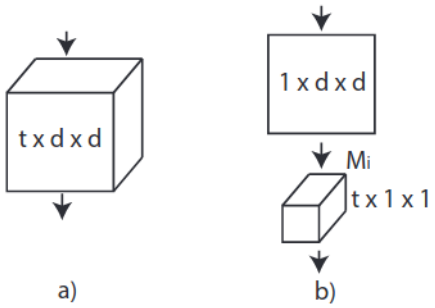
A Closer Look at Spatiotemporal Convolutions for Action Recognition

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri

Image-based 2D CNNs, while only operating on individual frames, still achieve competitive results on video understanding many benchmarks even if they are unable to model temporal information. However, temporal reasoning still remains an important component for effective action recognition. This paper explores two ways of effectively modeling the temporal information:

- using 3D convolutions only at the start or the end, thus considering the motion information either as a low level or high level one.
- considering (1+2)D convolutions as a middle ground between the simple 2D convs and the expensive 3D convs, which consists of factorizing 3D convolutions into separate spatial and temporal components and using it in a residual learning framework. This (1+2)D factorization adds additional nonlinear rectification between these two operations, doubling the total number of nonlinearities compared to the 3D version of the model, giving the model better representation capabilities. Additionally, such a decomposition simplifies the training process the results in lower training loss where the convolution is forced into separate spatial and temporal space.



The authors then compare different variants of the ResNet architecture adapted for video recognition, either fully 2D model, mixed with 3D convs at the start or at the end, fully 3D conv model, or fully (1+2)D model. The results show that R(2+1)D performs the best while having a similar computational cost.

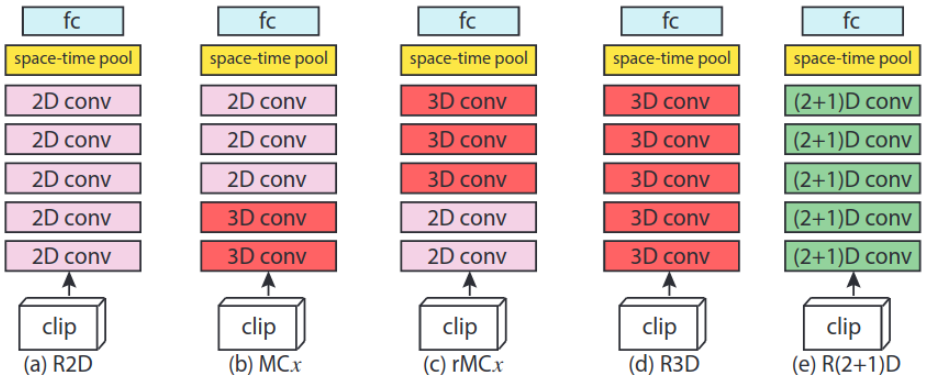


Figure 1. **Residual network architectures for video classification considered in this work**. (a) R2D are 2D ResNets; (b) MCx are ResNets with mixed convolutions (MC3 is presented in this figure); (c) rMCx use reversed mixed convolutions (rMC3 is shown here); (d) R3D are 3D ResNets; and (e) R(2+1)D are ResNets with (2+1)D convolutions. For interpretability, residual connections are omitted.

This is very similar to Pseudo-3D Convs, which also proposes an adaptation of the bottleneck block of ResNet 2D to video classification. Three different pseudo-3D blocks were introduced: P3D-A, P3D-B, and P3D-C. The blocks implement different orders of convolution: spatial followed by temporal, spatial, and temporal in parallel, and spatial followed by temporal with skip connection from spatial convolution to the output of the block, respectively. R(2+1)D differs by using the same convs all the way, while P3D starts with 2D convs then uses the proposed version.

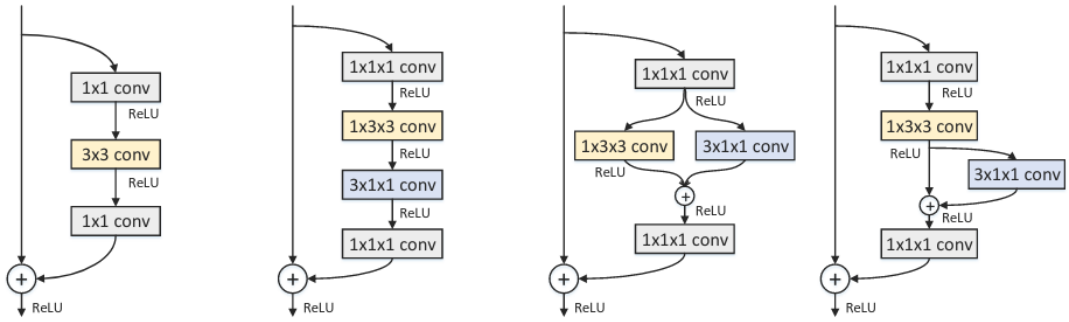


Figure 3. Bottleneck building blocks of Residual Unit and our Pseudo-3D.