Two-Stream Convolutional Networks for Action Recognition in Videos（双流网络 2014）

> 这是深度学习视频理解领域的开山之作

Abstract

This article developed a new architecture using a two-stream CNN. To be more specific, they demonstrated that using a multi-frame dense optical flow is well-performed.Besides, they also shows that applying to two different action classification datasets can be used to increase the amount of training data and improve the performance on both.

Introduction

在作者写下这篇文章期间，人体动作识别是个研究热点（2024也是）。相比于静止的图片，视频可以提供有关于动作的更多线索；同时，视频也提供了天然的数据增强，因为同一物体在视频的不同时刻会有不同的形状。

在先前的工作中，人们试图将视频的所有帧一股脑全部塞给CNN网络，但是效果并不好，甚至比不过人工特征。

而作者团队则认为，网络架构应该分别对空间信息和时间信息进行处理，因此他们使用了两个CNN网络分别进行处理，其中时间信息使用了光流处理，最后进行late fusion融合特征。

Method

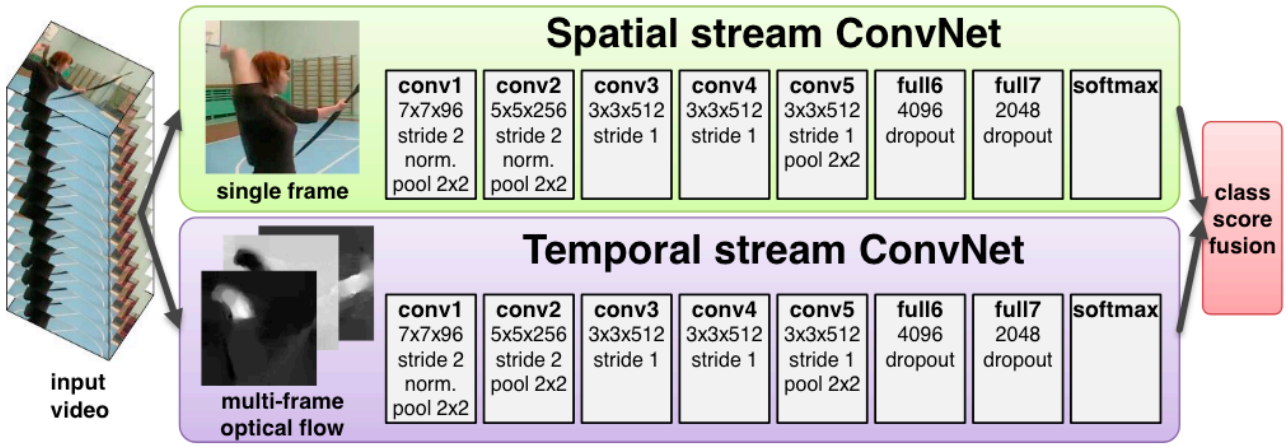


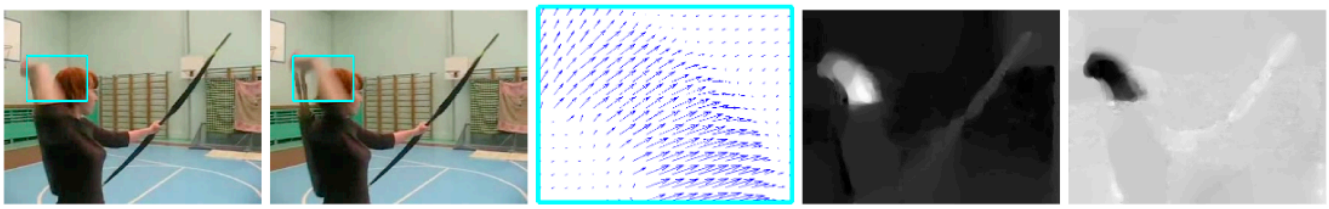
Figure 1: **Two-stream architecture for video classification.**

Two-stream architecture

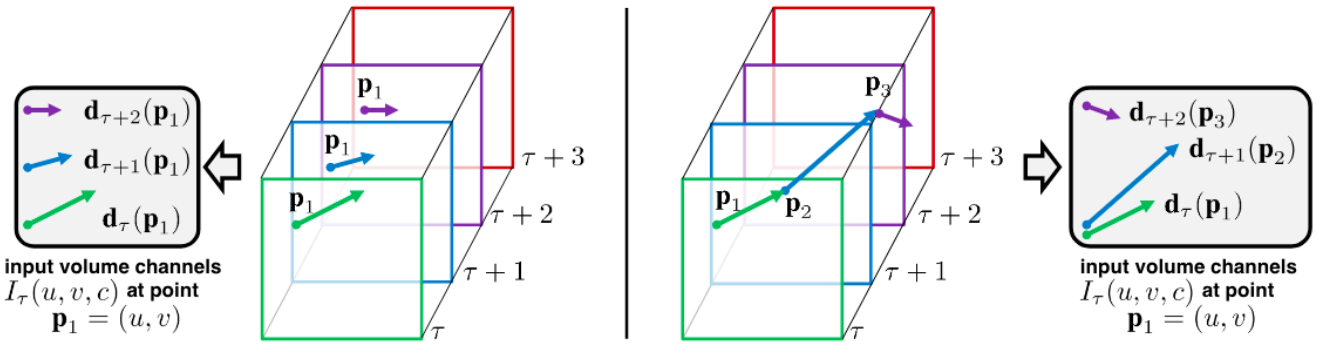
视频很自然的就可以分为空间和时间信息。在时间信息中，包含了物体的外貌特征和场景特征；而时间信息，则包含了物体的运动信息和相机的运动信息。实际上CNN就是分类框架，所以在处理空间信息的那边网络上，可以直接拿来当目标检测用。在使用时，把每一帧（其实是隔几帧）输入，让模型识别就OK。

Optical flow ConNets

在处理时间信息的那边网络上，作者使用了“光流‘这一方法，具体可以看这篇https://zhuanlan.zhihu.com/p/384651830。



作者隔几十帧取连续的11帧图像，（每一帧都做的话数据量太大了）每隔壁的两张做一次光流计算，所以总共得到10张光流图。每张光流图又可以分为x轴和y轴方向上的向量，因此处理时是20张。此外，作者保险起见，从第六帧为界限，前6帧做顺时间计算，后面6帧做逆时间计算。



作者比较过两种光流向量选取，第一种是原位置一直做光流计算；一种是更加合理，上一帧光流向量指向哪里，这一帧就从哪里开始计算，这个看起来更合理。但不知道为什么，后面的方式反而效果不好。