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

Cheng Guan

August 3, 2018

1. Mixed Convolution Tube

1.1. Concatenating Connections

As shown in Fig.1, it illustrates the concatenated connection of 2D and 3D convolutions in the MiCT. We use MiCT$_c on$ to represent the MiCT with only the concatenate connection hereafter. Denoting the feature map $\mathbf{O}$ at time t as $\mathbf{O}^t$, as shown in Eq.1

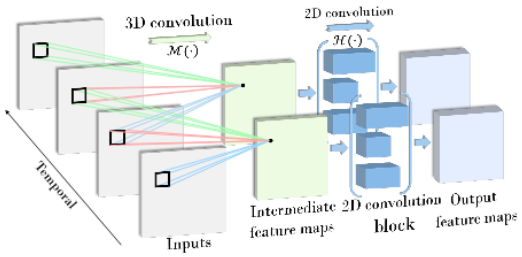

Figure 1. MiCT with concatenated connections. For an input 3D signal, the 3D convolution fuses spatio-temporal information and obtains intermediate feature maps which are fed into the 2D convolution block to generate the final feature maps.

$$\begin{aligned}\mathbf{O}^t =& \mathcal{M}\left(\mathbf{V}^t\right) \\ =& \mathcal{K} \otimes \mathbf{V}^t\end{aligned} \tag{1}$$

where $\mathbf{V}^t \in \mathbf{R}^{l_k \times h \times w}$ is the sliced tensor from time t to time $t + l_k$. Since $\mathcal{M}(\cdot)$ only outputs linearly fused spatio- temporal feature maps based on Eq.1, a 3D CNN has to stack enough of $\mathcal{M}(\cdot)$ for deep and high-level feature maps which requires dynamically increased memory usage, training samples, and training complexity. They thus propose enhancing $\mathcal{M}(\cdot)$ by a deeper and capable alternative $\mathcal{G}(\cdot)$ to extract much deeper features during every round of spatio-temporal fusion. $\mathcal{G}(\cdot)$ is supposed to meet three requirements. It should be computationally efficient, support end-to-end training, and be capable of feature learning for 2D and 3D signals. To meet these requirements, they design the function $\mathcal{G}(\cdot)$ by concatenating 2D CNNs after the 3D convolution to provide a very efficient deep feature extractor, denoted as Eq.2

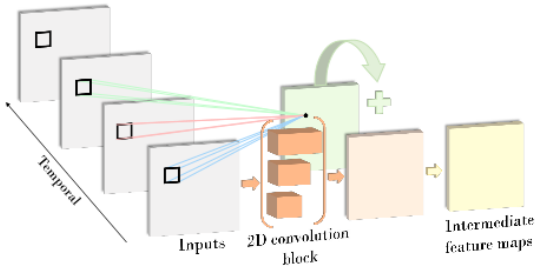

Figure 2. MiCT with a cross-domain residual connection. Spatio-temporal fusion is achieved by both the 2D convolution block to generate stationary features and 3D convolution to extract temporal residual information.

$$\mathcal{G}(\cdot) = \mathcal{H}(\mathcal{M}(\cdot)) \tag{2}$$

1.2. Cross-Domain Residual Connections

The MiCT with only a cross-domain residual connection, denoted as MiCT res , is illustrated in Fig.2. It introduces a 2D convolution between the input and output of the 3D convolution to further reduce spatio-temporal fusion com- plexity and facilitate the optimization of the whole network. Following the notations in Eq.1, They have the Eq.3

$$\begin{aligned}&\mathbf{o}'_{t_0,h_0,w_0} = \mathbf{o}_{t_0,h_0,w_0} + \mathbf{S}^{t_0}_{h_0,w_0} \\ &where \quad \mathbf{S}^{t_0} = \mathcal{H}'\left(\mathbf{V}^{t_0}\right)\end{aligned} \tag{3}$$

Here $\mathbf{V}^{t_0} \in \mathbf{R}^{h \times w}$ is the sliced tensor of input $\mathbf{V}$ at time t_0 , $\mathbf{S}^{t_0}_{h_0,w_0}$ refers to the value at (h_0, w_0) on $\mathbf{S}^{t_0}$ obtained by $\mathcal{H}'(\cdot)$, and $\mathcal{H}'(\cdot)$ denotes a 2D convolution block. Unlike the residual connections in previous work [2, 1], the shortcut in their scheme is cross-domain, where spatio-temporal fusion is derived by both a 3D convolution mapping with respect to the full 3D inputs and a 2D convolution block mapping with respect to the sampled 2D inputs. They propose a cross-domain residual connection based on the observation that a video stream usually contains lots of redundant information among consecutive frames, resulting in redundant information in feature maps along the temporal dimen-

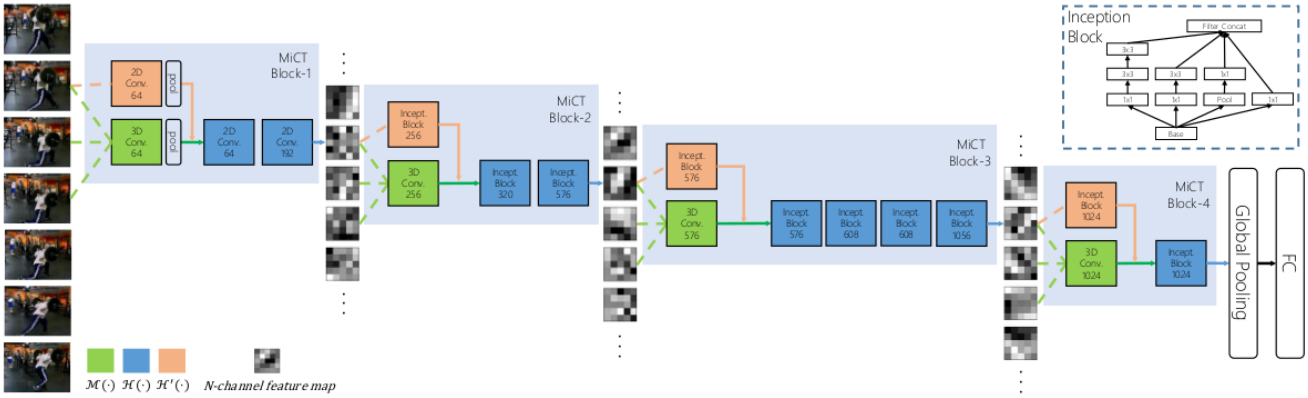

Figure 3. Illustration of the proposed MiCT-Net. Green blocks refer to 3D convolution. Orange blocks and blue blocks refer to 2D convolutions for cross-domain residual connections and concatenated connections, respectively. Each mosaic-like box denotes an n-channel feature map at time t (n=192, 576, 1056 in MiCT Block-1/2/3, respectively). The architecture details of each Incept. block are shown in the top-right area of the figure.

Table 1. Architecture of MiCT-Net.

Type	$\mathcal{M}(\cdot)$	$\mathcal{H}(\cdot)$	$\mathcal{M}(\cdot)$
block-1	3x7x7x64/$(1,2)$	1x1x64/1	7x7x64/2
block-2	3x3x3x256/$(2,1)$	2xInception	1xInception
block-3	3x3x3x576/$(2,1)$	2xInception	1xInception
block-4	3x3x3x1024/$(2,1)$	1xInception	1xInception
pooling	global pooling on spatial dimension		
fc	$1024 \times$ *num classes*		
pooling	global pooling		

sion. Their proposed MiCT combines the two connections and achieves the best performance among the three configurations MiCT con , MiCT res , and MiCT.

2. Deep MiCT Network

The authors propose a simple yet efficient deep MiCT Network (MiCT-Net in short) by stacking the MiCT together. The MiCT-Net takes the RGB video sequences as inputs and is end-to-end trainable. As shown in Fig.3, it consists of four MiCTs, which means only four 3D convolutions are em- ployed. For the 2D convolution blocks in each MiCT block, we partially follow the designs of BN-inception [3]. More details of the network architecture are provided in Table.1.

References

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

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

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2