# Title：

Two-Stream Convolutional Networks for Action Recognition in Videos

双流卷积神经网络来做视频里的动作识别

视频分类，视频理解领域的开山之作，双流网络。

视频本身是一个非常好的数据来源，它比2D的单个图像能包含更多的信息，比如物体之间移动的信息，长期的时序信息，音频信号等，非常适合当今的趋势-多模态学习。而且，作为人类，人眼看到的信号也是连续的，以视频形式展现的，而不是一个又一个图片。所以如何能更好的利用视频数据做视频理解，是有可能通向更强的人工智能的必经之路。故视频理解是个很好的方向，非常有前景。

之前CNN迁移到视频，早在2014年就有工作（Deep vedio），但是效果差强人意。双流网络是第一个能用CNN和之前最好的手工特征的网络打成平手的工作，从此深度学习的方法成为视频领域内的主流地位。

# Abstract：

在本文中，作者研究了如何使用CNN去做视频内的动作识别，主要难点是空间信息和物体之间移动信息。他们想把手工信息的思想带入以数据驱动的深度神经网络之中来。

贡献有三：1.提出了双流网络。2.少量数据情况下，直接在光流信息上学习的网络也能有好效果。3.使用multi-task learning的方法在两个数据集上学习同一个网络。

# Intro：

视频内的时序信息能为识别工作提供重要线索。

使用视频信息天生就是一个很好的数据增强。

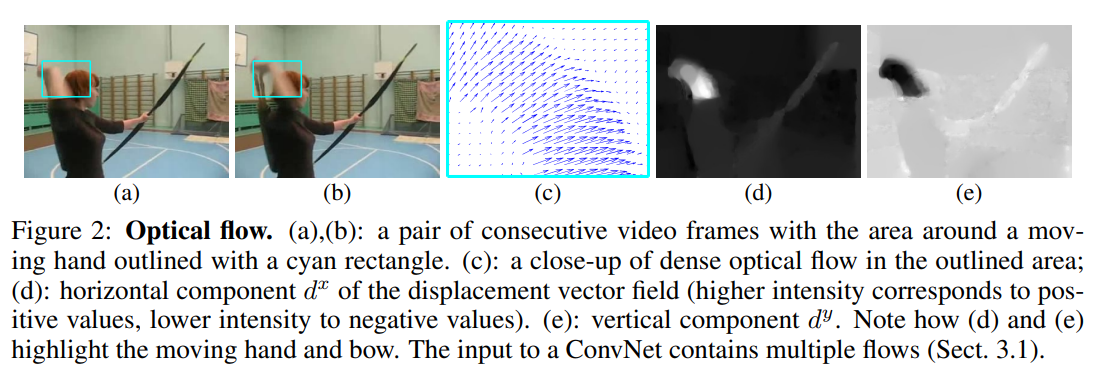
对于空间流网络来说，可以拿ImageNet预训练好的网络做初始化了，时间流可以直接拿光流来作为输入，从而预测动作，卷积神经网络本身不用去学这种motion information了，所以这个任务也变得简单了。

# Two-stream architecture for video recognition

早期工作是把视频抽出关键帧，然后把帧通过CNN，最后把结果合并起来，要么是把帧叠起来当作整体输入给网络，然后在网络里做一些Early Fusion或late Fusion，达到时空学习的效果，这些效果不好，比不上手工设计的特征。

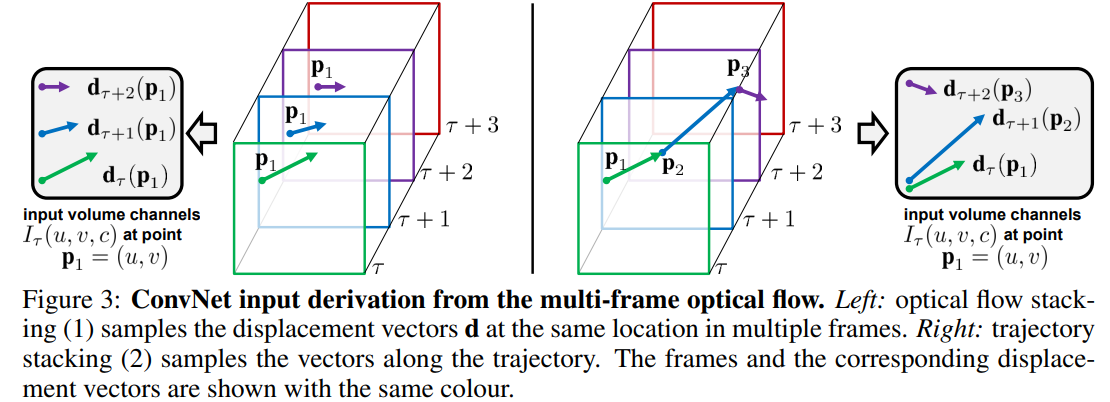
作者发现CNN不擅长学习动作信息（它擅长抽取局部信息），所以作者帮助网络把动作信息抽取好，就是光流抽取，只需要让神经网络学输入光流到动作分类映射就好了（深度学习擅长通过一系列矩阵乘法学习互相映射关系）。作者把双流神经网络分为：空间流神经网络（上）（AlexNet）和时间流神经网络（下）（AlexNet输入channel改成2L，水平在前叠加完然后叠加竖直）。前者输入一张图片，输出分类的概率，后者输入动作光流信息，输出分类概率，最后将二者加权平均，作为最后的输出（属于late fusion）。

# Optical flow ConvNets（光流）

前两张图是视频的两帧，图c是前两帧的光流的可视化，当用数学表示光流的时候，人们把它拆成了两个方向，分别是图d和图e，表示水平方向和竖直方向上的位移。每两张图得到一个光流，维度为H\*W\*2，两个维度分别为竖直和水平。通向神经网络的光流图是多个光流图叠加在一起的，如何叠加呢？

## ConvNet input configurations

**Optical flow stacking.**

一种是直接叠加在一起，另一种是在轨迹上进行数值的叠加。

左图为直接叠加，每一次问P1这个点下一帧往哪走。没有充分利用光流的信息。

右图是P1在上一帧移动到了P2，那就在P2寻找这个点的信息。对光流的利用好了很多。

但是最后实验的结果表面左图比右图好。

**Bi-directional optical flow.**

双向光流，BERT有用，涨点小trick。

计算了前向光流和后向光流，选取L帧视频，前半段算前向光流，后半段算后向光流，最后输入维度为H\*W\*2L。

# Implementation details

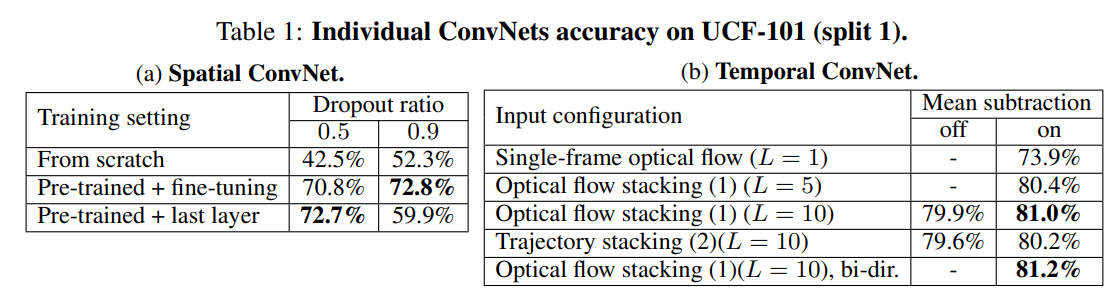
## Testing.

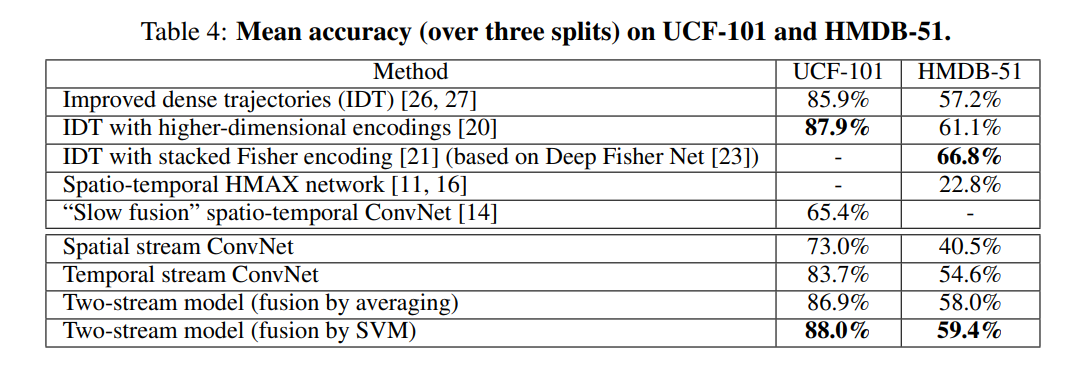
在测试时，给定一个视频，我们在它们之间等间隔地采样固定数量的帧（在实验中为25个）。然后从每个帧中，我们通过裁剪和翻转四个角和帧的中心来获得10个ConvNet输入。然后通过对采样的帧和其中的裁剪进行平均来获得整个视频的类别得分。最后相当于1个视频得到变250个输入（250view）。

## Optical flow

使用OpenCV工具箱中的的现成GPU实现来计算光流。尽管计算时间很快（一对帧需要0.06秒），但如果在飞行中进行，仍然会出现瓶颈，因此在训练之前预先计算了流。为了避免将位移场存储为浮点数，将流的水平分量和垂直分量线性缩放到[0, 255]范围内，并使用JPEG进行压缩（解压缩后，流会重新缩放回其原始范围）。这使UCF-101数据集的流大小从1.5TB减少到27GB。（将密集网络表示存储成JPEG图片）

# Evaluation

消融实验：

跟之前的方法对比：

# Conclusions and directions for improvement

我们提出了一种具有竞争力的性能的深度视频分类模型，该模型结合了基于ConvNets的单独的空间和时间识别流。目前看来，在光流上训练一个时间ConvNet明显优于在原始堆叠帧上进行训练。尽管使用光流作为输入，但我们的时间模型不需要大量的手工制作，因为流是通过基于通用假设的恒定性和平滑性的方法计算得出的。

目前我们的架构仍然缺少最先进的浅层表示的一些关键要素。其中最突出的一个是围绕轨迹的中心进行时空管道的局部特征池化。即使输入捕获了沿着轨迹的光流，我们网络中的空间池化也没有考虑轨迹。另一个潜在的改进领域是显式处理相机运动，而在我们的案例中，这是通过平均位移相减来补偿的。