Two-Stream Networks

Two-Stream Convolutional Networks for Action Recognition in Videos (2014) 7k

Pioneering Work of Video Understanding

When the neural network can't solve the problem, provide better data (scale, type)

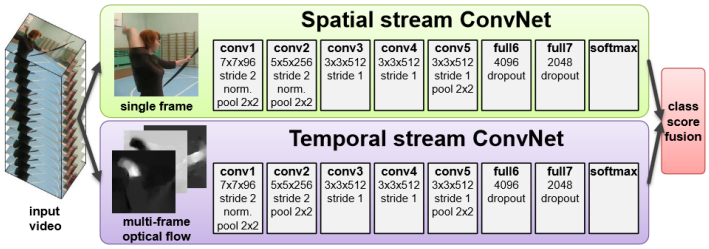

Figure 1: **Two-stream architecture for video classification.**

Spatial stream ConvNet

operates on individual video frames, effectively performing action recognition from still images

similar to AlexNet

pre-trained + fine-tuning (easy to overfit, dropout = 0.9) / last layer (dropout = 0.9)

Accuracy = 73%

Optical flow ConvNets (temporal)

Optical flow stacking / Trajectory stacking

Bi-directional optical flow: changing to bi-directional will generally improve accuracy, at least not worse

Optical flow is computed using the off-the-shelf GPU implementation from the OpenCV toolbox

Accuracy = 83.7%

Fusion by averaing / SVM

Accuracy = 88%