Learning Spatiotemporal Features with 3D Convolutional Networks

Liangjie Cao

Aug., 2018

Abstract

They propose a simple, yet effective approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large scale supervised video dataset. Our findings are three-fold: 1) 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets; 2) A homogeneous architecture with small 3 x 3 x 3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets; and 3) 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. In addition, the features are compact: achieving 52.8% accuracy on UCF101 dataset with only 10 dimensions and also very efficient to compute due to the fast inference of ConvNets. Finally, they are conceptually very simple and easy to train and use.

1. Introduction

Multimedia on the Internet is growing rapidly resulting in an increasing number of videos being shared every minute. To combat the information explosion it is essential to understand and analyze these videos for various purposes like search, recommendation, ranking *etc.* The computer vision community has been working on video analysis for decades and tackled different problems such as action recognition [3], abnormal event detection [2], and activity understanding [1]. Considerable progress has been made in these individual problems by employing different specific solutions. However, there is still a growing need for a generic video descriptor that helps in solving large-scale video tasks in a homogeneous way.

2. Learning Features with 3D ConvNets

They believe that 3D ConvNet is well-suited for spatiotemporal feature learning. Compared to 2D ConvNet, 3D ConvNet has the ability to model temporal information better owing to 3D convolution and 3D pooling operations. In 3D ConvNets, convolution and pooling operations are performed spatio-temporally while in 2D ConvNets they are done only spatially. Figure 1 illustrates the difference, 2D convolution applied on an image will output an image, 2D convolution applied on multiple images (treating them as different channels) [4] also results in an image. Hence, 2D ConvNets lose temporal information of the input signal right after every convolution operation. Only 3D convolution preserves the temporal information of the input signals resulting in an output volume.

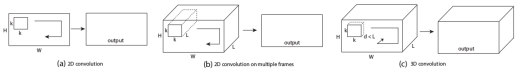

Figure 1. 2D and 3D convolution operations.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014. In *ECCV*, 2014. 1

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1