原有物体本身特征信息的基础上，对于物体特征进行泛化，在未见标签类别**(unseen)**和已见标签类别**(seen)**得到更好的学习效果。我们 在**Something-else**[19]以及 **Epic-kitchens**[6]数据集上进行了实验， 取得了很好的提升效果。我们后续的工作是在 **compaction** 的架构上引入外部语义信息进行更好的泛化。代码开源在：**http s://github.com/Randle-Github/statistical_learning_final**

关键词：视频理解，组合动作识别，tickets 交互

I. 引入

视频理解近年取得了较大的进展，如一些基于卷积的识别方法([15], [16], [17], [20])，基于 transformer 的方法([1], [3], [4])，以及基于图的方法([5], [18])取得了较大的进展。然而一般的动作识别问题没有关注组合的问题。对于一般的动作分类标签， 是将(noun1,verb,noun2) 视为一个整体，即测试集与训练集的分布是一致的。但是对于组合动作识别课题在[19] 中最开始提出， 在something- else 数据集中关注了这样的一种数据分布形式： (somethi ng,verb,something)。对于一个视频动作而言，它可能与 多种不同的 object 进行交互。多数情况下是与手和某种物体的交互。但是在[19]中提出了训练集分布与测试集 分布不同的情况，即对于同一个动作，与之交互的objec t 在训练集中完全处于未知(unseen)状态。此时“动作” 标签就比较抽象，而在缺乏一定推理能力，而仅仅适用于概率分析的深度学习框架很难去处理这种 Zero-Shot 的情况。

我们选择了 EPIC-KITCHENS2021 数据集上的挑战，即关于组合动作识别的问题。该数据集由来自 4 个城市的32 个厨房拍摄的第一人称视角视频组成，这些视频包含 1150 万张图像，展示了约四万个动作示例和五十万个物体对象，是有史以来最大的使用可穿戴相机的视频数据集。我们这次挑战的目的就是识别出其中的组合动作， 而不受对象的干扰。换言之，就是将动作分解为动词、主语、以及一个或者多个目标。在此之前，我们已经在一个较小的数据集——Something-Something V2 数据集取得成果，该数据集只有二十多万个视频，包含 174 个

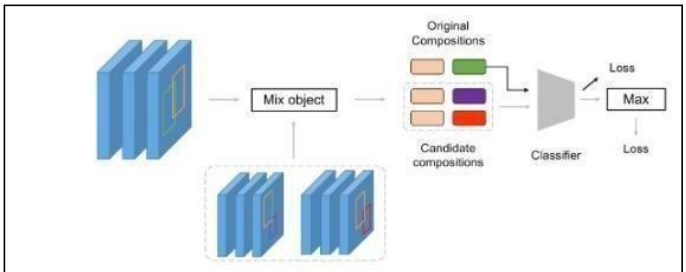


图 4 compaction 的外部结构。其中对不同视频样本的物体进行 Mixup 操作。Classifier 可以内接 IRN[8]的结构，但是输出层进行更改。

标签。其中的难点就是很多时候我们可以学得视频中的组合动作，但当在相同的动词以及模型没见过的名词上 面训练和测试，模型的准确率便大打折扣。模型泛化能 力弱。动作的预测最终可能被模型识别为一个对于物体 有偏的结果，实际上没有真正识别动作，而是模型记忆 了物体组合与动作之间的关系。我们从视频数据集中随 机间隔抽出一定帧数后先使用已训练好的 faster RCNN (即该 faster RCNN 已在别的数据集学得对物体的识别)在这些帧数做 detection，把其中与主语交互的物体识别出来，提取特征并标注，然后再用 transformer 学得物体与物体、物体与背景以及背景与背景之间的联系，这些学 习到的关系对组合动作的识别非常关键。

我们提出了一种 compaction 架构，使用了一种对于 o bject token 进行 mix up 操作的方法。我们的整个模型基于 IRN(Interaction Reasoning Network)[8]，我们使用的是 STLT[22]中提供的代码进行更改。创新点在于，对于 obje ct token，我们根据标签，在同一个 batch 中进行了组合交换的数据增强模式。后续的损失函数进行了相应的改进，而总体的 backbone 仍然是一个从空间域提取信息到时间域提取信息的多分支 transformer 结构。

II. 相关工作

2.1 传统动作识别

过去的工作中主要是在普通的动作识别问题中进行了研究。侧重了物体的交互，而没有对于动作这一抽象的标 签进行理解。最开始基于卷积的方法诸如 I3D[20], Slow Fast[15], TSN[16], TSM[17]等方法取得了很好的效果。在[4] 中首次将 transformer 结构引入视频，后续一些常见的架构有 VAT[1], ViViT[3], video transformer[4]，最近的一个统一架构是 Motionformer[21]，主要可以追踪整个 object 的运行轨迹，将时空域空间域进行了结合。一些基于图的 方法还有诸如[5], [18]则致力于将视频的人物进行交互建模为图结构，希望得到一些较好的推理能力。

然而传统的动作识别无法对于 Zero-Shot 进行处理，也不具备任何的推理能力。这个阶段的主要研究主要是时间域和空间域的信息聚合方式。训练的动作结果仍然严重地受到object 的有偏影响。

2.2 组合动作识别*Zero-Shot learning*

在[22]中ZSL(Zero-Shot learning)的survey 中提出，一般对于无样本的情况一般有三种形式： 比较早的有deep metric learning，后期的语义建图辅助推理，以及使用生成方法进行数据增强。

目前应用于组合动作识别的方法主要是后两种。使用度量学习的方法对于视频这种噪声极大且高维的特征暂时没有很好的解决方法。现在主流是使用生成方式进行数据增强的办法，也有使用知识图谱进行辅助建模的。

在[19]中提出的组合动作识别课题，并且提出了一种基本的通过 object token 进行后续处理的方法。在这篇文章中，提出了一种在空间上聚合每一帧的object token 的信息，再进行时序上的聚合方法，得到了一种基本的推理形式。这种方式是后续的一个很好的基准工作。

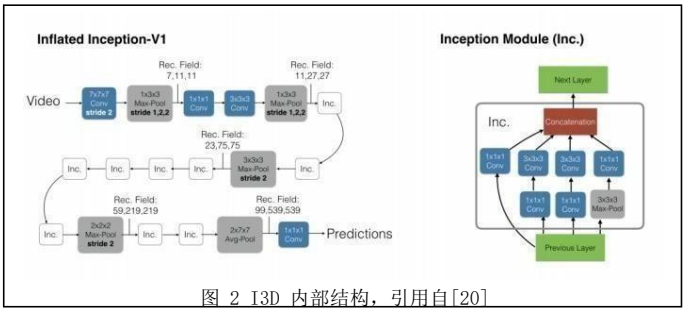


图 2 I3D 内部结构，引用自[20]

在后续一些比较有影响力的工作有Motionformer[21]。基于了VAT，对于 3D token 提取之后的聚合，不简单的使用时序与空间的结合，而是采取了对于轨迹的追踪。也就是说需要使用一个预训练的 object detector(采用的 faster-rcnn[23])对于每一帧进行追踪。选取 prototype(基于object 的 token 选取)，计算出所有 token 与 prototype 的关系，并且输入 transformer 进行特征提取。这种方法由于聚合方式中带有轨迹信息，所以在识别任务具备较强的效果。这个方法虽然没有对于组合动作进行研究，但是是后续工作的一个基准。

后续提出了 ORViT[24]以及 STLT[25]得到了一些好的推理结果。而对于 Zero-Shot 中使用外界数据集辅助推理的有 ConsNet[26]，使用了外接词袋数据集进行辅助推理。其使用了GAT 对于词向量(Word Vector)进行信息聚合。这个模型有两个流，一个是普通的预测模型，另一个是辅助的词向量聚类模型。

在human object reasoning network[8]中，也是与 ORViT 类似的模式，主要是聚合方式的不同。目前最主流的方法基本还是源自 MotionFormer 提出的架构。

对于Zero-Shot learning 至今做到的比较好的效果是能够完全去除 object 对于预测的影响，但是确实损失了所有的 appearance 信息。我们期望的解决方式是去除不必要的 object 的偏差，但是保留 object 对于模型的影响。从信息的角度上这样的方法是更加合理的。

COMPACTION 模型阐述

这一部分我们将会介绍我们提出的 compaction 模型，其中首先会介绍 SlowFast 模型。SlowFast 将作为整个模型的 3D 卷积特征提取部分。

3.1 *SlowFast*

这个部分我们介绍我们用于作为基准比对的 SlowFast [15]模型。我们复现 SlowFast 除为了进行对比，也是为了对于 epic-kitchens 部分的数据处理部分进行改造使用。

3.1.1 *I3D*

对于 SlowFast 我们首先引入 I3D[20]结构，2D 卷积网络主要是用作图像识别，对于视频的处理，单纯的 2D卷积网络肯定无法满足需求。视频实质上是由一帧帧 的图像构成的，比图像多出了的是时间序列信息。I3D 多出的一个维度用来处理时间序列信息，也就是将 2D 卷积网络中N * N 的 2D 卷积核在时间维度上复制 N 份，然后除以时间的维度 N，就可以得到扩展后的 3D 卷积核。其中网络具体架构图 x，包括卷积层和池化层，中间加入了插入残差模块。

3.1.2 *SlowFast*

SlowFast[15]， 模型是我们的baseline， 也是现有的一个效果比较好的模型。

在图像识别中，一般会对称地处理两个空间维度 x 和 y， 在长时间的统计分析也验证了这种操作的合理性。自然图像近似具备各向同性(即所有方向具有相同的可能 性)和平移不变性。但是对于视频信号，其拥有三个维度 x,y 和 t， 并非所有的时空方向都拥有相同的可能性。慢速动作比快速动作更有可能发生，就如现实生活 中我们肉眼看到的物体大多是静止的(包括环境)。更具体的， 拍摄飞机起飞的视频中，一般只有飞机在高速移动，其他如机场、树木等都是静止的。因此，对于视频 信号， 若采用图像识别那样对称地处理时间和空间， 效果可能不够好。而 SlowFast 便利用了上述的信息，强调快慢采样提取时空特征的思想分别处理视频中的慢动 作和快动作，并取得不错的效果。该模型示意图如图 x 所示。

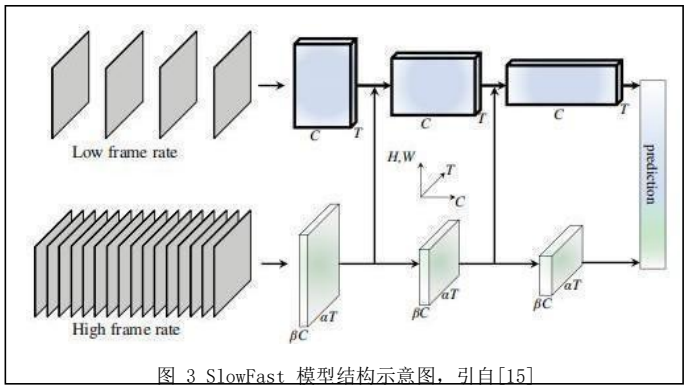


图 3 SlowFast 模型结构示意图，引自[15]

如图 x，输入分为两个路径，我们称上面的为 Slow 路径，下面的为 Fast 路径，最后两条路径的输出分别进行全局平均池化后组合的结果输入到一个全连接分类层来识别行为。

在此可以列出两条路径的具体区别：

1) 从采样看，他们采样的不同在于对视频中提取帧的频率不同，如 Slow路径以较低的帧率和较慢的刷

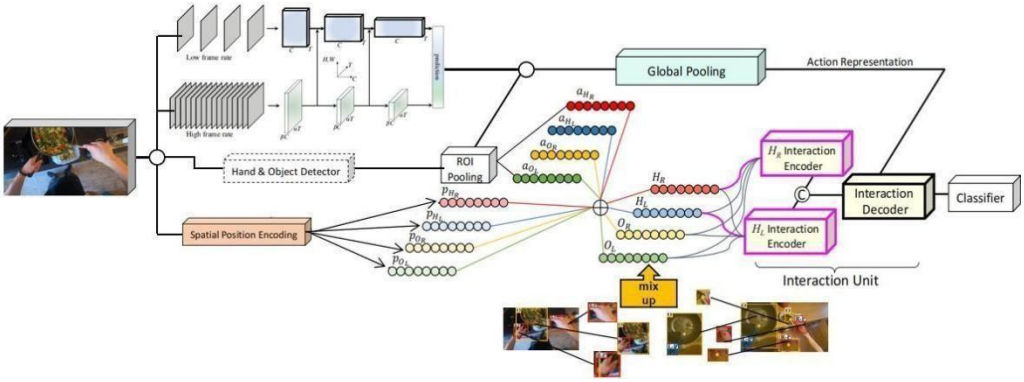


图 5 compaction 结构。对于视频使用 slowfast 进行特征提取，进行 detector 得到ROI 特征，后续进行 reasoning 并且加入 mix up 的方法进行模型泛化。

新速度运行，输入视频帧的时间跨度 stride τ 很大，也就是说每 τ 帧才处理一帧。典型的 τ 值可以取 16，也就是说对于 30fps 的视频，Slow 路径每秒大约采样两帧。而 Fast 路径以较快的更新速度和高时间分辨率运行， 其输入的时间跨度为 τ / α (其中 α 是快慢分路的帧率比， α >1)，也就是说对于同一个视频，Fast 路径处理的视频帧数是 Slow 路径的 α 倍，典型的 α 值可以取 8。

2) 从卷积层看，Slow 路径前几层使用 2D 卷积，后两层才用 3D卷积(实验发现比全用 3D卷积效果更好)；Fast 路径每一层都用的是 3D 卷积，但是各层维持时域维度大小不变以尽可能地保留时域信息。

3) 从效果上看，Slow 路径用来关注空间域，捕获图像或稀疏帧所提供的语义信息，另一条 Fast 路径保持着时间保真度，负责快速捕获运动的变化，但空间细节少。

再结合两者来看，虽然 Fast 路径提取的帧数更多，但其计算量大约只占整体计算量的 20%，这是因为该路径的通道数较少以及空间信息处理能力较差，而这些少的信息又可以由Slow 路径较少的冗余来提供。所以Fast 路径可以当作是以较小的代价来提供 Slow 路径精度的方法。

*3.2 Compaction*模型

后续我们将介绍我们的 compaction 模型架构，模型整体结构如图 X 所示。对于 classifier 是一个对于视频进行分类的结构 STIN[25]，其中是对于空间信息进行聚合再对于时间信息进行聚合的方法。

我们的 compaction 模型分为三个流，Appearance 特征流，object 流(此处 human 和 object 统称 object)，spatial 流。其中 object 流从 appearance 中提取ROI 特征，并进行编码；spatial 流对于每一帧的 object 进行编码，与 ROI 特征进行 c 后续进入 transformer 中进行 concatenate 操作，得到 object token。后续对于不同类型的 object token 进行 transformer 的交互，并且引入 mix up 机制，在解码阶段之后最终实现分类任务。

*3.2.1 Appearance*特征流

对原始 RGB 图像，通常的做法是使用卷积方法得到其 Appearance 特征。我们的模型使用了主干网络为在 Kinetics-400[13]数据集上预训练的Resnet101的 Slowfast网络结构。对于整个视频进行了编码得到appearance特征。

*3.2.2 Object*流

对于输入的视频进行采样后，我们在每一帧上面使用 一个 fixed Faster-RCNN[23]，对于得到的 bboxes(boundin g boxes)分为 left hand，right hand 和 object。对于 Appearance 中根据 bboxes 的坐标提取 ROI(region of interest)特征。此时每一帧记录置信度最高的 m 个 object。

*3.2.3 Position*流

Compaction 模型使用了一个 spatial 流进行空间特征的 提取 。在 每 一帧 中进 行空 间 特征 的编 码 。在[8] 中证明了使用 bboxes 中心坐标以及大小得到的元组进行映射具有较好的效果。我们同样沿用这个方法。使用一个可 学习的映射得到 position encoding。这种形式可以很好的记录物体的空间移动状态。

3.2.4 Mix Up

对于同一个 object得到的 ROI 特征与 position encoding 进行 concatenate，得到带有时序的 object 单元。对于同一个 object 的 t 个单元，进行一个 mapping 映射，得到 object token。三种类型的 bboxes 分别使用相同的映射权值，得到不同的 object token。我们认为这种模式能够很好的记录每一个 object 的状态。

不同于先前工作中的预测结构，我们在训练时会针对训练集 进行以标签为引导的(label-guidedpair)操作，从而交换不同视频样本的物体特征，即 Mix Up 操作。然后再使用组合的特征训练模型。其中对于同一个动作类别 中不同的 object 进行配对。

选择特征进行MixUp可以训练模型从未见过的组合，弥补数据集在物体-动作共现矩阵上的空缺。这个做法的动机也是基于 ZSL 中的“生成”数据集的思路。对于这种特殊的数据增强模式，对于 Zero-Shot learning 而言，可能得到的组合现实中是不存在的，但是也是基于 action 的一个好的泛化模式。

*3.2.5 Loss*评估

而输入模型的 loss 部分，我们采取了两个流。在训练的时候，对于 original 和 candidate 两个流都进行 loss 的计算，其中candidate 流多个组合中取损失最大的一个类

图 1 Epic-Kitchens 数据集展示

别(放弃其余组合) 进行梯度传播； 而在测试部分，仅仅保留 original 流。

III. 实验结果展示

在这一部分， 我们首先将介绍我们所使用的基准数据集(Something-Something v2 和 Epic-kitchens)，检验方式(evaluation)以及细节呈现(implementation details)。我们也会将我们的方法与目前最好的(state-of-art)方法进行 比较，并且对结果进行具体的分析。我们所有的实验在 8 块 NVIDIA GeForce RTX 2080 Ti 8GB 上进行。

4.1 实验数据集

4.1.1 Something-Something v2

一个有 220,847 个标记的视频剪辑的集合。这些视频剪辑是人类对日常物品执行预定义的基本动作。它旨在训练机器学习模型，以精细地理解人类的手势，例如将某些东西放入某物中，将某物颠倒过来并用某物覆盖某些东西。物体已被密集地注释用于训练，但是没有注释手的左右，因此物体不与手相关联。

这个数据集独特的地方在于， 测试集与测试集分别有 seen(测试集与训练集同分布)和 unseen(测试集与训练集不同分布，Zero-Shot)，对于 seen 的部分，测试集中所有的标签在训练集中都出现过。而对于 unseen，所有的标签都是用了不同的物体对于动作进行重新组合，极大的增加了任务难度。

4.1.2 Epic-kitchens-100

目前世界上最大的第一视角视频数据集， 记录了多个多角度、无脚本、本地环境中的厨房场景。数据集采集了 4 个城市、32 个厨房、总计 55 个小时的全高清视频数据，共包含动作边框数 39594 个，物体边界框数 454158 个。

这个数据集独特的地方在于，他具备了超巨大的规模， 每个标签对应的视频长度也较长，场景转换也较多。这 个数据集的长尾问题(long-tail)也比较严重。

*4.2 evaluation*模型检测

对于两个数据集，我们使用了同样的验证方法：top 1/5 准确率。我们对于(noun, verb, noun)的元组标签形式，只关注了verb 的检验结果。使用了一个单层结果输出。

4.3 细节呈现

4.3.1 采样方式

我们沿用了 baseline 的 SlowFast 中的采样方式，首先在 TSN[16]中提出。对于每一个视频平均分成 32 段，在每一段中随机采样一帧进行后续操作。其中 detector 也在采样的RGB frame 进行。

4.3.2 目标检测

我们使用了在 Kinetics-400[14]进行预训练的 Faster-R CNN 网络作为 detector。对于目标中的左手、右手以及物体三个类别进行选取。完整的可视化结果如图 X 所示。这是一个没有使用阈值限制的选框情况，按照 x 中的结果我们取 threshold_obj=0.01，以及 threshold_hand=0.1 得到的效果最好，如图 X 所示。

Method	Top-1(%)	Top-5(%)
STIN[19]	37.2	62.5
I3D[20]+STIN	51.5	77.1
STRG[18]+STIN	56.2	81.3
compaction	69.2	91.7
compaction+Mixup_Random(Ours)	69.7	91.7
compaction+Mixup_Oracle(Ours)	70.3	92.0

图 8 在something-else 进行的实验

Method	Top-1(%)
TSN[16]	33.19
SlowFast[15]	63.64(我们跑出来是65.17)
compaction	63.68
compaction+Mixup_Random(Ours)	还没跑
compaction+Mixup_Oracle(Ours)	还没跑

图 9 在Epic-kitchens 进行的实验

检测的错误根据[11]中得出，在 something-else 比基于 ground truth 的结果差 7%左右。

4.3.3 Mix Up

我们的融合方式对于训练集进行以标签为引导的(label-guided pair)操作。其中对于同一个动作类别中不同的 object 进行配对，在 detector 的输出部分，我们使用了 fixedFaster-RCNN，将权值固定，输出进行交换，进入 interaction block。一般是在同一个 batch 中进行交换，对于 batch 较小的情况下，可能交换的频率比较低，不能做到很好的 Mix Up，所以训练过程中尽量增大 batch_size 的设置。我们设置的 batch_size=128。

同时我们对于训练时进行了三中 compaction 的实验，使用了不加 Mix Up，添加 oracle Mix Up(action label guided), random Mix Up(random exchange)。我们认为 oracle 是一种理论上更有效的聚合方式。

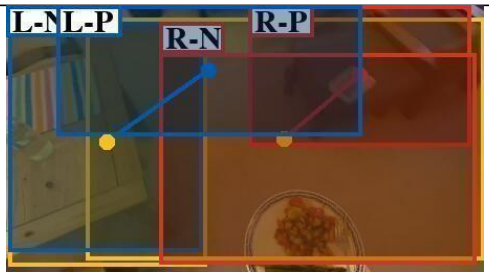


图 6 不设阈值的检测结果

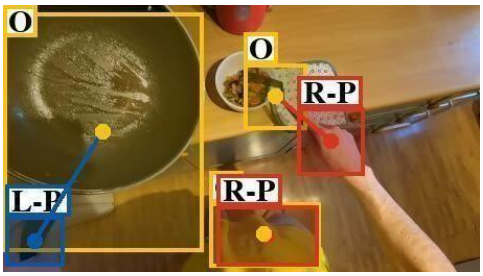
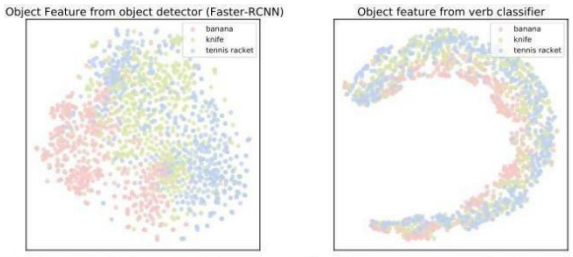


图 7 增设阈值的检测结果



(b) Object feature of *banana*, *knife*, *tennis racket*, from COCO pre-trained Faster-RCNN and the last FC of our verb classifier. Different colors indicate different object categories.

图 11 使用另一种 mixup 前后的可视化 token 对比，引自[13]

4.4 与先前的比较

我们在 something-else 数据集(后续增加了部分数据修复的 something-something v2 数据集)以及 Epic-kitchens 数据集进行实验。最终得到的结果如下：

在 Epic-kitchens 数据集由于计算资源不够一直没有进行实验。可以观察到在 something-else 数据集上，Mix Up 操作可以对于 baseline 有着显著的提升。但是也可以看到组合方式对于最后的结果影响并不大。Oracle 和 Random 的结果几乎相同。

4.5 实验结果分析

我们展现的实验结果表明了 Mix Up 方式具备一定的泛化性。对于组合动作识别数据集中的Zero-Shot问题，Mix Up 能够有效解决 object 对结果的有偏影响。然而 random Mix Up 方法并没有比oracle 方式差别过多。在[13]中提到了一种可能的解释。如图中提及，其结构使用了 一种类似于混合分类器的方式，与 Mix Up 操作可能消除物体的有偏影响。 但是action-guided方法由于大量 object 在不同 label 中间的重复，导致最终与 random Mix Up 相差不大。我们期望得到的一种对于 object 有偏但是减少差异的映射方式，但是这种 Mix Up 操作并不能得到这个效果，而是更加倾向于完全无偏的映射。

并且对于 Zero-Shot 的处理方式，由于这种 Mix Up 方法仍然是在组合内部进行，所以很容易会导致训练集的过拟合问题。还是没有真正解决模型泛化的问题。对于 unseen object 的是被仍然存在者一定的困难。

后续的工作可能会加入一些 Zero-Shot 相关的技术，如引入外部多模态语义数据集进行辅助训练，对于这种映射的方式有一个更好的处理。我们认为组合动作识别这种 Zero-Shot 任务需要基于逻辑推理才能真正解决。

REFERENCES

[1] Girdhar R, Carreira J, Doersch C, et al. Video action transformer network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:244-253.

[2] Herzig R, Ben-Avraham E, Mangalam K, et al. Object-region video transformers[J]. arXiv preprint arXiv:2110.06915,2021.

[3] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.

[4] Neimark D, Bar O, Zohar M, et al. Video transformer network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:3163-3172.

[5] Zhang F Z, Campbell D, Gould S. Spatially conditioned graphs for detecting human-object interactions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:13319-13327.

[6] Damen D, Doughty H, Farinella G M, et al. Scaling egocentric vision: The epic-kitchens dataset[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018:720-736.

[7] Peyre J, Laptev I, Schmid C, et al. Detecting unseen visual relations using analogies[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:1981-1990.

[8] 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. 2019:3585-3594.

[9] Liu Y, Yuan J, Chen C W. Consnet: Learning consistency graph for zero-shot human-object interaction detection[C]//Proceedings of the 28th ACM International Conference on Multimedia.2020:4235-4243.

[10] Materzynska J, Xiao T, Herzig R, et al. Something-else: Compositional action recognition with spatial-temporal interaction networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1049-1059.

[11] Ma J, Damen D. Hand-Object Interaction Reasoning[J]. arXiv preprint arXiv:2201.04906, 2022.

[12] Radevski G, Moens M F, Tuytelaars T. Revisiting spatio-temporal layouts for compositional action recognition[J]. arXiv preprint arXiv:2111.01936, 2021.

[13] Liu X, Li Y L, Lu C. Highlighting Object Category Immunity for the Generalization of Human-Object Interaction Detection[J]. arXiv preprint arXiv:2202.09492, 2022.

[14] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.

[15] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[16] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016:20-36.

[17] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:7083-7093.

[18] Wang X, Gupta A. Videos as space-time region graphs[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 399-417.

[19] Materzynska J, Xiao T, Herzig R, et al. Something-else: Compositional action recognition with spatial-temporal interaction networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1049-1059.

[20] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.

[21] Patrick M, Campbell D, Asano Y, et al. Keeping your eye on the bal l: Trajectory attention in video transformers[J]. Advances in Neural In formation Processing Systems, 2021.

[22] Pourpanah F, Abdar M, Luo Y, et al. A review of generalized zero-sh ot learning methods[J]. arXiv preprint arXiv:2011.08641,2020.

[23] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time objec t detection with region proposal networks[J]. Advances in neural infor mation processing systems, 2015,28.

[24] Herzig R, Ben-Avraham E, Mangalam K, et al. Object-region video tr ansformers[J]. arXiv preprint arXiv:2110.06915,2021.

[25] Radevski G, Moens M F, Tuytelaars T. Revisiting spatio-temporal lay outs for compositional action recognition[J]. arXiv preprint arXiv:211 1.01936, 2021.

[26] Liu Y, Yuan J, Chen C W. Consnet: Learning consistency graph for z ero-shot human-object interaction detection[C]//Proceedings of the 28 th ACM International Conference on Multimedia.2020:4235-4243.