Slowfast论文阅读

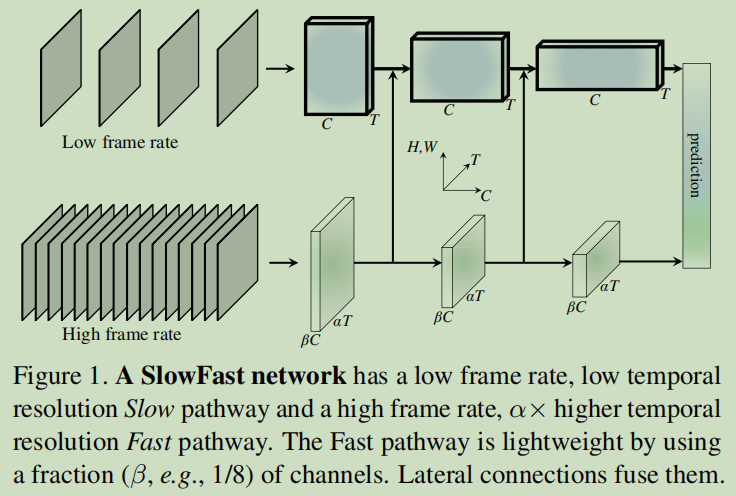
1. 要解决什么问题

提出了视频行为识别和检测的新思路和方法。在无预训练的情况下，****Kinetics****数据集上视频分类准确率达到了****79.0%****，在AVA action detection数据集上达到了当时最好的****28.3mAP****。

1. 用了什么方法

使用了双路径的模型。算是一个以两种不同的帧率运行的单流结构.

其中一条路径被设计用来捕获图像或少量稀疏帧所提供的语义信息，它以较低的帧速率和较低的刷新速度运行。相反，另一条路径负责捕捉快速变化的运动，它以快速刷新速度和高时间分辨率运行。



Slow路径：

       Slow pathway可以是任意一个将视频片段作为时空立方体输入的卷积模型.

      1）低帧率：slow路径输入帧数为T，在视频中每隔t帧取一帧图像(采样率为t)，t取16。即视频的总帧数为：t×T。

      2）高通道数：slow的图像帧数比fast路径少，但特征的通道数比fast路径多。

Fast路径：

 1）高帧率

        Fast路径采样αT帧作为输入，比Slow途径密集α倍。它的时间采样率为 τ/α，其中α>1，本文中α=8。

2）高时间分辨率特征

        整个Fast路径中均不使用时间下采样层（既不使用时间池化也不使用时间步长的卷积操作），特征张量在时间维度上总是具有αT帧，尽可能地保持时间保真度。

3）低通道容量

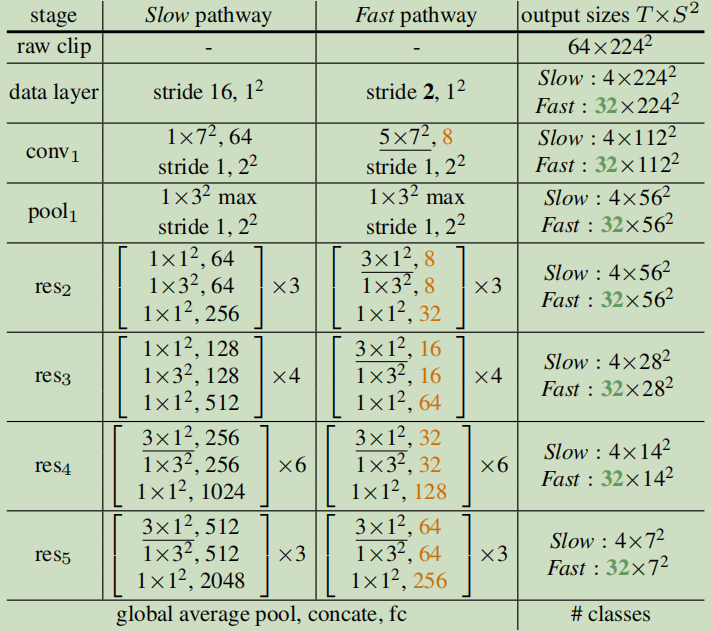
        Fast路径的通道数是Slow路径的β（β<1）倍。在文中β=1/8。

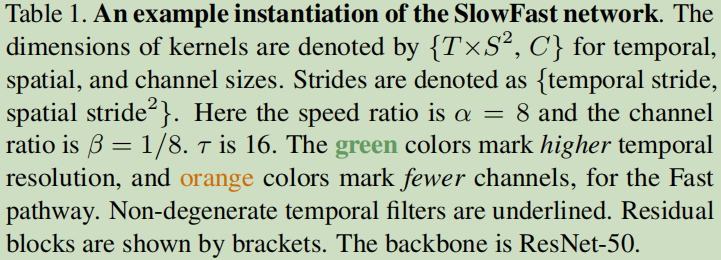
横向连接：

我们在每个阶段“stage”的两个路径之间都附加一条横向连接，特别的，如果是ResNets，这些横向连接紧跟在pool1,res2,res3和res4之后。这两个路径的时间维度不同，所以横向连接会先进行转换来匹配它们。我们使用单向连接来将Fast路径的特征融合到Slow路径中（做了双向连接实验，效果类似）。

最后，对两个路径的输出进行全局平均池化global average pooling，再将两个池化后的特征向量拼接在一起输入到全连接分类器中。

网络实例化：





用T×S²来表示时空尺寸，其中T是时间长度，S是方形空间裁剪的宽高。

步长为：{时间步幅，空间步幅2}。

速率比是α=8，通道比是β=1/8。τ是16。

Fast通道中的绿色表示高一些的时序分辨率，橙色表示较少的通道数。

三种横向连接方式：

横向连接将Fast的特征融合到Slow路径，不过在融合前需要进行特征尺寸的转换匹配（只能有一个维度尺寸不同）。将Slow路径的特征形状定义为{T,S²,C}，Fast路径的特征形状为{αT,S²,βC}。在横向连接中我们尝试了如下的转换方法：

（i）Time-to-Channel：我们把Fast的特征reshape成{T，S²，αβC}，也就是把所有帧的特征全部放到一帧的通道上。

（ii）Time-strided sampling：我们直接简单在Fast特征的时间维度上进行采样（每α帧采样1帧），即将Fast特征采样成{T，S²，βC}。

（iii）Time-strided convolution：用一个5×1²的卷积核进行3d卷积， 输出通道数为2βC(空间维度扩张2倍)，时序维度的步长= α(时序维度压缩a倍)。

最后，Fast路径横向连接的输出被融合到Slow路径中（求和或者拼接）。

1. 效果如何

行为识别上：

**Kinetics-400**。表2比较了我们的模型和SOTA方法的结果，我们的SlowFast模型分别使用了不同的输入采样量（T×τ）和不同的backbones（Res50/101[24]， Nonlocal[56]）。

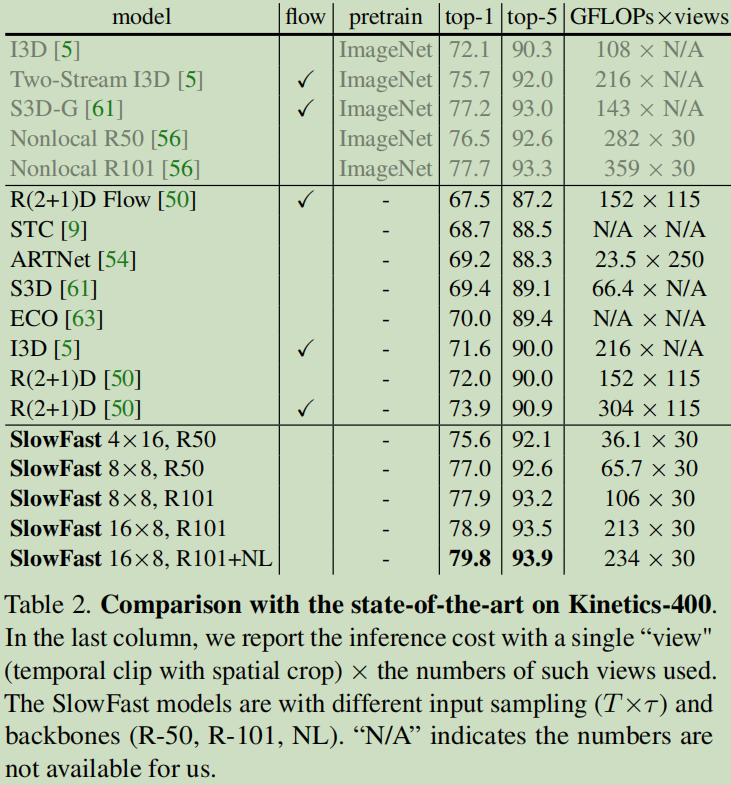
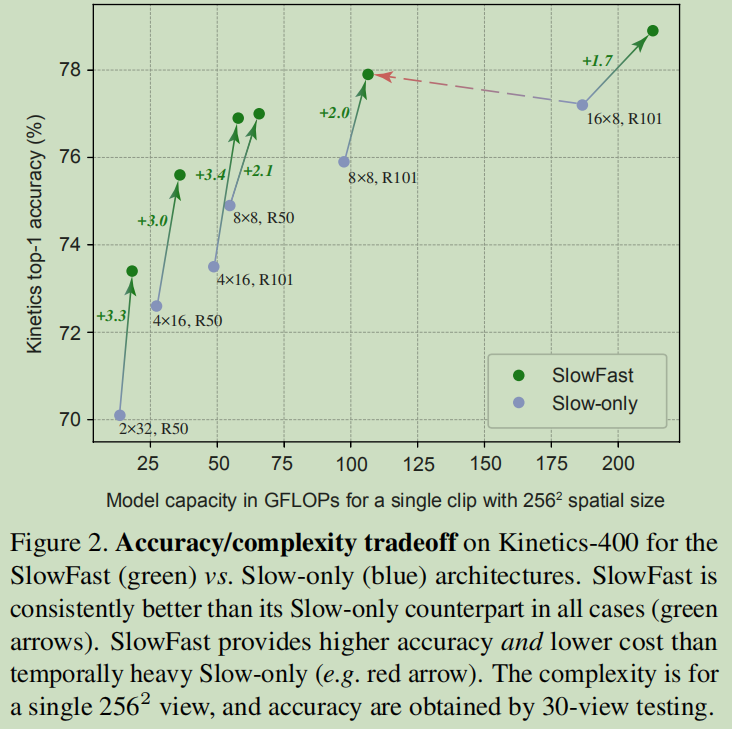
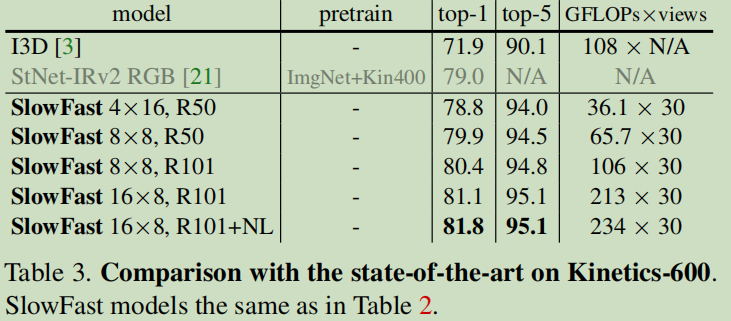
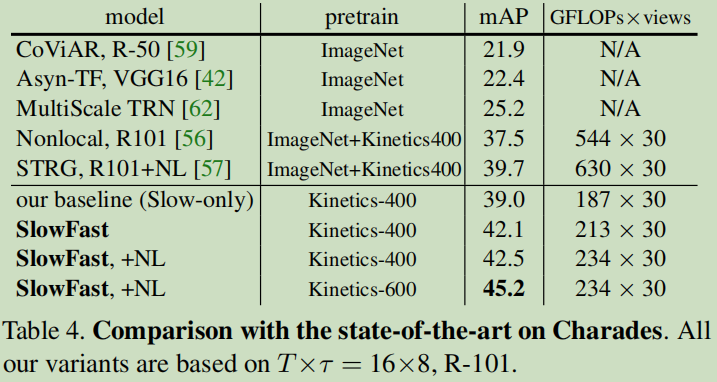


图2表示在Kinetics-400上slowfast和slowonly精度和速度的权衡。

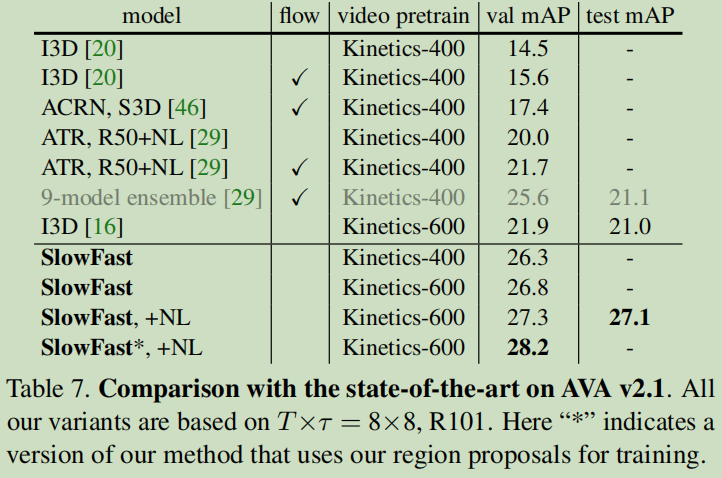


**Kinetics-600**。这个数据集是一个相对比较新的，现有的结果也不够好的数据集，所以我们的目标主要是提供一些可供将来参考的结果，如表3所示。

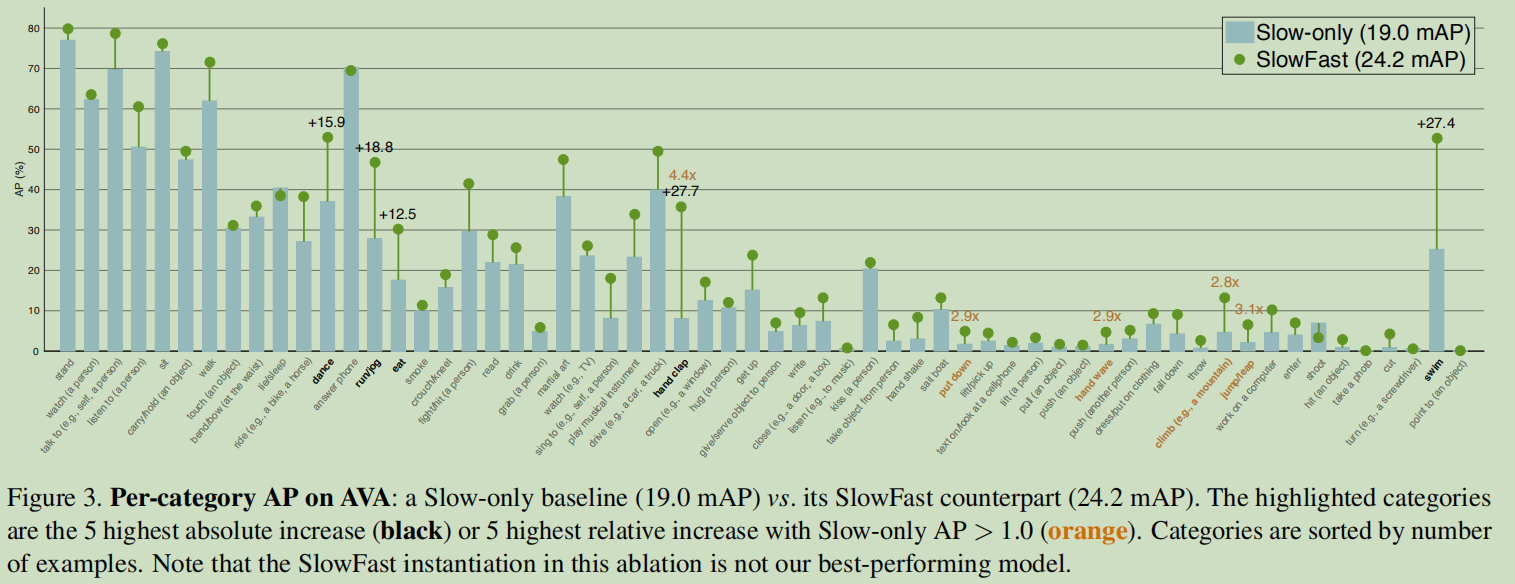
**Charades**。这个数据集的动作持续时间更长，表4展示了我们的SlowFast在该数据集上的结果。

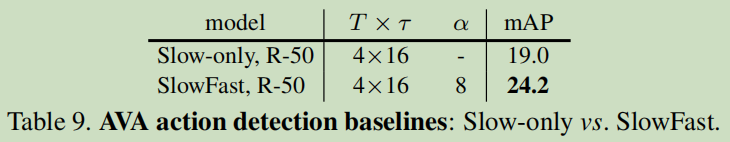


在ava行为检测上：



增加了光流的网络还是比不过slowfast的baseline。





从分类的角度看，slowfast比slowonly提升了60个类别中的57个类别。SlowFast模型只在3个类别中更差：“接听电话”（-0.1 AP）、“撒谎/睡觉”（-0.2 AP）、“射击”（-0.4 AP），与其他类别相比，它们的下降幅度相对较小。

1. 还存在什么问题和可借鉴的地方

Slowfast充分利用时间轴，使用slow分支和fast分支，取得了比rgb加光流分支更好的效果。

SlowFast模型其主要的工作流程大致如下所示：

step1：用快慢两种速率采样输入视频

step2：采样后的视频帧对应输入到Slow/Fast两个分路

step3：Slow分路使用ResNet 2D Conv + 3D Conv提取视频空间语义特征，size {T,S²,C}

step3：同时Fast分支使用ResNet 3D Conv提取视频时域运动特征，size{αT,S²,βC},其中α>1,β<1

step4：横向连接统一两个分路的特征

step5：Softmax进行分类

问题和不足：

1、在行为检测中，人员位置检测和行为检测还是分开训练的。是两个模型。

2、使用关键帧上的2D ResNet-50模型输出作为区域建议网络的输入，通过沿时间轴方向复制每个帧上的2D RoI，将其扩展为3D RoI。通过RoIAlign在空间上计算RoI特征，并在时间上计算全局平均池。最后，对RoI特征进行max-pooled，输入到每个类基于sigmoid的分类器中进行多标签预测。