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

Cheng Guan

August 1, 2018

1. MiCT and Deep MiCT Network

In 2D CNN, to explore the spatio-temporal informa- tion in human actions, the two-stream architecture is first proposed in [4] where two 2D CNNs are applied to the appearance (RGB frames) and motion (stacked optical flow) domains, respectively. Based on this architecture, several mechanisms are presented to fuse the two networks over the appearance and motion [3, 1, 2]. In this section, the authors start with a brief introduction of the 3D convolution. They then give a detailed description of the proposed MiCT. Lastly, their simple yet efficient deep network, MiCT-Net, is presented for human action recognition.

1.1. 3D Convolution

A 3D spatio-temporal signal, *e.g.* a video clip, can be represented as a tensor with a size of $T \times H \times W \times C$, where T, H, W, C denotes the temporal duration, height and width in the spatial domain, and number of channels, respectively. Kernels of a 3D convolution layer are then formulated as a 4D tensor $\| \in \mathbf{R}^{n_k \times t_k \times h_k \times w_k}$ (they omit the channel dimension hereafter for simplicity), where l_k, h_k, w_k are the kernel size for the T, H, and w dimensions, and n_k denotes the number of kernels. As illustrated in Fig. 1, a 3D convolution layer takes the input 3D spatio-temporal features $\mathbf{V} = \{\mathbf{v}_{t,h,w}\}$ and outputs the 3D dimensional feature map $\mathbf{O} = \{\mathrm{o}_{t,h,w}\}$ by implementing convolution along both the spatial and temporal dimensions of the inputs, which can be formulated as Eq.1

$$
\begin{aligned}
\mathbf{O} &= \mathcal{K} \otimes \mathbf{V}, \quad where \\
\mathbf{o}_{t_0,h_0,w_0} &= \left[q^1_{t_0,h_0,w_0}, q^2_{t_0,h_0,w_0}, ..., q^{n_k}_{t_0,h_0,w_0}\right]^{\mathrm{T}} \\
q^n_{t_0,h_0,w_0} &= \sum_{t,w,h} \mathcal{K}_{n,t,w,h} \cdot \mathbf{V}^{t_0 h_0 w_0}_{t,w,h}
\end{aligned}
\tag{1}
$$

Here $\mathbf{V}^{t_0,h_0,w_0}$ is the sliced tensor that starts from the location (t_0, h_0, w_0) in V and has the same size as the kernel $\mathcal{K}^n$. $q^n_{t_0,h_0,w_0}$ denotes the value at (t_0, h_0, w_0) on the n th feature map output by the n^{th} 3D convolution kernel.

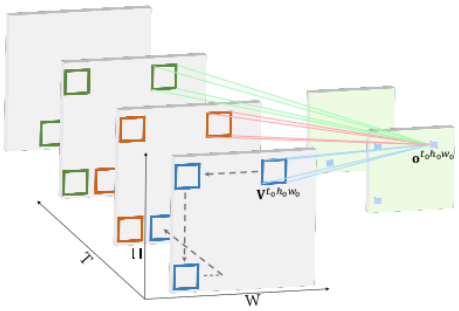

Figure 1. Illustration of a 3D convolution. The convolution kernels slide along both the spatial and temporal dimensions of the input 3D signal and generate the 3D spatio-temporal feature maps.

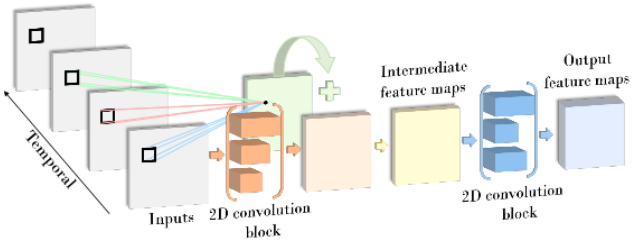

Figure 2. Illustration of MiCT that integrates 2D CNNs into 3D convolution for feature learning. In each MiCT, feature maps generated by the 3D convolutional module (green) are added to the ones produced by the residual 2D convolutional module (orange) on sampled 2D inputs. The combined feature maps are then fed into the concatenated 2D convolutional module (blue) to obtain the final feature maps.

1.2. MiCT

A 3D convolution couples spatio-temporal signals in an effort to effectively extract spatio-temporal features. However, when stacked together to form 3D CNNs, it also increases the difficulty of optimization, hinders 3D CNNs from generating deeper feature maps for high-level tasks due to unaffordable memory usage and high computational cost, and raises the demand on huge training sets. All these facts together limit the performance of current existing 3D CNNs on action recognition. In order to address

these problems, we propose introducing 2D CNNs, which can be trained effectively, constructed deeply, and learned with huge datasets, to 3D convolution modules and form a new 3D convolution unit MiCT to empower feature learning, as illustrated in Fig.2.

References

[1] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1

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

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[4] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1