Temporal Segment Networks时域分割网络

1. 要解决什么问题

深度学习在静止的图像上取得了巨大的成功，但是在视频行为识别上，却没那么有用。

在本文之前主流的方法都是着重于提取图像空间信息以及短期的运动信息，不能很好地处理长期时间依赖问题（long-range temporal structure），所以本文需要解决这个问题。然后考虑到密集采样会造成很多的冗余计算，本文采用了稀疏时间采样策略。

1. 用了什么方法

提出了一个新的基于视频的动作识别的框架（TSN）。基于长范围的时间结构建模，结合了稀疏时间采样策略和视频级监督，保证了整段视频学习时的有效和高效。

TSN使用从整个视频中稀疏地采样一系列短片段，每个片段都将给出其本身对于行为类别的初步预测，从这些片段的“共识”来得到视频级的预测结果。在学习过程中，通过迭代更新模型参数来优化视频级预测的损失值（loss value）。

下面是网络结构图

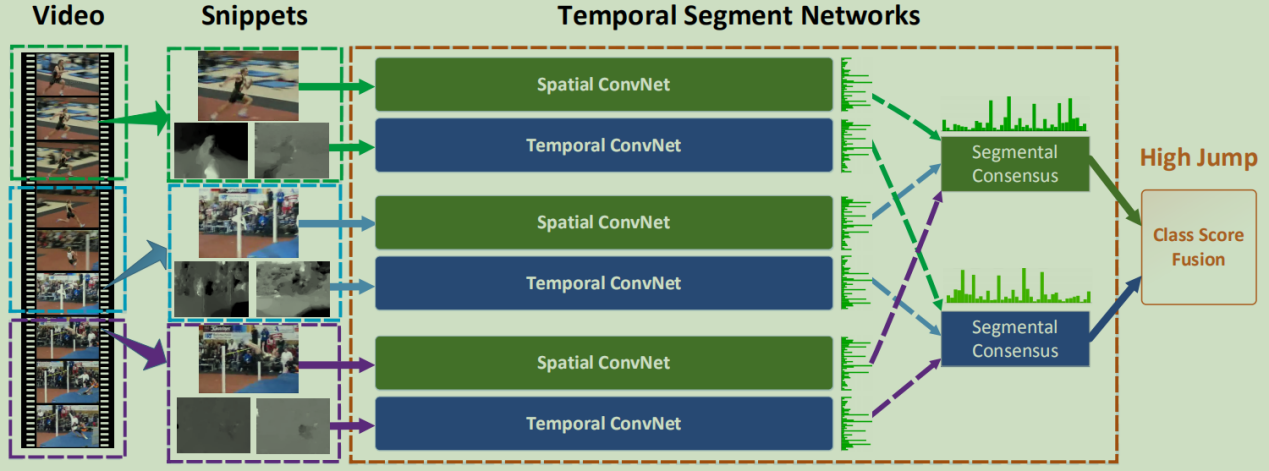


图1.时间分段网络：一个输入的视频分为K个片段（segments），从每个分隔片段中随机选取一小段（snippet）。不同小段的类别得分使用分隔共识函数来融合生成分隔共识，这是一个视频级的预测。所有模式的预测融合产生最终结果。在所有小段中卷积网络共享参数。

给定视频V，我们将其分为相等持续时间的K个段{S1，S2，…，SK}。 TSN对一系列小段进行建模：

TSN(T1, T2, · · · , TK) = H(G(F(T1;W), F(T2;W), · · · , F(TK;W)))

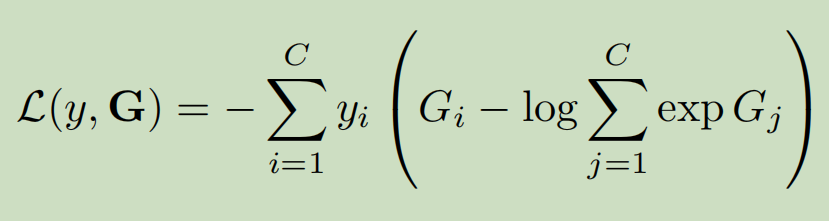
(T1, T2, · · · , TK)是一系列小段。每个小段TK从对应的片段SK中随机采样。

F(TK;W)表示参数W的卷积网络，作用在小段TK上，为所有类别生成类别得分。

分隔共识函数G组合多个小段的输出获得对于类别的共识。

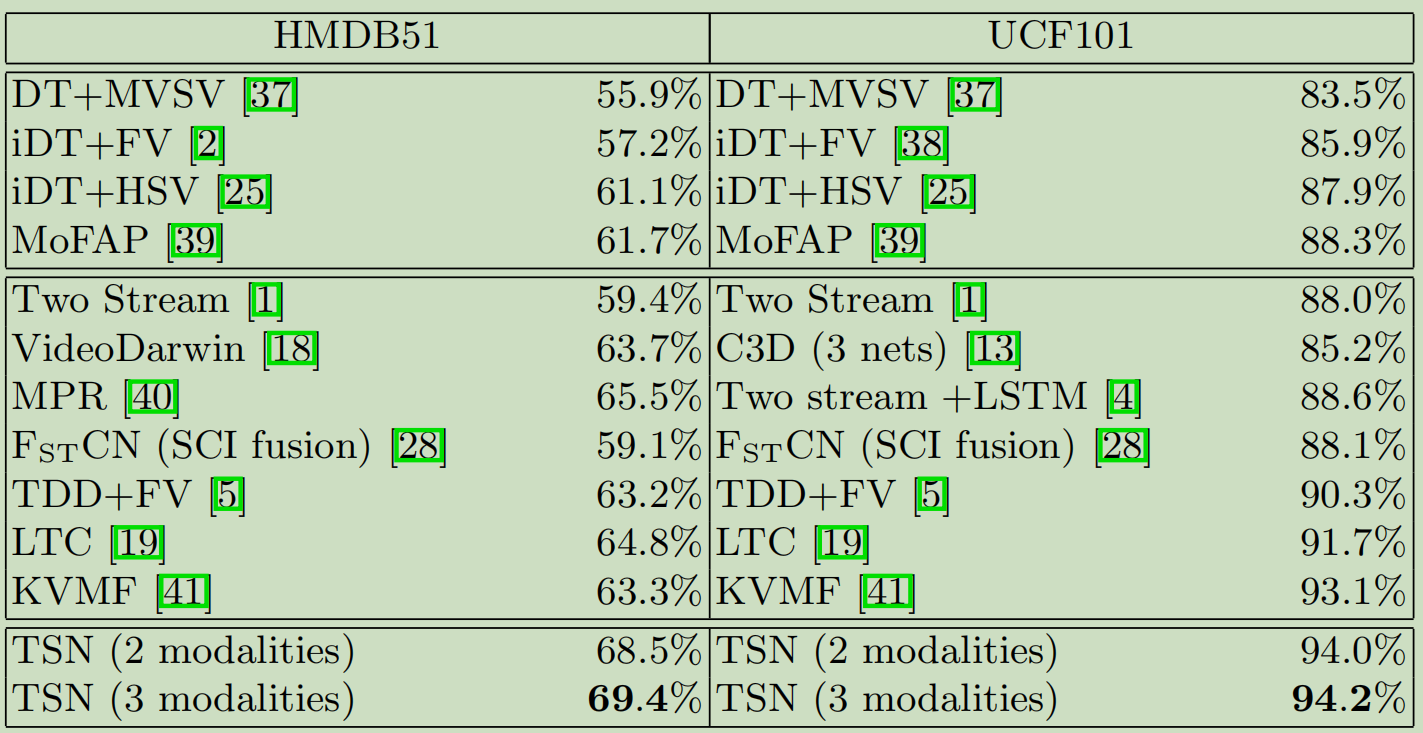
基于该共识，预测函数H预测在整个视频中每个行为动作的概率。这里使用softmax函数作为H。

结合标准的分类交叉熵损失，最终的分段函数损失共识是



其中C是行为类别总数，yi是类别i的真实标签，在实验中，小段的数量K设为3，共识函数g的形式仍没有解决。本文中，使用最简单的g形式，Gi =g(Fi(T1), . . . , Fi(TK))。使用聚合函数g从所有片段相同类别的分数推断出一个类别分数Gi。聚合函数g实验了平均，极值和加权平均。最终采用平均。

1. 效果如何



论文发表时，在两个小数据集上达到了state-of-the-art

1. 还存在什么问题和可借鉴的地方

Tsn是双流网络，要训练两个模型。

光流不能实时。