EE655: Computer Vision & Deep Learning

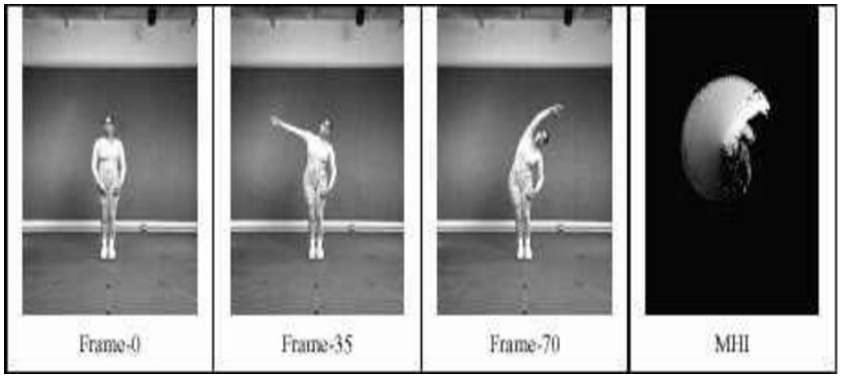
Lecture 14

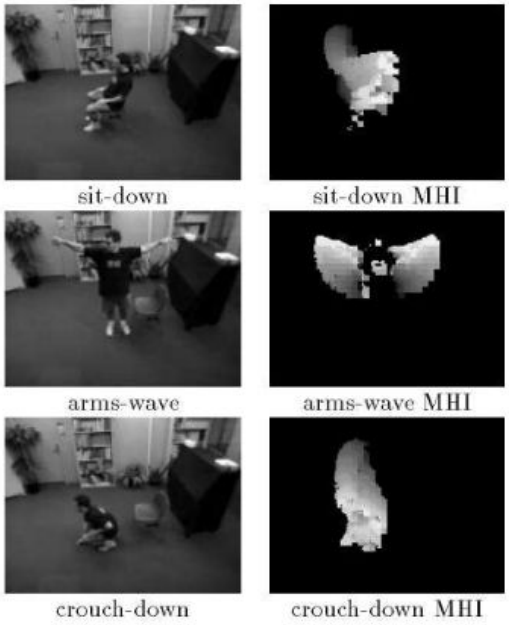
Koteswar Rao Jerripothula, PhD
Department of Electrical Engineering
IIT Kanpur

Outline:

- Motion History Images
- CNNs for Videos

Examples


Frame-0
Frame-35
Frame-70
MHI


sit-down
sit-down MHI
arms-wave
arms-wave MHI
crouch-down
crouch-down MHI

Motion History Images

How recently the pixel has moved?

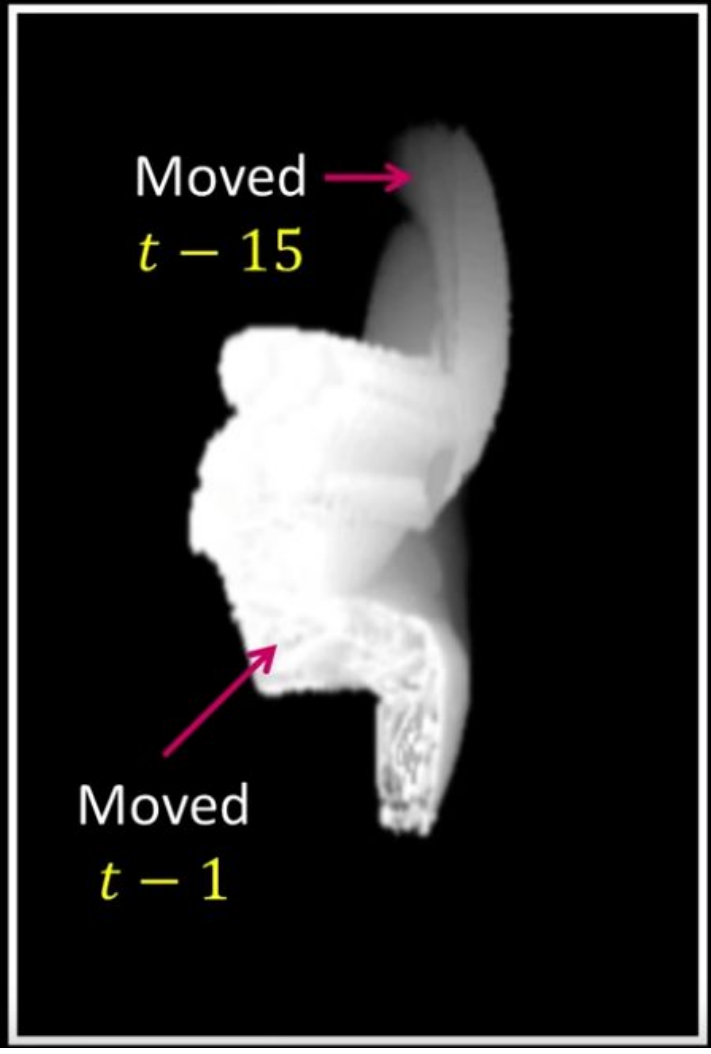
First, we need to determine whether a pixel has moved in the current frame or not.
At current frame ‘t’, we compute MHI as follows

if moving:

$$I_\tau(x, y, t) = \tau$$

otherwise:

$$I_\tau(x, y, t) = \max(I_\tau(x, y, t-1) - 1, 0)$$



Two-stream architecture

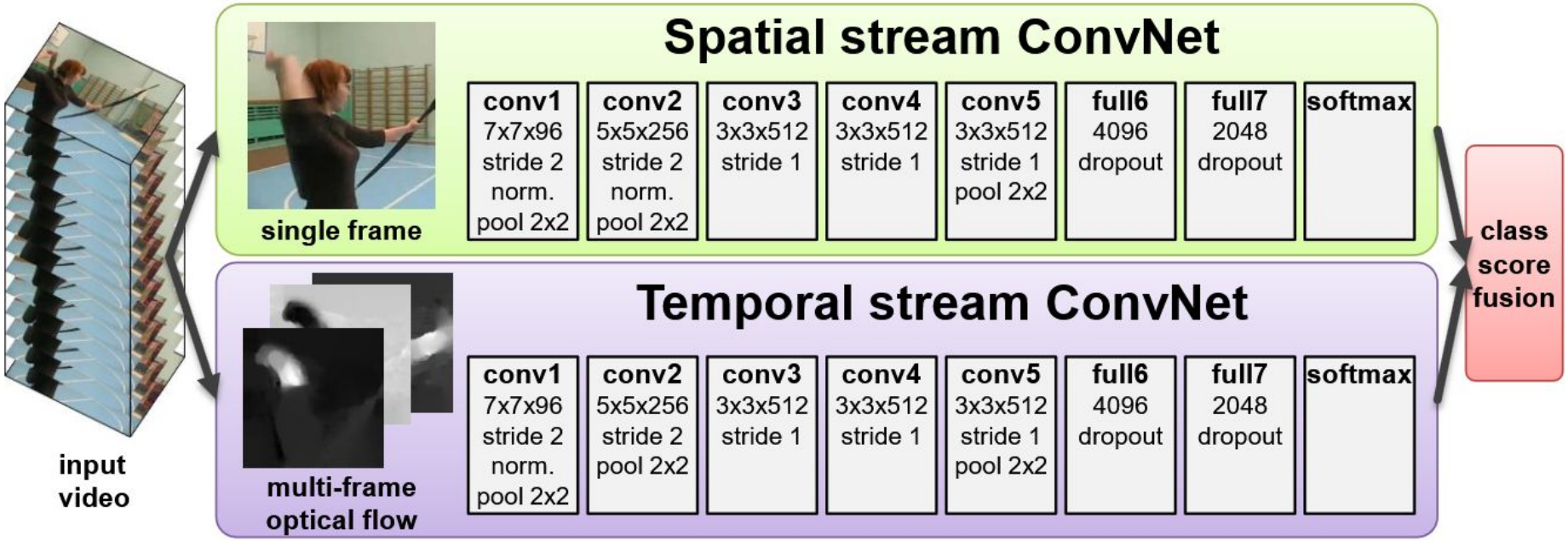
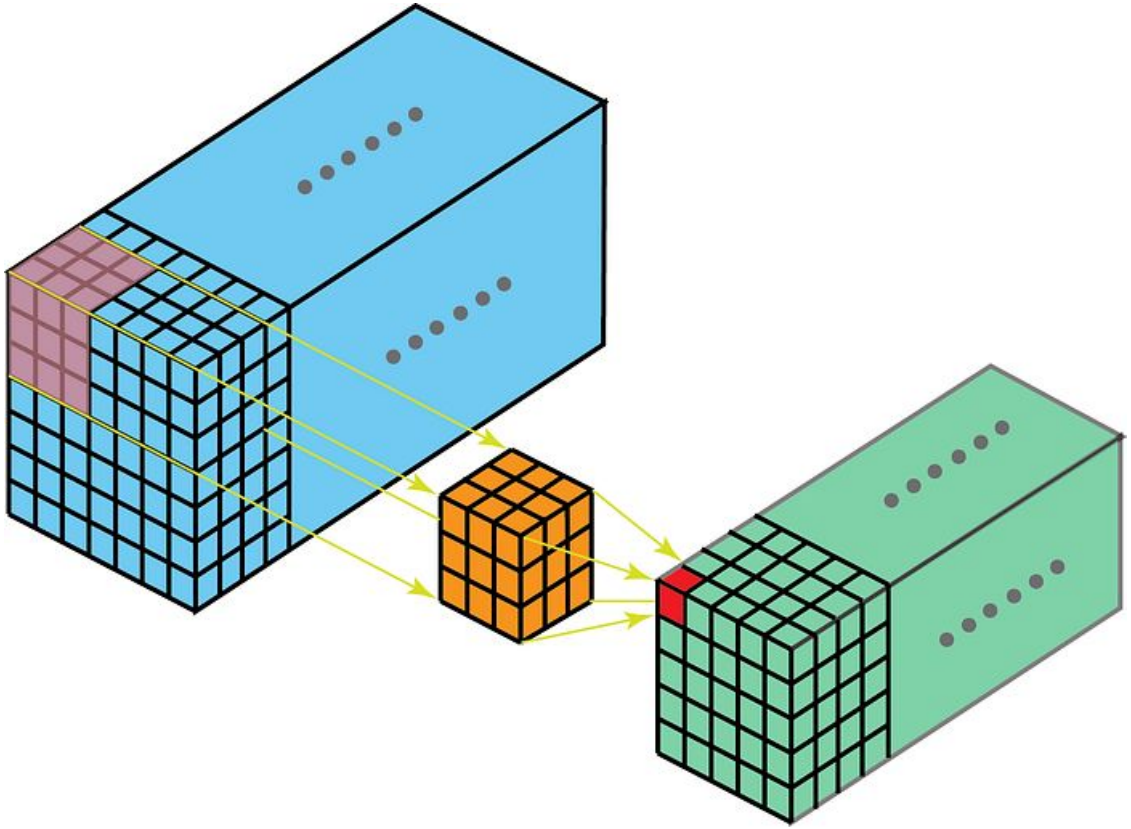


Figure 1: **Two-stream architecture for video classification.**

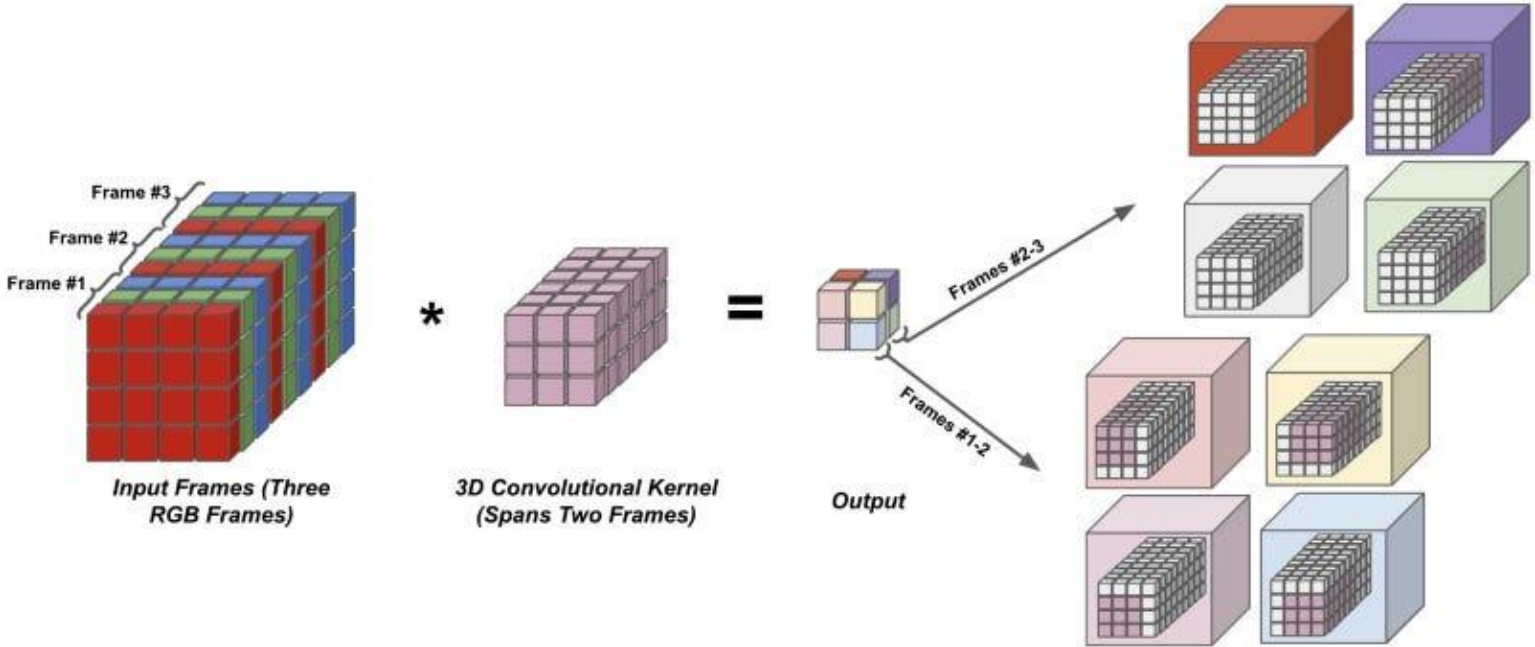
Idea behind 3D convolution



3D Convolution for videos

4 dimensional input and kernel:

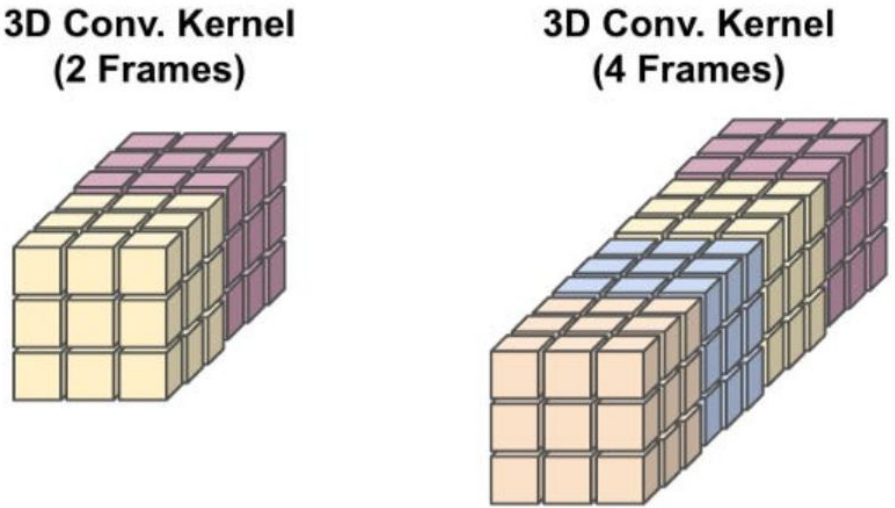
2 spacial, 1 for channels, and 1 for frames



Spatiotemporal representation

The output representation of a 3D convolution is, by nature, spatiotemporal (i.e., captures both spatial information within each frame and temporal information between adjacent frames).

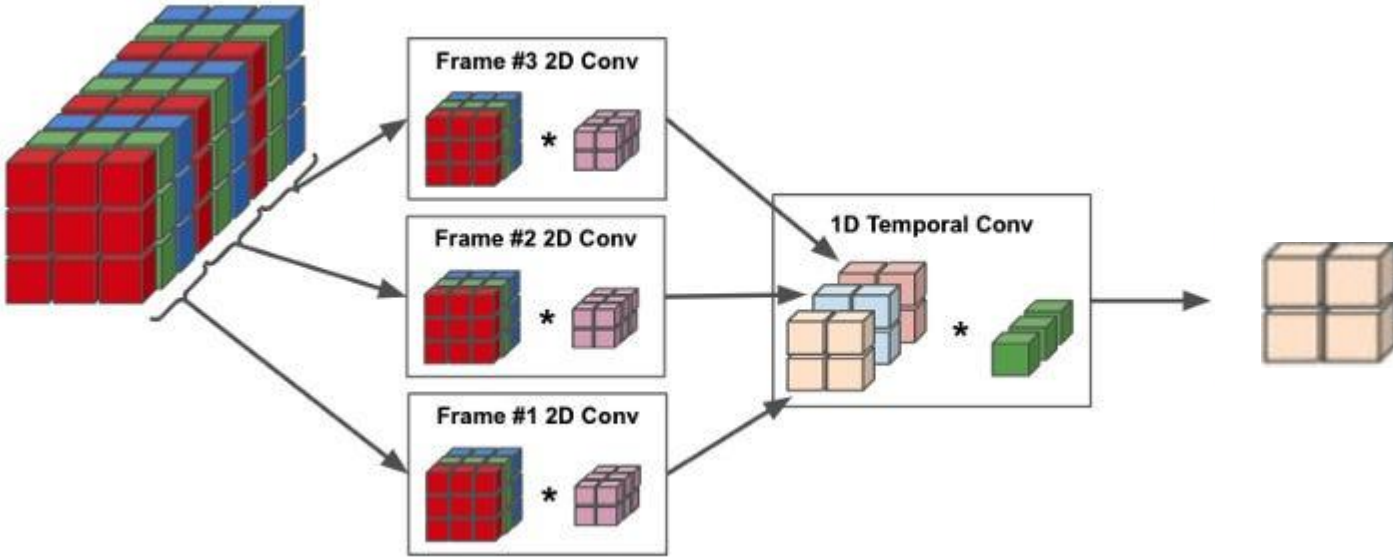
This added temporal dimension causes 3D CNNs to have higher computational costs and heavier data requirements for achieving acceptable training and generalization performance.



Alleviating issues with 3D Convolutions

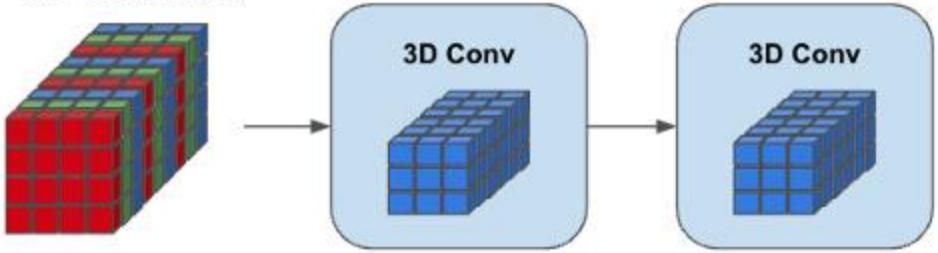
- Only using 3D convolutions in a smaller number of networks layers, and allowing remaining layers to perform 2D convolution operations.
- Factorizing 3D convolutions into separate 2D spatial and 1D temporal convolution operations applied in sequence

Factorized 3D Convolution

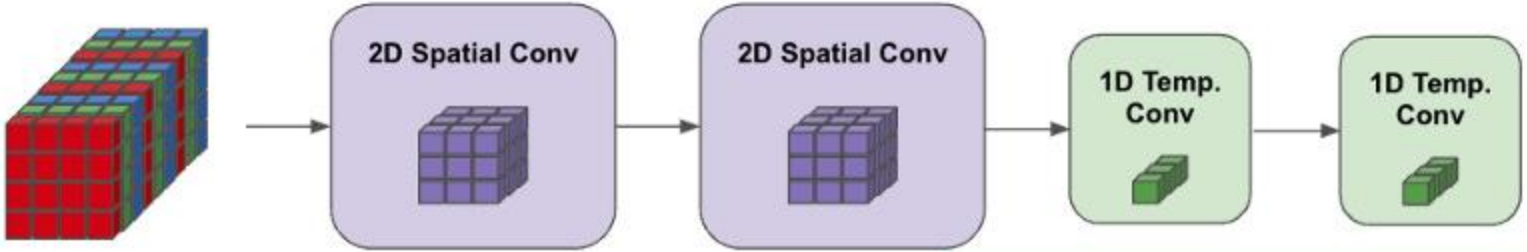


It reduces the number of trainable parameters significantly. Implementation-wise, two 3D convolutions with filters of sizes 1x3x3x3 and Fx1x1x1, thus reducing the number of trainable parameters from Fx3x3x3 to F+3x3x3

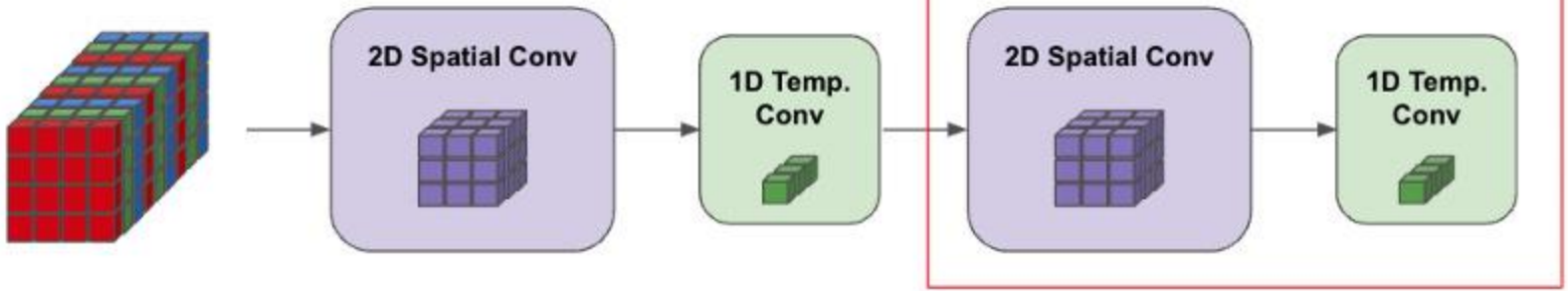
3D ConvNet



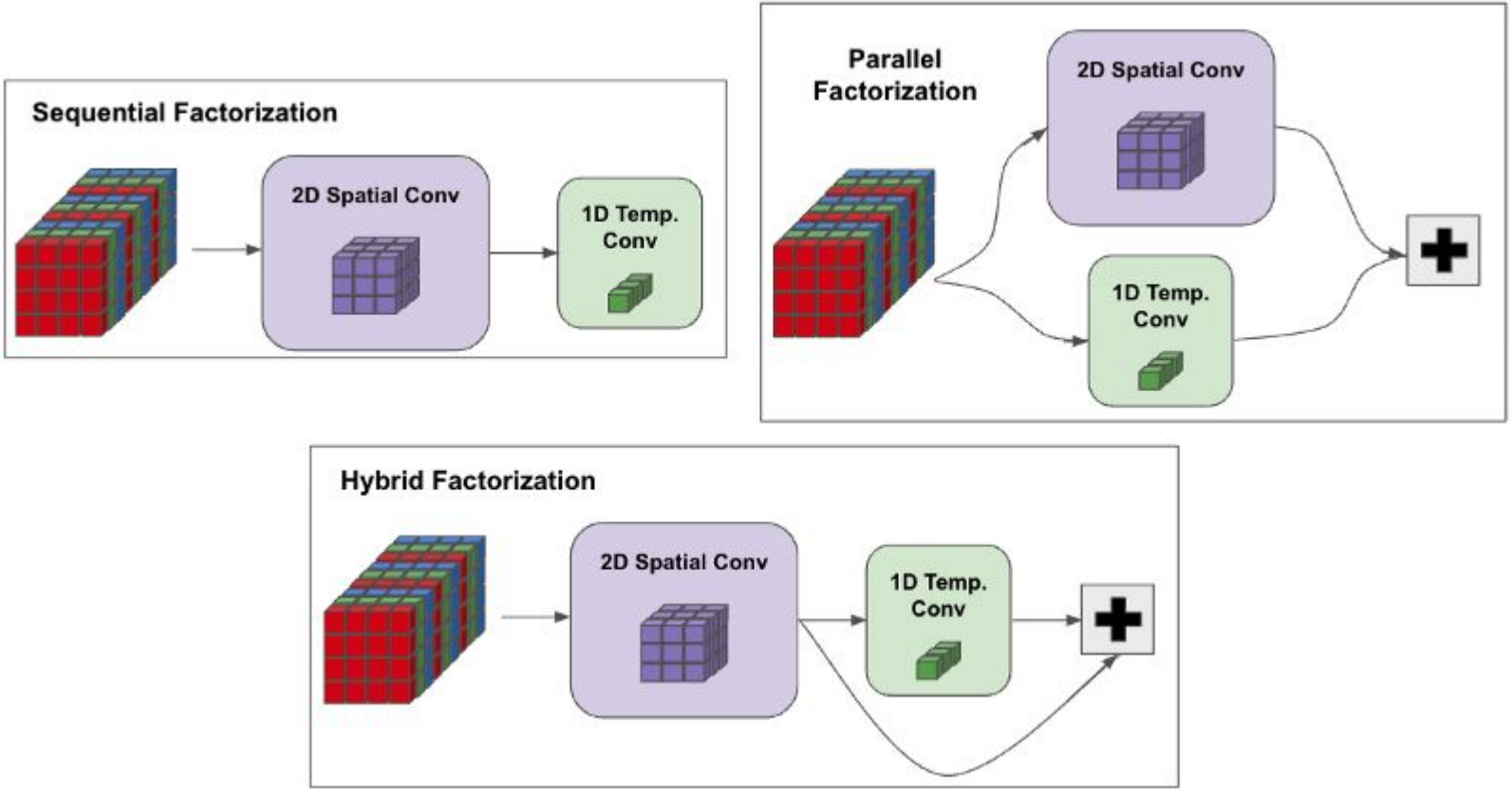
3D ConvNet with Factorized Architecture



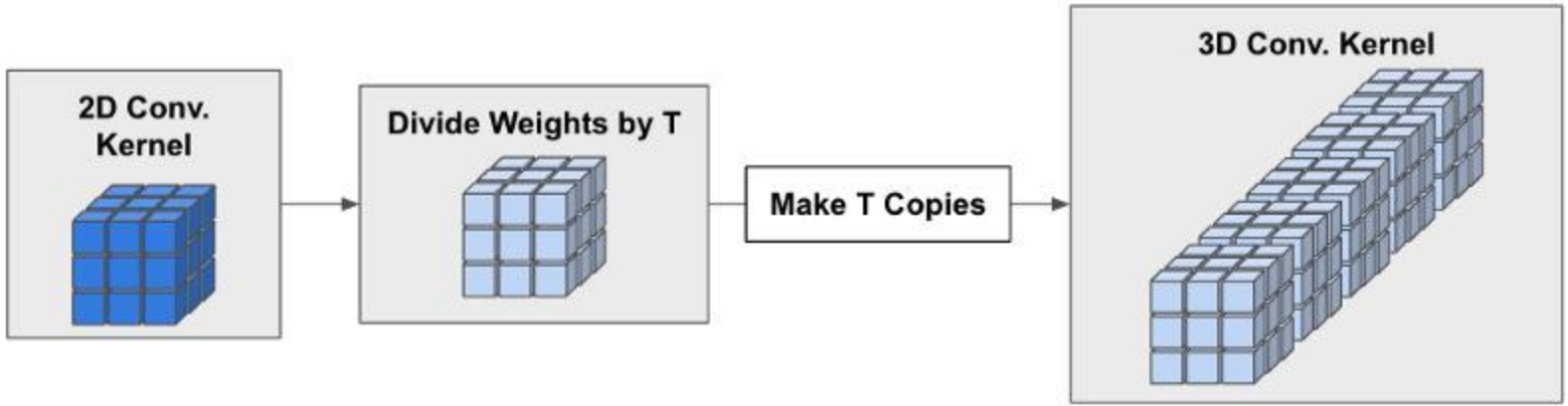
3D ConvNet with Factorized 3D Convolutions



Variants of Factorization



Inflating 2D Networks (Transferring Image-based learning to video domain)



Reference

https://cameronrwolfe.substack.com/p/deep-learning-on-video-part-three-diving-deeper-into-3d-cnns-cb3c0daa471e