Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization

用于时空动作定位的Actor-context-actor（感觉是一种将图片中的人物与上下文环境相联系的一个动作识别） 关系网

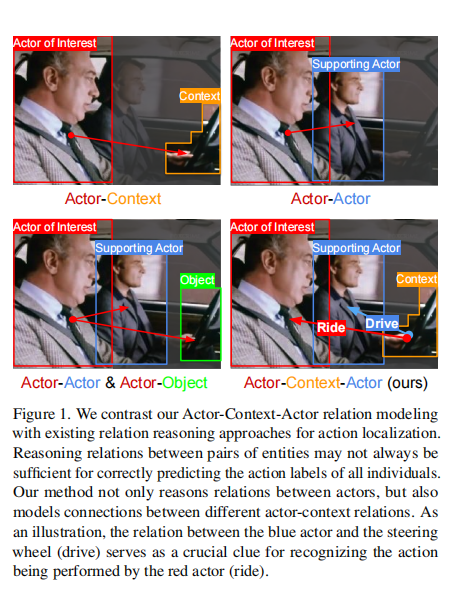
摘要：

对于高阶视频理解来说，从视频中定位人物以及认识理解他们的动作是一项具有挑战性的任务。最近的进展是建立实体之间直接的成对联系来实现的。在这篇文章中，我们迈了更远的一步，不仅是模型直接的承兑联系，而且还考虑到在各种元素上建立的间接更高等级的联系。我们打算将这个Actor-Context-Actor 关系网进行显式建模，这个关系网就是这两个参与者与他们周围环境进行交互的关系。为此。我们设计了这么一个ACAR-N(这个网)，它构建了一个Reasoning Opreator and an Actor-Context Feature Bank去启用间接关系,这个东西是基于新奇的高阶关系推理算法和一个Actor-Context Feature Bank（参与者-上下文特征库）的。在AVA（时空行为检测数据集，就是检测出视频中所有人的位置以及对应的行为类别，要标记的内容包括人物、每个人的行为类别，同一个人同一时间可能会有多个行为）和UCF101-24数据集（具有101个动作类别的真是动作视频的动作识别数据集）的一些实验说明了Actor-context-actor关系建模的一些优势，以及热力图的可视化能够更深入地验证我们的模型是能够找到有关联的高阶关系去支持动作检测。值得注意的是，我们的方法在2020ActicityNet挑战赛（是视频动作识别领域的重要竞赛，涵盖了视频动作分类，时序与时空动作定位，视频时间理解等多个方向，在国际会议CVPR上举办） AVA-Kinetics动作定位任务中以显著的优势排名第一。该代码可以在网上找到。

摘要总结：该模型（包括新的高阶关系推理算法和一些特征库）可以运用图中参与人与上下文环境的一个间接关系来做一个时空行为检测。

1. 介绍

时空动作定位需要定位人以及识别它们在视频中的动作，这个任务在近些年越来越受关注（然后列举了一些文献，要是想做这个领域就去看看呗，做个文献综述）。不像目标检测那样可以仅仅通过观察视觉表现就完成，活动识别通常需要参与者与周围环境的联系来推理，而这些参与者与环境的联系包括了环境、其他参与者以及目标对象。以图一为例子，为了去识别红色框中这个人的“搭车”动作，我们需要去观察他是在一辆车里，然后还有一个司机在他旁边。因此，近些年，联系模型的成功推动了在时空动作检测中的大部分进步发展。这些方法关注于根据实体间的成对联系建立关系船模型。



上面那句话的大概意思就是我们可以通过判别旁边那个人与方向盘之间的关系（即开车）来得到红框中那个在乘车的动作。

然而，各元素之间的关系并不总是被规划为成对的；通常，高阶的联系会为精准的动作识别提供至关重要的线索。在图一中，我们很难仅仅通过红框人旁边的篮筐里的人或者方向盘来推断出他的动作。为了去识别推断出红框人的搭车动作，必须得从篮筐人以及他的动作内容来推断。换句话来说，捕捉图中两个参与者之间的隐含联系是必要的。而这种隐含联系需要基于它们各自与上下文环境的显式关系。

有一些过去的研究工作运用图神经网络去隐式建模参与者与相关目标对象之间的高阶联系。然而，在这些方法需要一个额外的预训练目标检测，并且只有已经定位了的目标能够用来当作上下文场景。由于标识物体的边界框在时空动作定位数据集当中通常是没有提供的，这个预训练目标检测就会被限制在它原来的物体类别并且可能容易丢失场景中的各种其他物体。另外，在这些方法里的高阶关系被限制在只能由有关联的物体推断。这就可能会在动作分类中错过重要的环境和背景线索。

为了解决上面所说的问题，我们提出了ACAR-NET，这个网络关注于建模在ACTOR-CONTEXT-ACTOR关系中的第二层关系（大概就是那层隐式关系的意思吧）。这个网络为动作定位推断出了各个人物角色以及上下文环境之间的隐式关系。ACAR-Net把人物和环境特征作为输出。我们把行为者特征定义为从感兴趣的行为者区域里汇集而来的特征，对于上下文环境特征来说，我们直接使用来自我们主干网络中的时空网格特征图。我们采纳的这个商家问环境并不依赖于任何额外有预定义策略的目标检测，从而使得我们所有的设计变得更加简单灵活。而且，网格特征图此外能够表示各种级别的场景元素（例如实例级别和部件级别）以及类型 （例如背景、物体和物体部分），这对于细粒度的动作区分很有用。ACAR-Net先编码了第一层人物环境关系，然后应用了一个高阶关系推断算法去对第一次关系间进行建模。为了支持不同时间段的人物与环境之间的关系推断，我们建立了一个人物-上下文环境特征库，这个库可以包含在整个视频中的不同时间的人物-背景 关系。

我们在 AVA和UCF两个数据集上做了大量的实验为了去进行时空动作定位。我们提出的ACAR-NET能够在识别人物-对象和人物-人物联系上取得非常显著的提升。定性的可视化说明了我们的方法能够致力于与感兴趣动作有关的那些关系区域。

我们的贡献可以总结为以下三个方面：

1. 我们提出可以对人物-上下文-人物关系进行建模进行时空动作定位。这样的关系总是被过去的实验方法所忽略但是对实现精准的动作定位又非常重要。
2. 我们提出了一种新的 参与者-上下文-参与者关系网络，这个网络可以通过对参与者与上下文之间更高阶的关系显式推理去提高时空动作定位效果。
3. 我们在AVA和UCF数据集上取得了显著的进展。在提交的时间，我们的方法在ACTIVITYNET leaderboard上排名第一。
4. 相关工作：

动作识别。在动作识别上的调查研究工作基本上可以分为三个类别：动作分类，时间定位（包含两个任务，1个是找到感兴趣的时间片段，2.对其进行分类）和时空定位（解决复杂场景的多人多动作问题，包含的工作，1.找出感兴趣的时间片段，2.类别。3.动作发生的空间区域）。早期的工作主要关注于将一段简短视频片段的动作进行分类。3D-CNN;two-stream NETWORK;2D-CNN都是这个任务的三大主要网络模型。当剪辑好的短片段分类取得了进展之后，大量研究就开始朝着解决长的未剪裁视频的方向去了，这个就不仅仅需要对每个动作实例进行识别分类了，还包括去定位这个动作的开始、结束时间。在一维时间中，少量的工作通过扩大目标检测框架把这个问题考虑为一个检测问题。

时空动作定位，目前，时空动作定位的问题已经引起了科研界的关注。类似于AVA数据集的数据集们出现了，这些数据集都为视频中的所有任务以及所有单个动作都打上了标签。它使得动作检测问题到了一个更精细的层面，因为动作实例需要在时间上和空间上都进行定位。早期工作运用的典型方法都在3的CNN上应用了R-CNN检测器。WU发现在原视频上剪裁出参与者所在的区域后，运行3D-CNN主干网络达到的效果比ROI-POOLED 池化参与者特征更好，也就是有更好的空间细节。不过，它还是有一定的局限性，比如说它的计算开销和推断时长总是会随着参与者数量的增多而线性增长。一些最近的工作已经开发出了图结构网络去利用那些上下文信息。

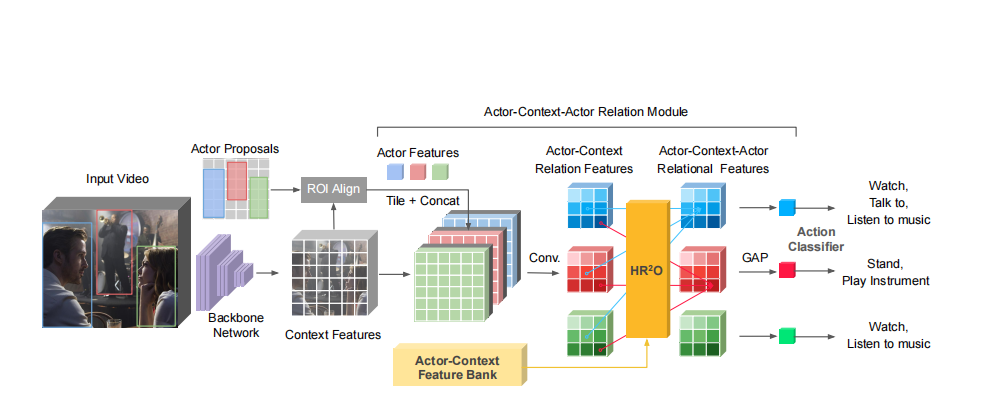
用关系推断理解视频。关系推理已经在视频理解领域内进行了研究。这是自然的因为要识别一个人的动作就要一句与他与其他人物或者对象之间的关系。ZHou为了建模在时间上的视频框架之间的关系扩展了关系网络。未定位的网络利用了self-attention机制去捕捉不同实体之间的一系列依赖关系。WANG说明了有着时空范围图的视频能够提高动作分类的精确度。在时空定位的内容中，有着许多传统的方法，这些方法致力于捕捉视频当中的时空关系，对于在这些方法上的神经网络来说，SUN等人提出了以人物为中心的关系网络，这个网络它可以整合参与者以及背景特征。Girdhar 重新提出了转换网络去编码每两个提出的参与者之间的成对关系。同时，Wu 使用了LFB（long-ter feature banks 长时间特征库？）去提供达到60s的时间支持信息，用这些信息去计算任务见的一个长时间跨度关系。Zhang提出了显式建模参与者们与对象之间的关系。然而，他们的方法分开关注于人物-对象或者人物-人物关系之间的建模。当推断一个任务的动作时，其他人与他相关联的一些环境目标之间的关系都被互虐了。换句话说，它们没有显式建模actor-context-actor关系。对比起来，我们的方法强调了对这些高层关系进行建模。也许我们的工作与38文献中地工作很相似，他们地成果把各堆叠起来的单元都整合起来类似于变换网络。尽管如此，这个方法也是支持ACTOR-CONTEXT-ACTOR 关系的，它把目标检测当成一个背景内容，这就需要额外的在固定对象类别中的预处理目标检测，并且胡怒了气筒重要的上下文环境类型(比如说背景，未定义在预处理类别当中的飙，还有一些目标的具体部位)

1. 方法

在这一部分中，我们提供了ACAR-NET的详细细节描述，我们的ACAR-NET的目标是高效建模并且利用以人物-人物和人物-背景关系为基础的高阶关系去使得动作定位更加精确。

所有的框架：

我们首先介绍了这个动作定位的所有框架，提出的ACAR建模是最关键的模块。这个框架用来检测一段输入视频段中的所有任务并且估计他们的动作标签。像图二中那样，



一段视频首先经过Backbone Network的处理，产生了一系列的时空上下文特征。我们把要研究的那些人物标出来，得到一些人物边界框，使用那些环境特征来提取出人物的一些特征。然后在将这些人物和背景环境的特征整合得到各个人物与上下文环境的一些关系特征，最后利用ACAR-NET计算每两个人物之间的二阶关系。比如说，在这个视频中，

像一些SOTA方法，这个框架以一个现成的人物检测器（Faster R-CNN）和一个backbone network（I3D）为基础建立。人物和背景特征都使用提出的用长时ACTOR-CONTEXT feature Bank 的ACAR模块来处理，并用于最终的动作预测。

详细说就是，这个人物探测器对这个输入的视频片段中的关键帧一阵操作，然后得到了N个检测到的演员。这个检测框框就被复制到与这些关键帧相邻的别的帧里去，同时，这个Backbone network 从输入的视频片段中提取出时空特征量。然后执行一个沿着时间维度平均池化操作去节省一些相应的计算开销，以此得到了一个特征图，这个C、H、W分别代表通道数、高度、宽度。我们应用了RoIAlign 7x7的空间输入，对N个演员的特征最大池化，得到一系列N actor features. 每一个都描述了每个感兴趣区域的时空表现以及运动。

这个网络做了以下两个主要工作：

一阶 人物-背景关系编码。我们使用了这个ACRN网络作为一个模块去编码一阶的人物-上下文关系，这个一阶关系是通过结合ROI A1,A2, ---- An(就是那些检测出来的actor)以及上下文 X来得到的.更详细地说，就是不断地去复制连接每个人物特征与上下文特征X中的所有空间位置（长×宽个），由此得到一系列链接好的feature map，这个map的维度为2C\*H\*W.

再经过卷积操作得到后面的这个feature map Fi.x,y。它对一阶的人物-背景关系进行了编码，然后依据这个关系，去更深入整合了高阶关系，啥叫高阶关系，就是那些隐藏的、没有直接交互的、容易被忽略需要用一阶关系去推理的关系

高阶关系推理：之前得到了这个Fi,x,y。这个Fi,x,y记录了人物和环境信息之间的一阶关系并且对应了相应的空间位置（x,y）。现在就需要对这些一阶关系进行建模得到一些高阶关系以此获取更多的图像信息。然而，因为一阶关系feature太多了。它们可能的成对关系太大了。于是就设计了一个HR2O（高阶关系推理运算）,这个运算就是想要得到再相同空间位置上成对一阶关系之间的所有高阶关系。也就是，两个参与者i和j可以通过相同的空间上下文信息联系起来，即，

然后这个推理操作是使用了ACFB去得到的。这个库包含了在不同时间里的人物-上下文关系，并且可以提供更多完整的时空上下文信息