CVlab Progress

2 stream Networks for action recognition and classification

- Earlier works focused on a 2-stream method, one 2D CNN for time and one for spatial domain

- This idea worked because of imagenet dataset's ability to pretrain spatial convnets.

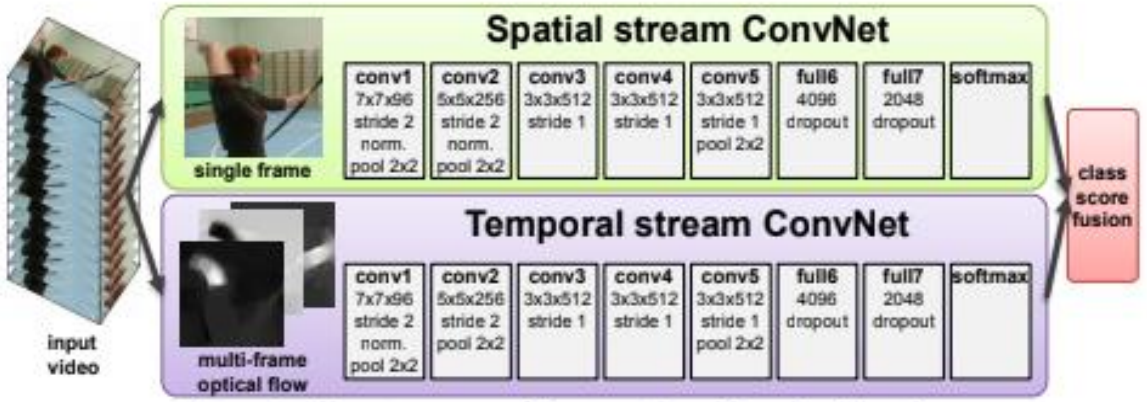
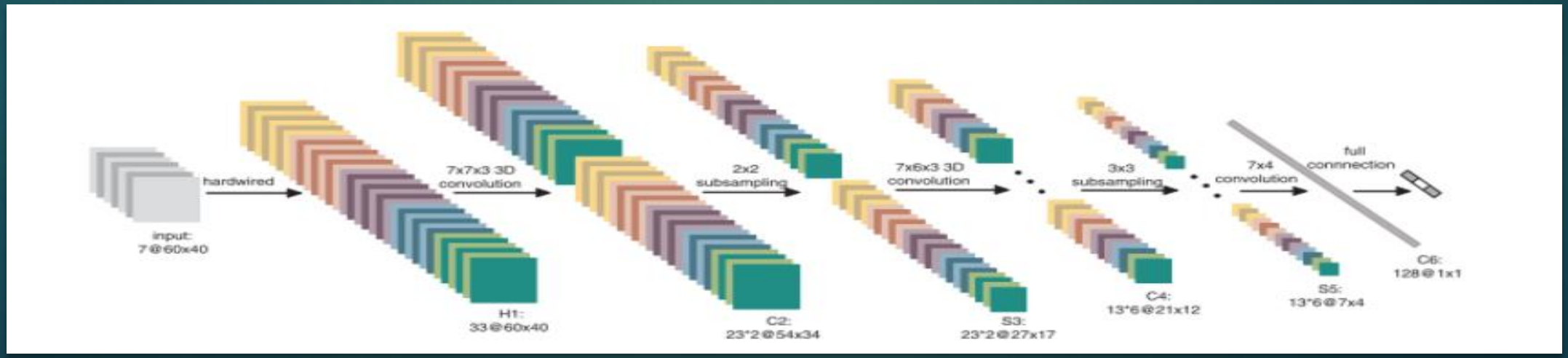


Figure 1: **Two-stream architecture for video classification.**

"Two-Stream Convolutional Networks for Action Recognition in Videos"-Karen Simonyan Andrew Zisserman

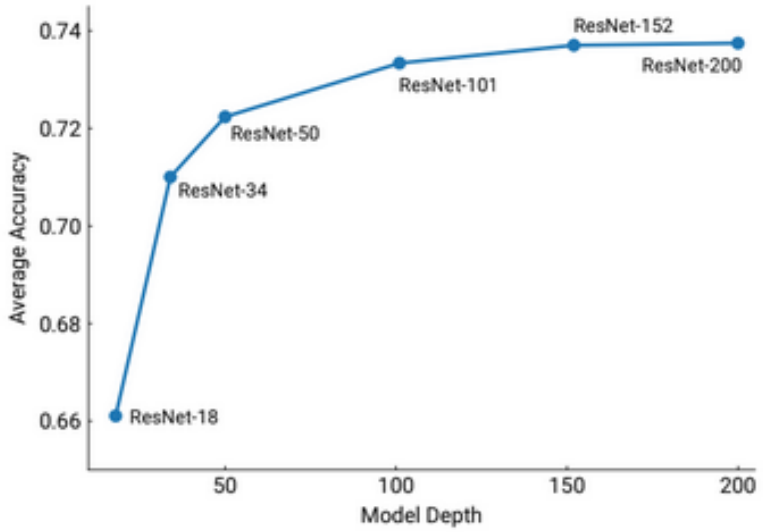
3D CNNs form action recognition

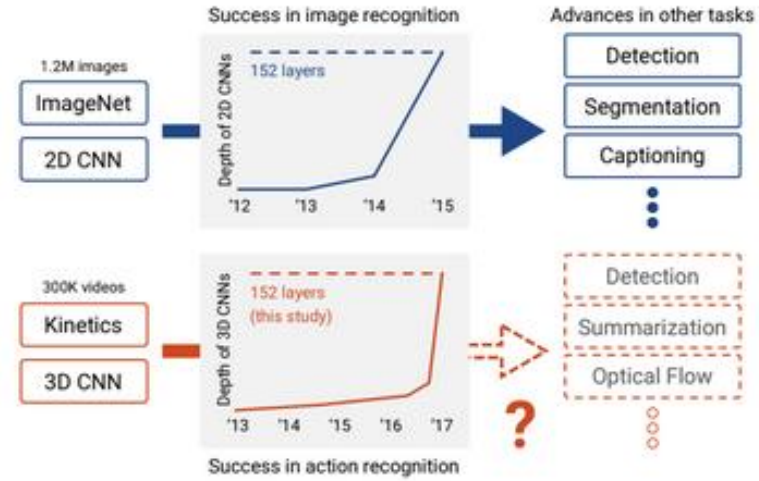
- 3D CNNs have been proposed for this task in Yu et.el. , Karpatjy et.et
- This approach consists of training the network in an end to end manner using video datasets such as UCF-101 and HMDB datasets
- This approach seemed like the natural way to go
- Struggled to beat state of the art results of the 2-stream netwroks



Kinetics dataset and its implications

- The kinetics dataset is huge, containing more than 300K videos
- In the paper "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet ?", by Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh, it is shown that kinetics dataset holds more than enough data to train 3DCNNs
- This huge dataset means less chances of overfitting
- Comparable to imagenet dataset for images
- Can train 3D resnet-152 without overfitting





Our Ideas and approaches

We are coming up with ideas for using this pretrained 3D-Resnet to some other task

One potential area could be action similarity assessment, which uses higher level feature generated from a deep layer of the 3D resnet to decide whether two actions are similar or not

Another Idea is Action Quality Assessment, explored in the paper "What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment ", by Parmar and Morris. This paper uses a technique called Multitask AQA which is interesting

Thank You