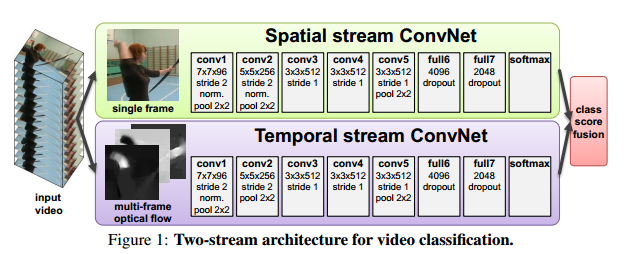
# 导读：行为识别是指通过分析视频、深度[传感器](http://www.iot-online.com/chuanganqi/" \t "http://www.iot-online.com/IC/tech/2017/_blank)等数据，利用特定的算法，对行人的行为进行识别、分析的技术。这项技术被广泛应用在视频分类、人机交互、[安防](http://www.iot-online.com/xingyeyingyong/sa/" \t "http://www.iot-online.com/IC/tech/2017/_blank)监控等领域。行为识别包含两个研究方向：个体行为识别与群体行为（事件）识别，而基于深度学习的人体行为识别成为了最近图像识别的标配。深度学习的方法包括基于无监督学习的行为识别、基于卷积神经网络的行为识别、基于循环神经网络以及一些拓展模型的方法，深思考小夏带给你最前沿的算法总结。

# 一、Two-stream

## 1 、Two-Stream Convolutional Networks for Action Recognition in Videos

<http://www.robots.ox.ac.uk/~vgg/software/two_stream_action/>

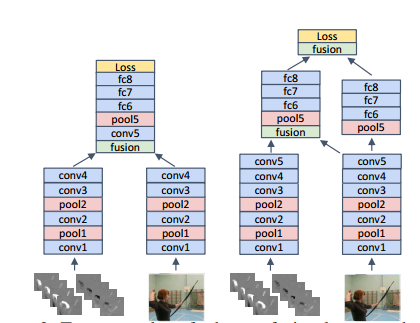


14年提出双流，利用帧图像和光流图像作为CNN的输入得到很好的效果。光流能够描述出视频帧的运动信息，一路是连续几帧的光流叠起来作为CNN的输入；另一路就是普通的单帧的CNN。其实就是两个独立的神经网络了，最后再把两个模型的结果平均一下。另外，它利用multi-task learning来克服数据量不足的问题。其实就是CNN的最后一层连到多个softmax的层上，对应不同的数据集，这样就可以在多个数据集上进行 multi-task learning。

## 2、 Convolutional Two-Stream Network Fusion for Video Action Recognition

**<http://www.robots.ox.ac.uk/~vgg/software/two_stream_action/>**

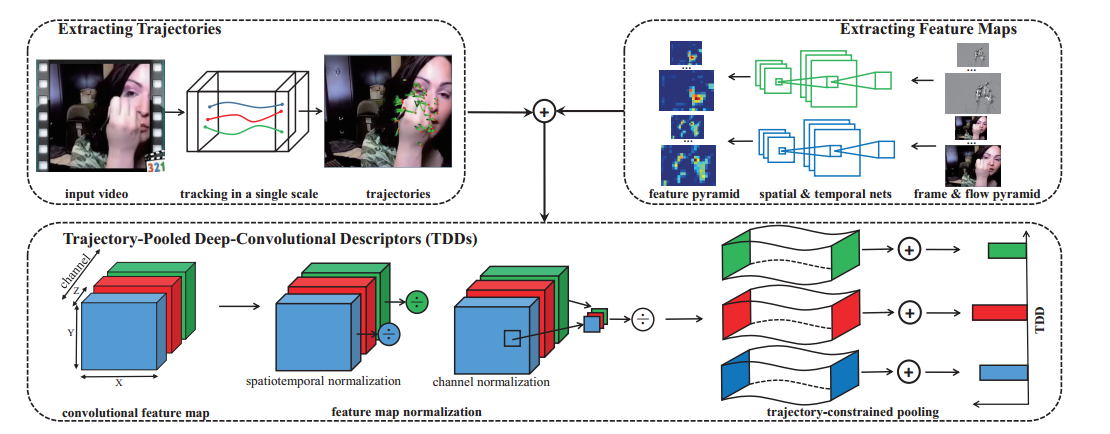
**<https://github.com/feichtenhofer/twostreamfusion>**



16年针对双流融合问题进行研究，得到卷积结束之后在全链接之前融合效果比较好，左边是单纯在某一层融合，右边是融合之后还保留一路网络，在最后再把结果融合一次。论文的实验表明，后者的准确率要稍高。

## 3、 Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors

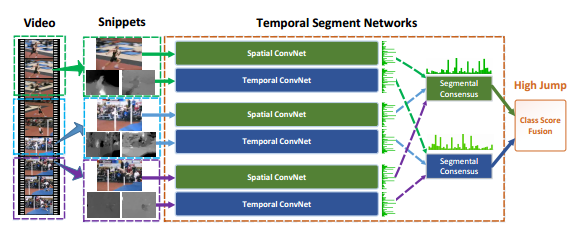
<https://wanglimin.github.io/tdd/index.html>



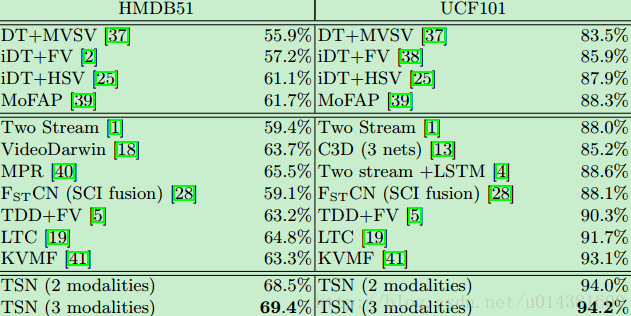
论文考虑了时间维的特性，引进了轨迹控制策略来采样，将手工设计的特征和深度学习结合。首先多个空间尺度上密集采样特征点，然后特征点跟踪得到轨迹形状特征，同时需要更有力的特征来描述光流，Fisher Vector方法进行特征的编码，最后svm采用one-against-rest策略训练多类分类器。

## 4、 Temporal Segment Networks: Towards Good Practices for Deep Action Recognition

<https://github.com/yjxiong/temporal-segment-networks>

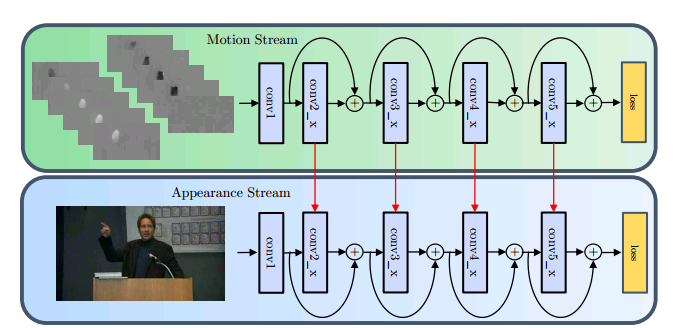


16年香港中文大学针对双流不能很好利用长时间信息，提出segment思路，将视屏分为前中后三段，每段经过双流然后融合结果。其中港中文还做了很多其他工作，<https://arxiv.org/abs/1507.02159> 比较了各种网络在action recognition中的效果，<https://wanglimin.github.io/> <http://yjxiong.me/>

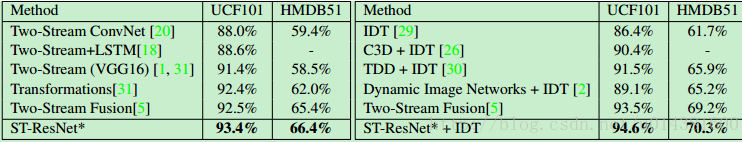


## 5、 Spatiotemporal Residual Networks for Video Action Recognition

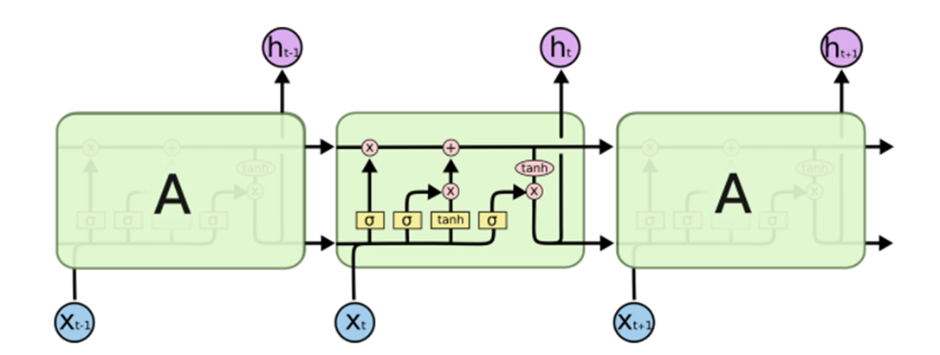
<https://feichtenhofer.github.io/>



使用了两个流，但是名字不是取为空间流和时间流，而是motion stream和appearance stream，但是本质不变，运动流接收的输入依然是堆叠的多帧光流灰度图片，而appearance stream和原来的空间流一致，接收的输入都是RGB图片，但是这里使用的 双流的两个流之间是有数据交换的，而不是像TSN网络一样在最后的得分进行融。单帧的潜力挖尽之后自然就会有人上3D Conv，Recurrent CNN，Grid RNN之类的东西。虽然深度学习大法好，不过也得按基本法来，直接上fancy的模型有较大概率吃力不讨好。



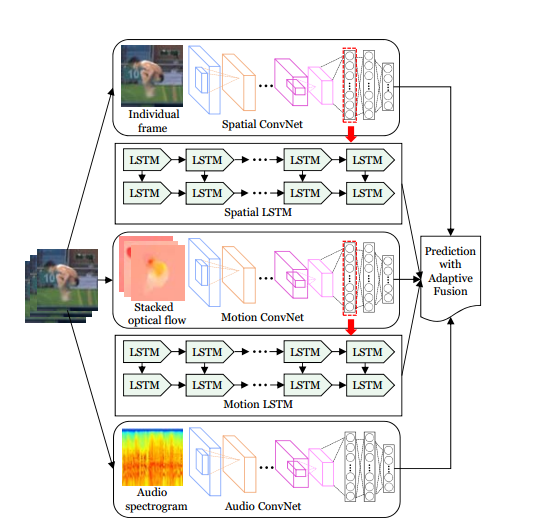
# 二 、LSTM 结构



Long Short Term 网络一般就叫LSTM，它是一种 RNN 特殊的类型。LSTM 通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM 的默认行为，而非付出很大代价才能获得的能力！

## 1、 Fusing Multi-Stream Deep Networks for Video Classification

<https://arxiv.org/abs/1509.06086>

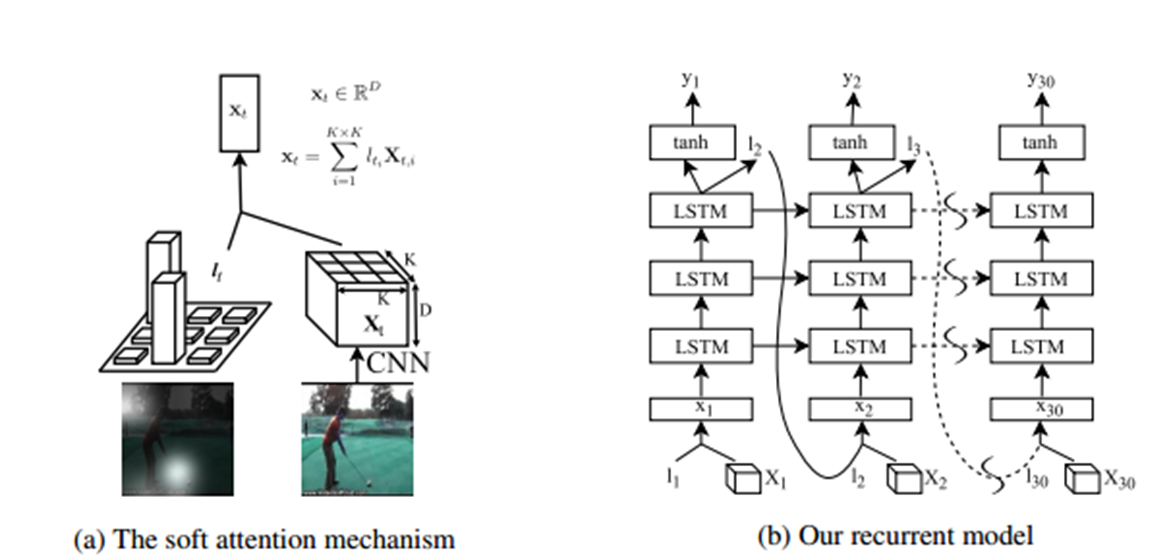


文章先CNN提取特征，包括rgb图光流图和语音频谱图，然后经过lstm最后融合。

## 2、 Action Recognition using Visual Attention

<http://shikharsharma.com/projects/action-recognition-attention/>

<http://www.cs.cmu.edu/~rsalakhu/>

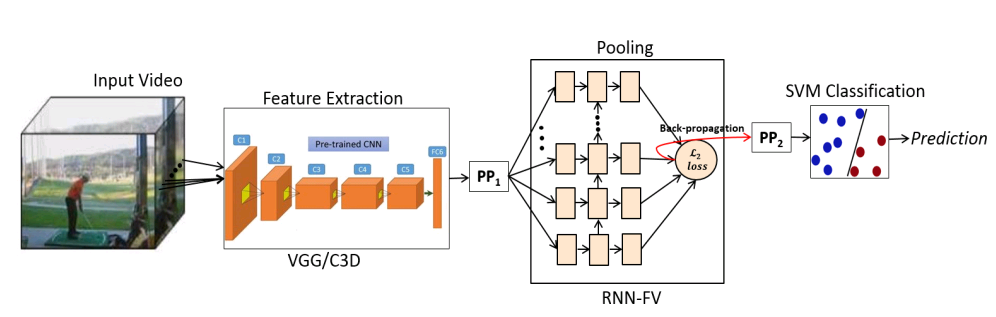


注意力模型终于来啦，人在看东西的时候，目光沿感兴趣的地方移动，甚至仔细盯着部分细节看，然后再得到结论。Attention就是在网络中加入关注区域的移动、缩放机制，连续部分信息的序列化输入。采用attention使用时间很深的lstm模型，学习视屏的关键运动部位。Attention相关：

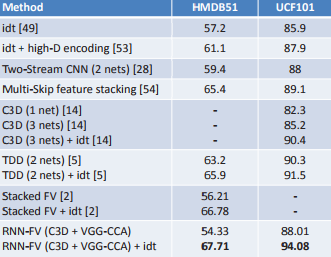
<http://www.cosmosshadow.com/ml/%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/2016/03/08/Attention.html>

## 3、 RNN Fisher Vectors for Action Recognition and Image Annotation

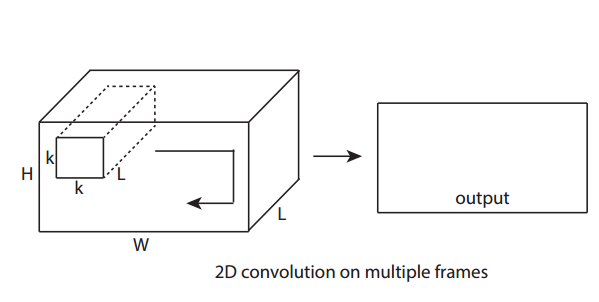
<http://www.eccv2016.org/files/posters/P-4A-30.pdf>

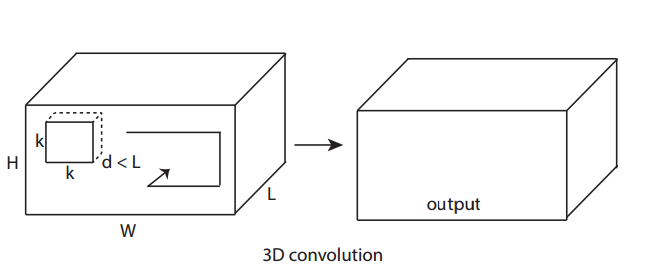


文章典型的特征提取，分类思路文章采用卷积网络提取特征之后进过pca降维，然后Fisher Vector编码扔给RNN再pca降维，最后svm分类。Ucf101上实验结果到了94%



# 三、C3D



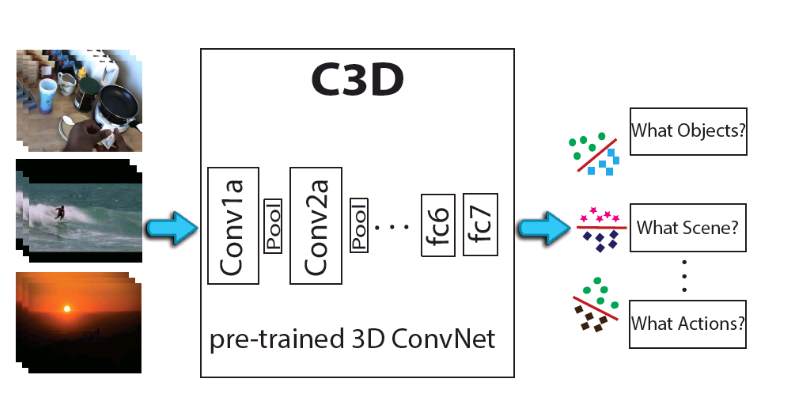


3D CNN 应用于一个视频帧序列图像集合，并不是简单地把图像集合作为多通道来看待输出多个图像（这种方式在卷积和池化后就丢失了时间域的信息，如图6上）， 而是让卷积核扩展到时域，卷积在空域和时域同时进行，输出仍然是有机的图像集合

## 1、Learning Spatiotemporal Features with 3D Convolutional Networks

<https://github.com/facebook/C3D>

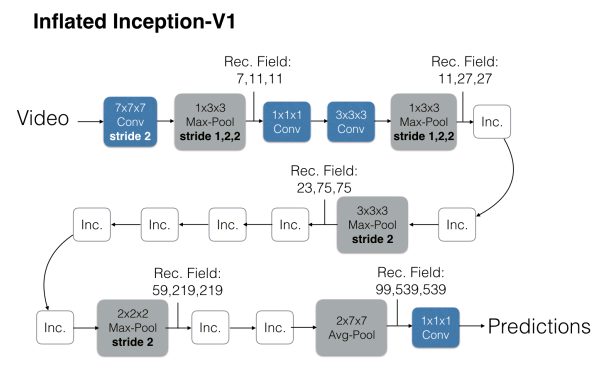
<https://gist.github.com/albertomontesg/d8b21a179c1e6cca0480ebdf292c34d2>



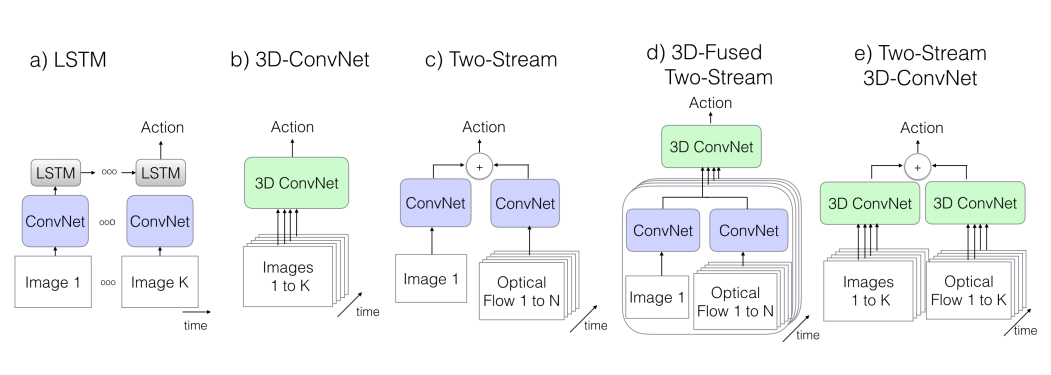
将视频分成多个包含16帧的片段作为网络的输入。第一个池化层d=1，是为了保证时间域的信息不要过早地被融合，接下来的池化层的d=2。有所卷积层的卷积核大小为3x3x3，相对其他尺寸的卷积核，达到了精度最优，计算性能最佳。

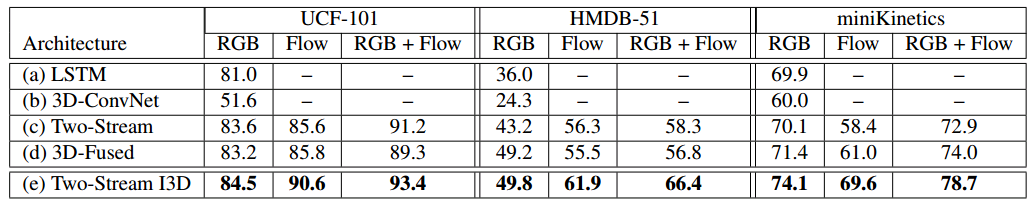
## 2、 Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

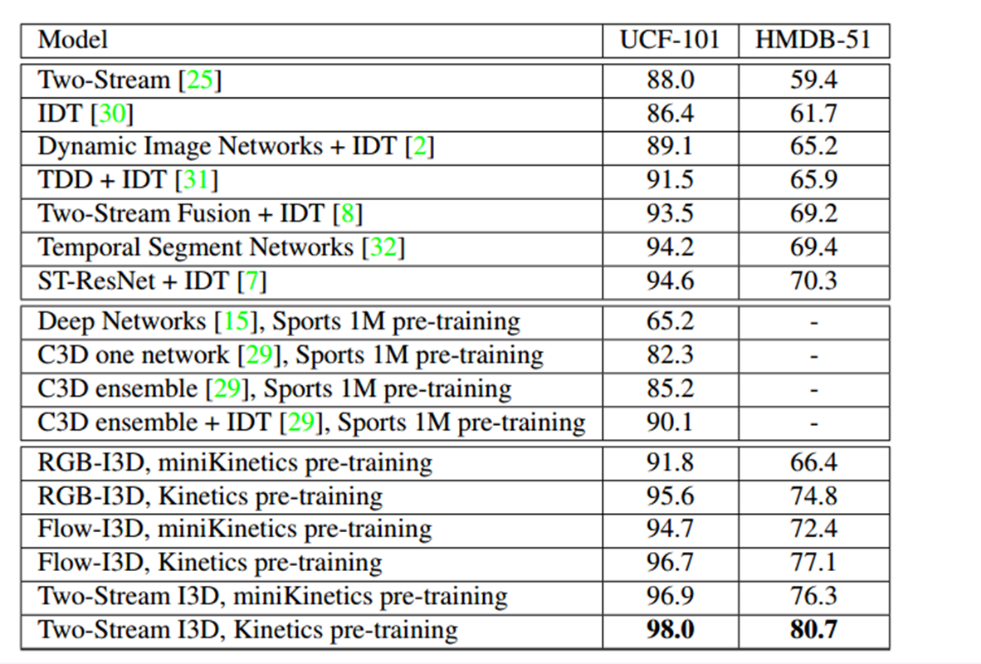
<https://deepmind.com/research/publications/quo-vadis-action-recognition-new-model-and-kinetics-dataset/>



以往的Conv3D效果很差的原因之一就是数据集太小，喂不饱网络。文章中的3D网络并不是随机初始化的，而是将在ImageNet训好的2D模型参数展开成3D，之后再训练。因此叫Inflating 3D ConvNets. 本文选用的网络结构为BN-Inception(TSN也是)，但做了一些改动。如果2D的滤波器为N\*N的，那么3D的则为N\*N\*N的。具体做法是沿着时间维度重复2D滤波器权重N次，并且通过除以N将它们重新缩放. 在前两个池化层上将时间维度的步长设为了1，空间还是2\*2。最后的池化层是2\*7\*7。训练的时候将每一条视频采样64帧作为一个样本，测试时将全部的视频帧放进去最后average\_score。除最后一个卷积层之外，在每一个都加上BN层和Relu。对于I3D的效果为什么好，作者解释说I3D有64帧的感受野。可以更好地学习时序信息。再就是先用ImageNet的模型做了预训练



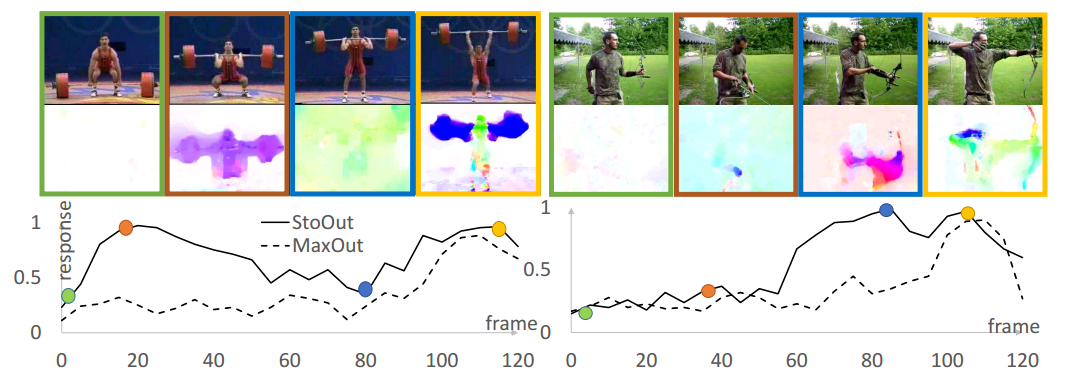




I3D这个网络结构的提出是很显然，但用2D的ImageNet模型做预训练以及参数展开分配还是挺具有创新性的，虽然在TSN中处理光流的第一个卷积层时就有使用过类似的方法。这个实验室真有能力，以往的数据集上效果很难提升，自己就搞了个大数据集。那个Kinetics的I3D模型是在**64块GPU**上跑出来的。

**四 、其他**

**1、 A Key Volume Mining Deep Framework for Action Recognition**



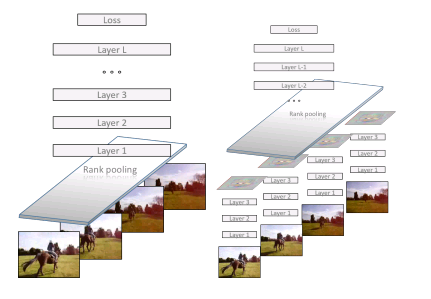
现即便是 trimmed video （例如 UCF101 数据集），实际的动作发生的时空位置也是非常不确定的：我们既不知道做动作的人在什么空间位置，也不知道真正的动作发生的精确时间位置。更糟糕的是，和动作类别直接相关的，具有区分性的 （discriminative）key volume 往往占比非常小，这在 flow stream 上表现得尤为突出。

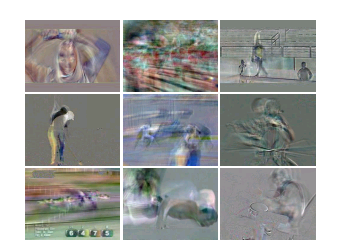
于是我们就想能否先把这些 key volume 找出来，直接用以训练分类器，这样可以免受噪声数据的干扰，更加聚焦在动作本质上。但实际上，在得到一个好的分类器之前我们是很难自动地将 key volume 挑出来的。于是我们陷入了一个鸡生蛋，蛋生鸡的困境。

借鉴 Multiple Instance Learning 的思想，我们把鸡和蛋的问题放在一起来优化解决：在训练分类器的同时，挑选 key volume；并用挑出来的 key volume 更新分类器的参数。这两个过程无缝地融合到了 CNN （卷积神经网络）的网络训练的 forward 和 backward 过程中，使得整个训练过程非常优雅、高效。

**2 、 Dynamic Image Networks for Action Recognition**

<https://github.com/hbilen/dynamic-image-nets>





CNN的输入一般是图片，针对输入时视屏，能否将视屏压缩成一张图，以图来表征视屏的信息？答案是可疑的，针对对视频中的RGB图像进行rank pooling处理，以此作为cnn的输入。虽然最终的效果不是特别好，但是想法很nice。