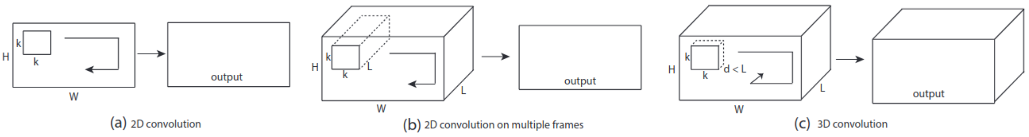
Learning Spatiotemporal Features with 3D Convolutional Network

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri

The main component of the model is 3D convolutions. With 2D convolutions, the kernel is displaced only over the spatial dimensions while the computation is done over the whole depth, either RGB for images or the temporal dimension in the form of the number of frames for videos. However, with 3D convolutions, the displacement is also done in the temporal dimension. For example, with a filter of size H x W x d, and d is over the temporal dimension of L frames, the kernel is applied (L - d / 2) / strides instead of a single time.



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

What is the correct depth? For d x k x k 3D convolutions, we need to find the correct temporal depth. For 5 layers of convolutions, the authors propose two variations:

- Homogeneous: all of the conv layers have the same depth: 1, 3, 5, or 7.
- Variable: either increasing 3-3-5-5-7 or decreasing: 7-5-5-3-3.

By testing these different variations. The authors find that the best choice is a fixed depth with a depth of 3, so 3-3-3-3-3. Given these results, the final C3D architecture has 8 convolutions, 5 max-pooling layers, and 2 fully connected layers, followed by a softmax output layer. All of the 3D convolutions are 3x3x3 with stride 1 in both spatial and temporal dimensions. All pooling kernels are 2x2x2, except for the first pooling layer which is 1×2×2 to avoid removing the temporal information too early.

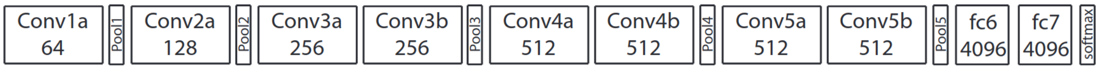


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.