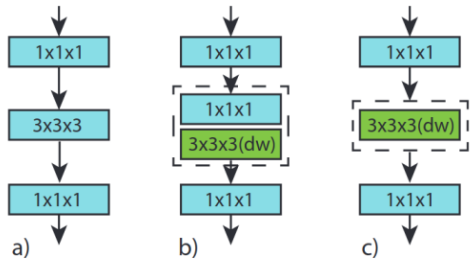
Video Classification with Channel-Separated Convolutional Networks

Du Tran, Heng Wang, Lorenzo Torresani, Matt Feiszli

This paper studies the effects of different design choices in 3D group convolutional networks for video classification. By testing different designs, they show that factorizing 3D convolutions by separating channel interactions (a 1x1x1 convolution) and spatiotemporal interactions (channel separated 3x3x3 convolution) lead to improved accuracy and lower computational cost. 3D channel-separated convolutions can also provide a form of regularization, yielding lower training accuracy but higher test accuracy compared to 3D convolutions.



layer name	output size	ResNet3D-simple	ResNet3D-bottleneck
conv1	T×112×112	3×7×7, 64, stride 1×2×2	
pool1	T×56×56	max, 1×3×3, stride 1×2×2	
conv2_x	T×56×56	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}\times b_1$	$\begin{bmatrix} 1\times1\times1, 256 \\ 3\times3\times3, 64 \\ 1\times1\times1, 256 \end{bmatrix}\times b_1$
conv3_x	$\frac{T}{2}\times28\times28$	$\begin{bmatrix} 3\times3\times3, 128 \\ 3\times3\times3, 128 \end{bmatrix}\times b_2$	$\begin{bmatrix} 1\times1\times1, 512 \\ 3\times3\times3, 128 \\ 1\times1\times1, 512 \end{bmatrix}\times b_2$
conv4_x	$\frac{T}{4}\times14\times14$	$\begin{bmatrix} 3\times3\times3, 256 \\ 3\times3\times3, 256 \end{bmatrix}\times b_3$	$\begin{bmatrix} 1\times1\times1, 1024 \\ 3\times3\times3, 256 \\ 1\times1\times1, 1024 \end{bmatrix}\times b_3$
conv5_x	$\frac{T}{8}\times7\times7$	$\begin{bmatrix} 3\times3\times3, 512 \\ 3\times3\times3, 512 \end{bmatrix}\times b_4$	$\begin{bmatrix} 1\times1\times1, 2048 \\ 3\times3\times3, 512 \\ 1\times1\times1, 2048 \end{bmatrix}\times b_4$
pool5	1×1×1	spatiotemporal avg pool, fc layer with softmax	

The paper proposes two versions of CSNs:

- Interaction preserved CSN (ipCSN): in this version, the 3x3x3 standard convolutions are replaced by a 1×1×1 traditional convolution and a 3×3×3 depthwise convolution. This reduced the number of parameters and FLOPs but preserved all channel interactions with the added 1x1x1 convolutions.
- Interaction reduced CSN (irCSN): in this version, the 3x3x3 standard convolutions are replaced with just a 3×3×3 depthwise convolution. In this case, there is a reduction in channel interaction. The experiments show that both can perform better than standard ResNet3D, however, with irCSN, the model needs to have more depth to overcome the reduced number of channel interactions.