Two-Stream Convolutional Networks for Action Recognition in Videos

(NIPS 2014)

1. 视频领域具有的优势

视频本身就是一个很好的数据来源，相比于单个二维图像可以包含更多信息，比如物体之间移动的信息，还有长期的时序信息，音频信号，适用于多模态学习。针对人类本身而言，人眼看到的事物也都是连续的，以视频的形式展现。

2. 把深度学习运用到视频分类

双流网络不是第一篇，之前的 Deep Video 就已经实现，并且提出了一个特别大的数据集(Sports One Million)，包含 100 万个视频，但是训练出来的结果不如预期，在常用数据集上的测试效果甚至不如之前手工设计的特征。

3. 视频理解的开山之作：双流网络

是第一个能让 CNN 的效果与之前基于最好的手工特征的方法打成平手，让大家意识到深度学习不是不能解决视频理解，而是之前的打开方式不对。自此之后，深度学习的方法基本就在视频理解领域占据主流地位了。

4. 正文

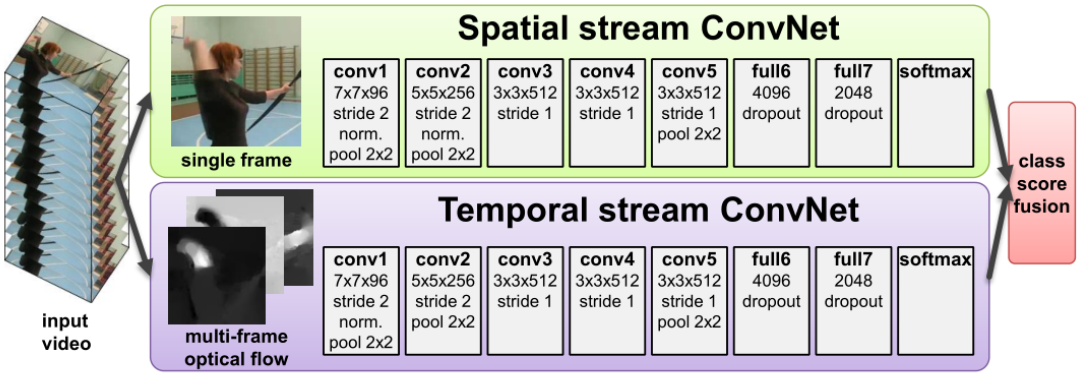
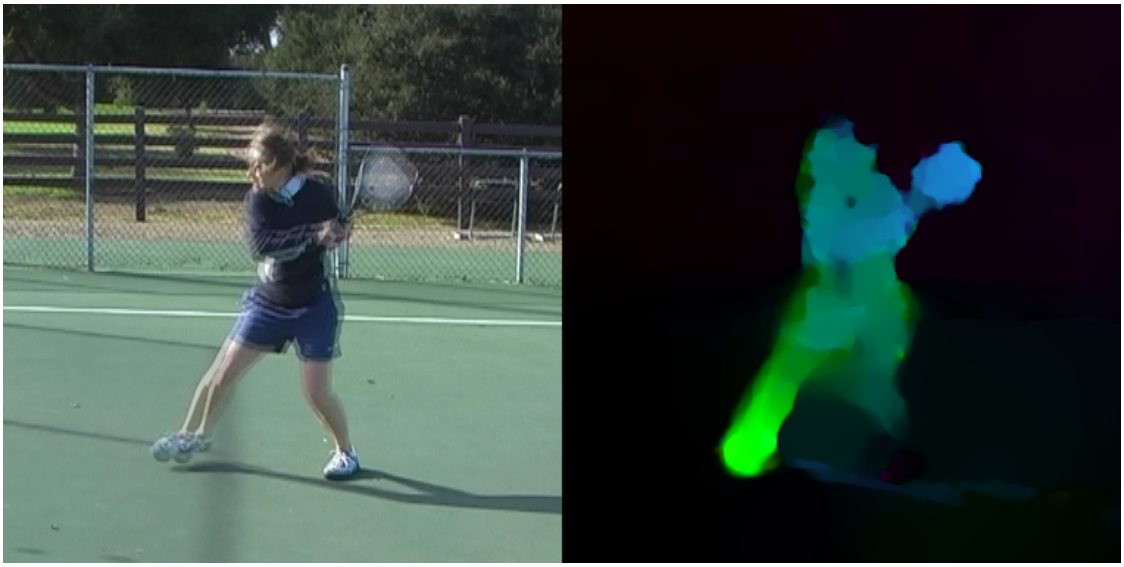


Figure 1: **Two-stream architecture for video classification.**

卷积神经网络比较擅长学习局部的特征，而不擅长视频中物体的移动规律。作者将运动的光流信息提前抽取好，只需要让神经网络学习输入光流到最后的动作分类之间的映射，这恰好是深度卷积神经网络最擅长的部分（通过一系列的矩阵乘法，去学习这种输入输出之间的映射关系）。

Figure 1 中，空间流的输入就是一张单帧的图片，输出分类概率；时间流的输入是一系列的光流图片，输出分类概率；最后，对上述两个分类概率取加权平均得到最终的预测。



光流：是一个非常有效的描述物体之间运动的特征，上图右侧光流捕捉人物运动，运动越明显的地区颜色越亮，代表运动幅度越大。

4.1 Abstract

CNN 用于动作识别的难点：①从静止的图像上获得外部信息，包括物体形状，大小，颜色以及整体的场景信息；②物体的移动信息，可以理解为时序信息。

文章的三点贡献：①提出双流网络：空间流和时间流；②证实了即使在少量的训练数据下，一个直接在光流数据上训练的神经网络也能取得很好的效果；③利用多任务学习，在两个数据集上同时训练一个骨干网络，这样不仅骨干网络可以训练的更好，而且在这两个数据集上都能有效提高性能。

4.2 Introduction

视频数据天生就能提供一个很好的数据增强，一个视频中，同一个物体会经历各种各样的形变，或者遮挡、光照的改变，这种改变是多种多样且十分自然的，比使用生硬的数据增强手段(cutout)好的多。

Late fusion: 指的是在网络最后的 logits 这个层面上去做合并。在深度学习中，logits 就是最终的全连接层的输出。通常神经网络中都是先有 logits，而后通过 sigmoid 函数或者 softmax 函数得到概率。

典型的 fusion 方法有 early fusion 和 late fusion。顾名思义，early fusion 就是在特征上（feature-level）进行融合，进行不同特征的连接（concatenate），输入到一个模型中进行训练；late fusion 指的是在预测分数（score-level）上进行融合，做法就是训练多个模型，每个模型都会有一个预测评分，我们对所有模型的结果进行 fusion，得到最后的预测结果。常见的 late fusion 方法有取分数的平均值（average）、最大值（maximum）、加权平均（weighted average），另外还有采用 Logistics Regression 的方法进行 late fusion。

4.2.1 Related work

视频领域的进展一般由图像领域推动，比如之前针对图像的 SIFT 这个特征，

对应到视频中的 STIP(spatio-temporal interest points)，

传统手工方法的两个方面：①局部的时空学习(spatio-temporal interest points)；②基于光流和轨迹的方法 dense point trajectories (利用视频前后帧中点和点之间联系得到的这些轨迹信息)，改进比较好的是 idt 特征(improved dense trajectory)。

聚焦如何能在深度神经网络中学习到有用的 motion 信息。

4.3 Two-stream architecture for video recognition

双流网络的两种 late fusion 方法：①加权平均；②把得到的 softmax 分数作为特征再去训练一个 SVM 多分类器。

Spatial stream ConvNet: 输入视频帧，相当于从静止的图片中做动作识别，即图像分类任务，跟视频没什么关系。此外，单帧图片作为输入时，空间流的网络可以利用 ImageNet 去做预训练。

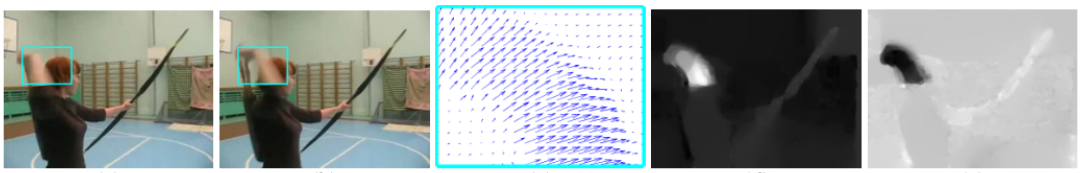


Figure 2: **Optical flow.** (a),(b): a pair of consecutive video frames with the area around a moving hand outlined with a cyan rectangle. (c): a close-up of dense optical flow in the outlined area; (d): horizontal component d^x of the displacement vector field (higher intensity corresponds to positive values, lower intensity to negative values). (e): vertical component d^y. Note how (d) and (e) highlight the moving hand and bow. The input to a ConvNet contains multiple flows (Sect. 3.1).

Optical flow: 两个图片得到一张光流，L 帧视频最终得到 L-1 帧光流图片。本文中使用 11 帧视频，得到 L=10，对应时间网络的输入维度 2L=20（空间网络输入为 RGB 图像，维度为 3），2L 是因为光流分解为水平方向和竖直方向。

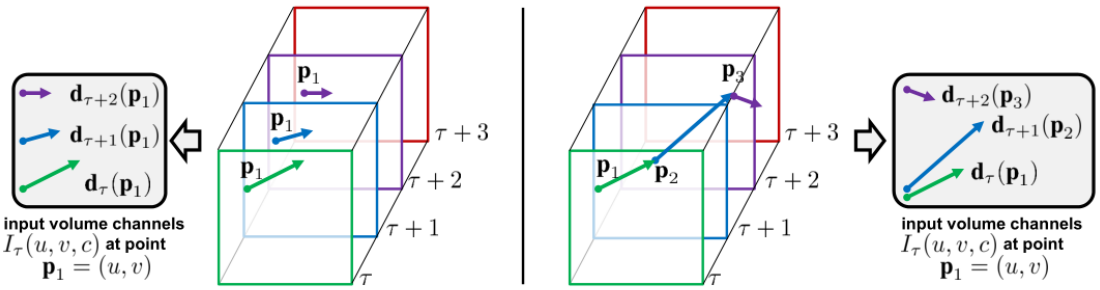


Figure 3: **ConvNet input derivation from the multi-frame optical flow.** *Left:* optical flow stacking (1) samples the displacement vectors **d** at the same location in multiple frames. *Right:* trajectory stacking (2) samples the vectors along the trajectory. The frames and the corresponding displacement vectors are shown with the same colour.

保持时序信息，选用多张光流帧。两种叠加光流帧的方式：①每次都关注同一个点的光流信息；②根据光流的轨迹，进行这种光流数值的叠加。

光流方法的缺点：时间长，存储空间大。作者把密集的光流值 rescale 到[0，255]中的整数，并且存成 JPEG 图片实现压缩。

4.4 Evaluation

Table 1: **Individual ConvNets accuracy on UCF-101 (split 1).**

(a) **Spatial ConvNet.**

Training setting	Dropout ratio	
	0.5	0.9
From scratch	42.5%	52.3%
Pre-trained + fine-tuning	70.8%	**72.8%**
Pre-trained + last layer	**72.7%**	59.9%

(b) **Temporal ConvNet.**

Input configuration	Mean subtraction	
	off	on
Single-frame optical flow ($L = 1$)	-	73.9%
Optical flow stacking (1) ($L = 5$)	-	80.4%
Optical flow stacking (1) ($L = 10$)	79.9%	**81.0%**
Trajectory stacking (2)($L = 10$)	79.6%	80.2%
Optical flow stacking (1)($L = 10$), bi-dir.	-	**81.2%**

Table 4: **Mean accuracy (over three splits) on UCF-101 and HMDB-51.**

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

前三种是最好的手工特征，第 2，3 个是在第 1 个 IDT 基础上增加了全局信息，通过一些 encoding 使得特征更具有全局性，更适合做视频（比如 fisher vector encoding）。第 4，5 是之前利用深度学习做视频分类的效果，比较差。空间流网络因为利用 ImageNet 预训练模型，效果要比第 5 个 Deep Video 好

本文训练测试数据集：UCF-101 and HMDB-51

常见数据集：kinetic 数据集，something something 数据集，两者大概都包含 20 万视频

启发：当单一神经网络表现不好时，人们会想到双流网络。双流网络也可以作为一个多模态学习的先例，RGB 图像和光流图像其实就属于不同的模态，代表的实际意义不同，最后在 logits 层面合并。

CLIP 上面是图片，下面是文本，最后算一个相似性。

代码：https://github.com/woodfrog/ActionRecognition