MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition

Cheng Guan

July 30, 2018

Abstract

From this week, I decide to do some scientific research task with elder student and start to know about some knowledge about video action. Human actions in videos are three-dimensional (3D) signals. Recent attempts use 3D convolutional neural networks (CNNs) to explore spatio-temporal information for human action recognition. Though promising, 3D CNNs have not achieved high performance on this task with respect to their well-established two-dimensional (2D) counterparts for visual recognition in still images. The authors argue that the high training complexity of spatio-temporal fusion and the huge memory cost of 3D convolution hinder current 3D CNNs, which stack 3D convolutions layer by layer, by outputting deeper feature maps that are crucial for high-level tasks. They thus propose a Mixed Convolutional Tube (MiCT) that integrates 2D CNNs with the 3D convolution module to generate deeper and more informative feature maps, while reducing training complexity in each round of spatio-temporal fusion. A new end-to-end trainable deep 3D network, MiCTNet, is also proposed based on the MiCT to better explore spatio-temporal information in human actions.The dataset they use is the three well-known datasets (UCF101,Sport1M and HMDB51).

1. Introduction

Human action recognition is a fundamental yet challenging task with considerable efforts having been investigated for decades. Motivated by the notable-success of convolutional neural networks (CNNs) for visual recognition in still images, many recent works take advantage of deep models to train end-to-end networks for recognizing actions in videos [3, 1, 7, 8, 4], which sig- nificantly outperform hand-crafted representation learning methods [6, 5, 2]

Reconsidering current 3D CNN networks for action recognition, they notice that most of these methods share the same architecture that stacks 3D convolutions layer by layer, as proposed in C3D [30]. Since the spatial and temporal signals get coupled with each other through each 3D convolution, it becomes much more difficult to optimize the network with dozens of such 3D convolution layers because of the exponential growth of the solution space with respect to the case of 2D CNNs. Besides, the memory cost of 3D convolution is too high to be afforded in practice when constructing a deep 3D CNN, which makes the features of the current 3D CNNs usually not deep enough.

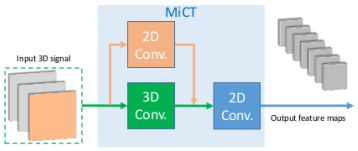

Figure 1. Illustration of their proposed MiCT that integrates 2D CNNs with the 3D convolution for the spatio-temporal feature learning.

In this paper, the authors present a new deep architecture to address this problem and improve the performance of 3D CNNs for action recognition with their proposed Mixed 2D/3D Convolutional Tube (MiCT). The MiCT enables the feature map at each spatio-temporal level to be much deeper prior to the next spatio-temporal fusion, which in turn makes it possible for the network to achieve better per- formance with fewer spatio-temporal fusions, while reduc- ing the complexity of each round of spatio-temporal fusion by using the cross-domain residual connection. In contrast to the 3D CNNs that stack the 3D convolution layer by layer, the proposed MiCT, as shown in Fig.1, integrates 3D CNNs with 2D CNNs to enhance the feature learning with negligible increase in memory usage and complexity.

2. Related Work

There exists an extensive body of literature on human action recognition. Here the authors outline work involving deep features and classify the related work into two categories, 2D CNN and 3D CNN based approaches, according to the convolutions used in feature learning.

References

[1] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, 2016. 1

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[3] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 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

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 1

[6] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1

[8] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 1