Learning Spatiotemporal Features with 3D Convolutional Networks

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

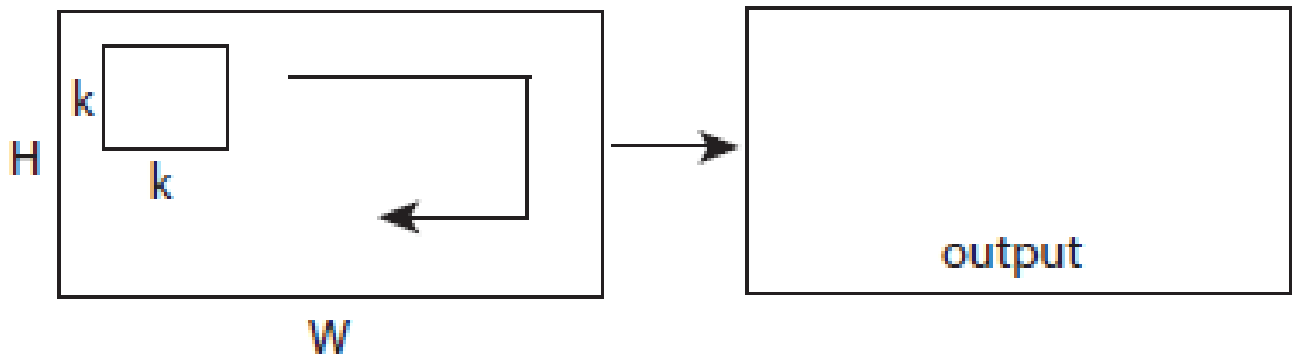
Abstract

- 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets.
- A homogeneous architecture with small 3 × 3 × 3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets.
- Our learned features, namely C3D (Convolutional 3D), with a simple linear classifier outperform state-of-the-art methods on 4 different benchmarks and are comparable with current best methods on the other 2 benchmarks.

Introduction

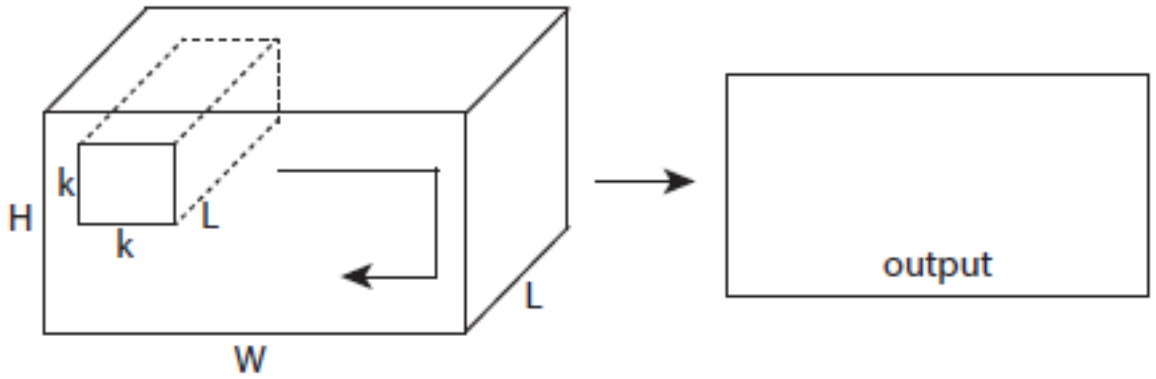
- Video analysis
- Four properties of video descriptors
  - Need to be Generic, compact, efficient, simple
- Spatio-temporal using 3D CNN

2D Convolution



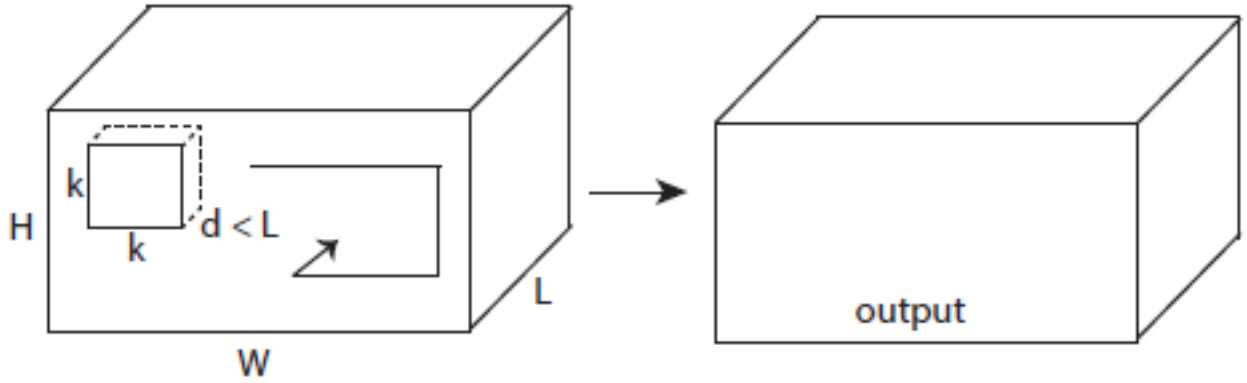
(a) 2D convolution

2D Conv on multiple frames



(b) 2D convolution on multiple frames

3D Convolution



(c) 3D convolution

Temporal search

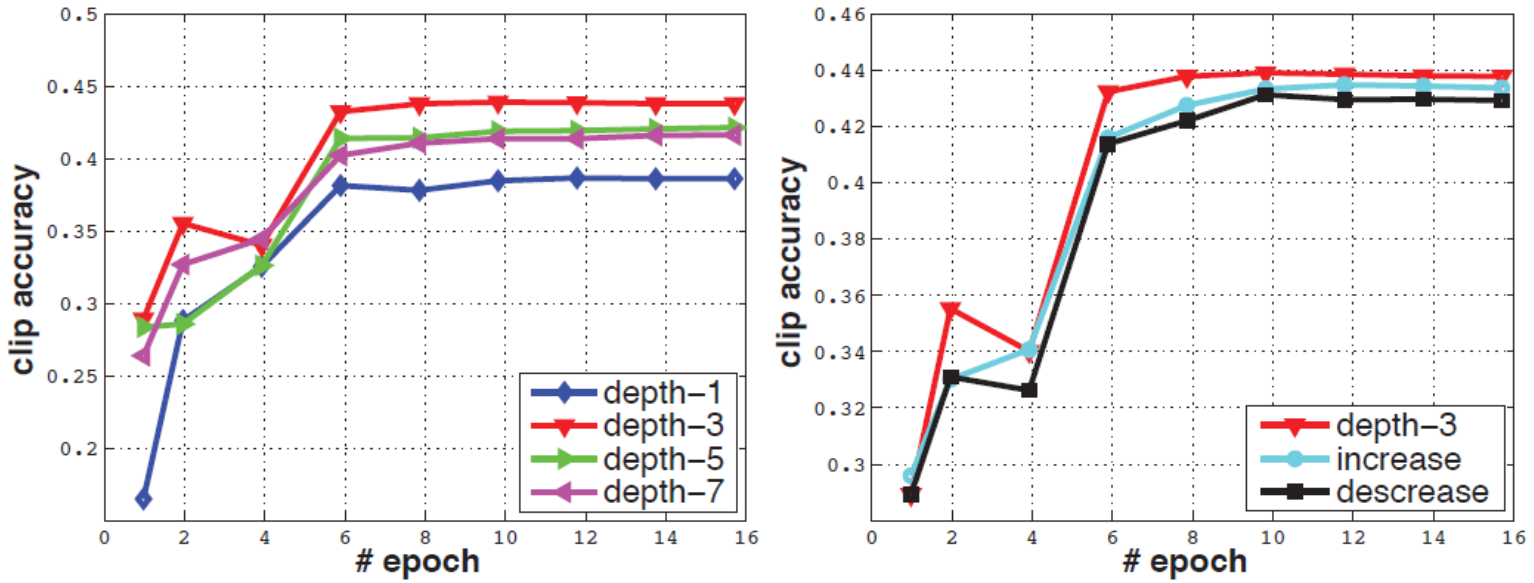


Figure 2. **3D convolution kernel temporal depth search.** Action recognition clip accuracy on UCF101 test split-1 of different kernel temporal depth settings. 2D ConvNet performs worst and 3D ConvNet with $3 \times 3 \times 3$ kernels performs best among the experimented nets.

Architecture

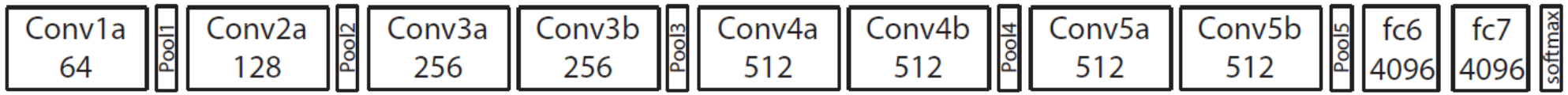


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling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

Results

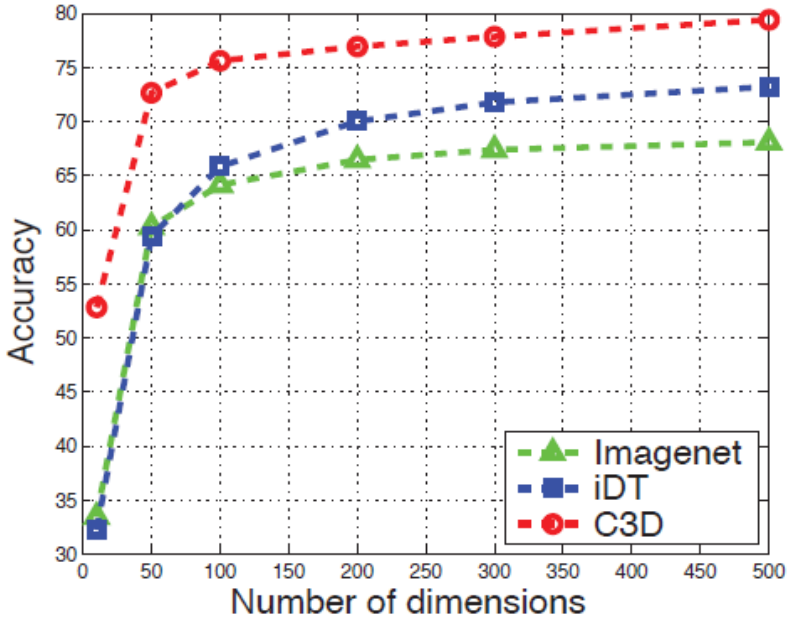


Figure 5. **C3D compared with Imagenet and iDT in low dimensions**. C3D, Imagenet, and iDT accuracy on UCF101 using PCA dimensionality reduction and a linear SVM. C3D outperforms Imagenet and iDT by 10-20% in low dimensions.