通过中心引导的判别学习进行弱监督视频异常检测

# 

# Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

**行动识别，何去何从？新模型和动力学数据集**

## 1 主要贡献

**(1)提出了一个新的大型视频行为识别数据集 "Kinetics Human Action Video dataset"**

现有的数据集，如UCF-101和HMDB-51的视频数量都比较少，很多模型因此都获得了比较接近的效果，没法有效的对模型性能进行评价。

因此作者提出了新的数据集，这个数据集比UCF-101和HMDB-51大了两个数量级。包括400个人类动作的类别，每个类别都有超过400个clips。数据集是从Youtube中搜集的真实场景下的视频。

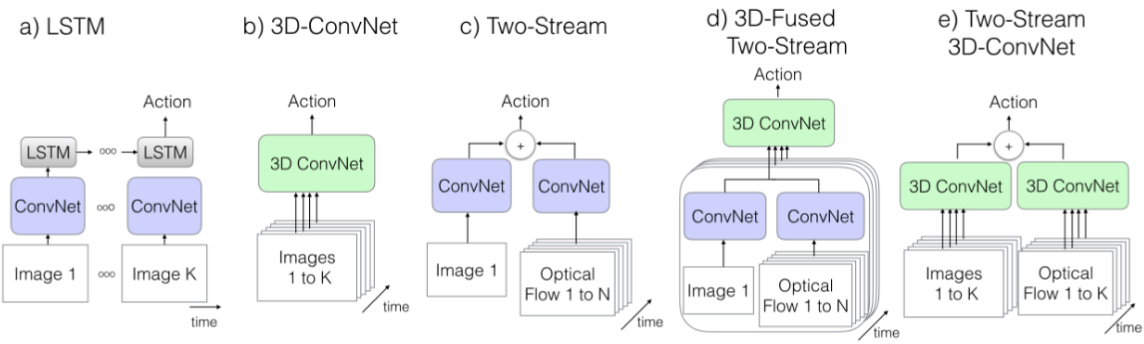
**(2)在2D图像识别模型上增加时间维度(Temporal)，满足视频行为识别任务**

对于视频行为识别任务来说，相较于图像分类等任务，多了一个**时间维度**。而[Imagenet挑战赛](https://github.com/apachecn/ml-mastery-zh-pt2/blob/master/docs/cv/introduction-to-the-imagenet-large-scale-visual-recognition-challenge-ilsvrc.md)中的各种2D卷积网络没有办法同时提取时空信息，因此作者提出**Two-Stream Inflated 3D Convnets(I3D)**模型

## 2 主要方法

### 2.1 benchmark

作者对以往的识别模型进行基准测试，并且取其精华得出I3D网络



#### 2.1.1 **2D CNN + LSTM**

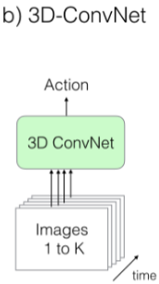
使用2D CNN的好处是可以直接从Imagenet的预训练权重迁移过来，并且使用LSTM提取时序特征。

其基本流程是：首先使用2D网络分别提取每一帧图像的特征，然后对于所有特征使用LSTM来获取帧之间的时序信息。

**优点**：是可以直接使用2D网络的预训练权重；

**缺点**：只有在LSTM的部分才能够基于高维抽象的特征进行运动信息的提取，因此损失了很多低等级的运动信息，此外LSTM训练时需要在每个step都进行反向传播。

#### 2.1.2 **3D CNN**



**优点**：可以同时提取时空信息；

**缺点**：也显而易见，即具有更庞大的参数数量从而更难训练，因此一般3D网络的深度都较浅，但这样又影响了模型的表达能力，此外，没有办法能够有效的把2D网络的预训练权重迁移到3D网络。

本文作者在C3D的基础上进行了改进：（1）在所有卷积和全连接层之后都假如了BN，（2）原始C3D的第一个池化步长是 1×2×2 ，本文改为了 2×2×2 ，因此能够更加节省显存占用（可以使用更大的Batch，对于BN来说这也是非常重要的）。

#### 2.1.3 **Two-Stream**

双流网络通过提取光流提取时序信息，避免了LSTM进行时空模拟和3DCov进行时空学习

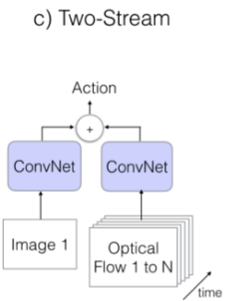
**[光流（oprical flow）](#optical_flow)**

**光流**是指视频中连续两帧图像中物体运动所引起的位移变化。光流是视频中物体运动的一个重要特征，可以用来描述物体在空间中的运动轨迹和速度。

**光流法**是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系，从而计算出相邻帧之间物体的运动信息的一种方法。







双流网络中:

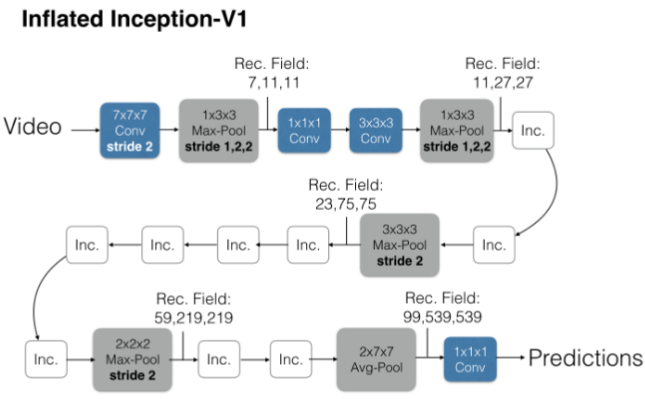
* 一个分支作为空间流，输入RGB图像来提取物体和场景外观特征；
* 一个分支作为时间流，输入光流来提取运动特征作为时序。

通常两个网络分开训练，只有在测试时才会平均两个网络的预测。因此，也产生了时空信息在浅层网络中无法有效融合的问题。

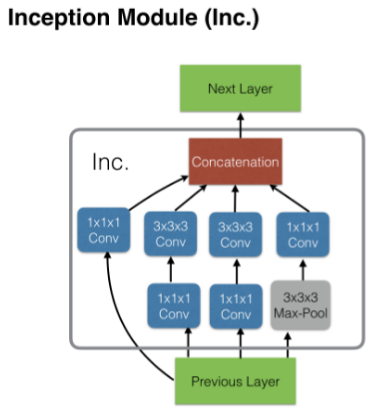
### 2.2 Two-Stream Inflated 3D ConvNets （I3D）

**双流膨胀3D卷积网络**

作者使用[Inception-V1](%20获得2014年ImageNet分类竞赛的冠军，主要创新点在于同时拓展网络的宽度和深度（设计了Inception结构），网络结构中采用了辅助分类器%20)作为骨干网络（后面更改为ResNet），其网络结构如下：



**其中Inc.模块的具体模型图**



#### 2.2.1 2D到3D

由于Inception模型是针对图片的2D模型，所以变为3D版本要干两件事情：

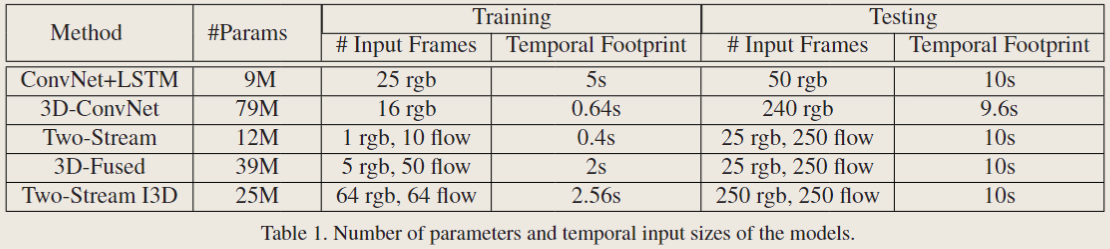
* 模型从2D变成3D
* 权重从2D变为3D

文章中，作者直接把 2D卷积或2D池化的 N×N的kernel变成 N×N×N的kernel。

对于预训练权重，首先把2D kernal在时间维度上复制N份，然后除以表示时间维度N，这样做是为了扩展到3D卷积之后，每一层都仍然获取到类似大小的输出相应。

## 4.evaluation

首先评估模型参数和输入时序间隔。可以看出3DConv参数非常大，而采用双流网络的模型要更小。



然后是在不同数据集上的评估。

* 首先可以看到，提取双流(RGB+Optical Flow)特征都提升了模型的分类
* 作者提出的I3D网络表现更出色

