MixTConv: Mixed Temporal Convolutional Kernels for Efficient Action Recognition

Kaiyu Shan, Yongtao Wang, Zhi Tang, Ying Chen and Yangyan Li

Wangxuan Institute of Computer Technology, Peking University Alibaba Cloud Intelligence Business Group

Email: {shankyle, wyt, tangzhi}@pku.edu.cn {chenying.ailab, yangyan.lyy}@alibaba-inc.com

1. Introduction

To efficiently extract spatiotemporal features of video for action recognition, most state-of-the-art methods integrate 1D temporal convolutional filters into 2D CNN backbones. However, they all exploit 1D temporal convolutional filters of fixed kernel size (i.e., 3) in their network building block, thus have suboptimal temporal modeling capability to handle both long-term and short-term actions. To address this problem, we first investigate the impacts of different kernel sizes for the 1D temporal convolutional filters. Then, we propose a simple yet efficient operation called Mixed Temporal Convolution (MixTConv), which consists of multiple depthwise 1D convolutional filters with different kernel sizes. By plugging MixTConv into the conventional 2D CNN backbone ResNet-50, we further propose an efficient and effective network architecture named MSTNet for action recognition, and achieve state-of-the-art results on multiple large-scale benchmarks.

2. Intuition

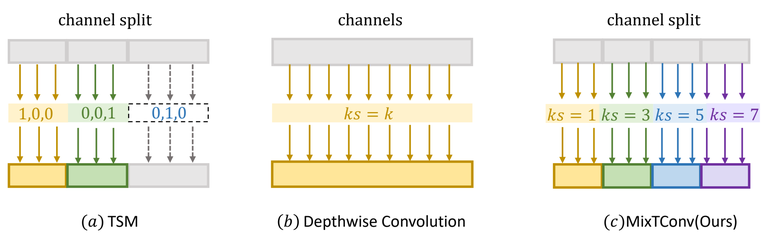

Figure 1: Our method is related to the current state-of-the-art method TSM[11]. In fact, the *shift* operation is a special case of our proposed MixTConv, more specifically, equal to a *fixed weight* depthwise 1D convolution with *fixed kernel size* of 3 (bi-directional shift), where temporal kernel is fixed as: [0, 1, 0] for *static channels*(3/4 of total channels), [1, 0, 0] for *backward-shift channels*(1/8 of total channels), and [0, 0, 1] for *forward-shift channels*(1/8 of total channels), shown in Figure (a). Our experiment shows that, using depthwise 1D convolution with *learnable weight* (Figure (b)) and *multiple kernel sizes* (Figure (c)) along the temporal dimension is more effective than these handcrafted temporal kernels to capture pyramidal temporal contextual information.

3. Method

The pipeline of our method is shown as follows:

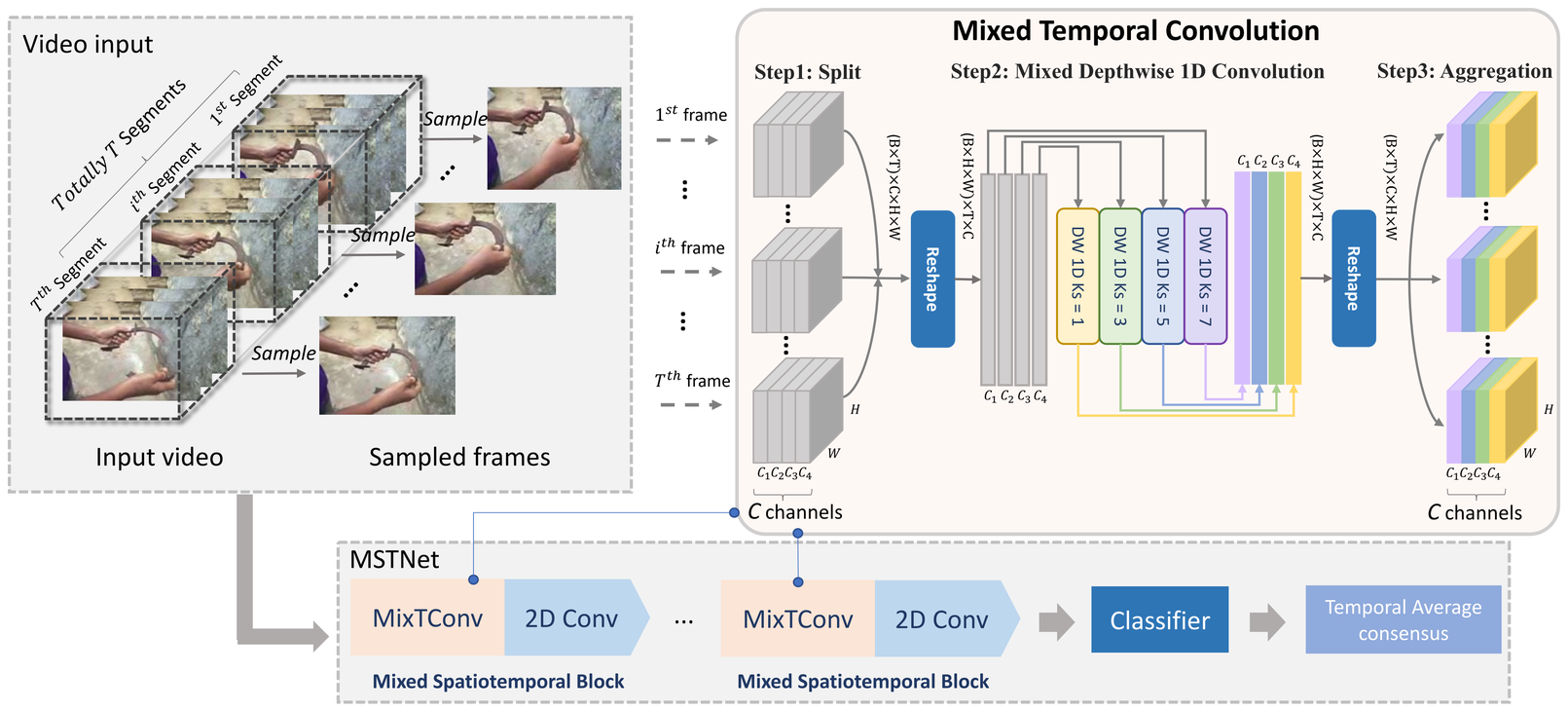

Figure 2: The pipeline of the proposed video action recognition network Mixed Spatiotemporal Network(*MSTNet*), based on the Mixed Temporal Convolution. "Ks" means kernel size, and "DW" means depthwise.

We denote the input feature map for MixTConv operation as $F \in \mathbb{R}^{(B\times T)\times C\times H\times W}$, where B, H, W, T, C is the batch size, height, weight, number of sampled frames and channel size. As illustrated in Figure 2, we firstly reshape F as: $F \in \mathbb{R}^{(B\times H\times W)\times C\times T}$, and then apply the depthwise 1D convolution with g different kernel sizes $\{k_1, ..., k_g\}$ on the temporal dimension. Let W_m denote a depthwise 1D convolutional kernel with kernel size of k_m. Unlike vanilla depthwise convolution, MixTConv partitions channels into g groups $\{\hat{F}^1, ..., \hat{F}^g\}$ and applies depthwise 1D convolution with different kernel sizes to each group, where c_m denotes channels in the *m-th* group. Formally, the mixed 1D convolution is defined as:

$$\hat{Z}^m_{i,t} = \sum_j \hat{F}^i_{t+j} W_{\frac{k_m-1}{2}+j}, m = 1, ..., g, \tag{1}$$

where $j \in [-\frac{k_m-1}{2}, \frac{k_m-1}{2}]$ and $\hat{Z}^m_{i,t}$ is the value of $\hat{Z}^m$ at the *t-th* frame and *i-th* channel.
The final output tensor is a concatenation of all the output tensor $\{\hat{Z}^1, ..., \hat{Z}^g\}$:

$$Z = Concat(\hat{Z}^1, ..., \hat{Z}^g). \tag{2}$$

4. Experiments

Comparison with the State-of-the-Art We compare MSTNet with state-of-the-art methods on Something-Something v1 and v2 in Table 3. The comparison details are as follows:

Method	Backbone	Modality	Frames	Params	FLOPs	Something-Something v1		Something-Something v2	
						Val Top-1	Val Top-5	Val Top-1	Val Top-5
TSN[3]ECCV'16	BNIception	RGB	8	10.7M	16G	19.5	-	-	-
TSN(baseline)[3]ECCV'16	ResNet-50	RGB	8	24.3M	33G	19.7	46.6	27.8	57.6
TRN Multiscale[4]ECCV'18	BNInception	RGB	8	18.3M	16G	34.4	-	44.8	77.6
TRN Two-steam[4]ECCV'18	BNInception	RGB+Flow	8+8	36.6M	-	42.0	-	55.5	83.1
I3D[6]CVPR'17	3D ResNet-50	RGB	32×2clips	28.0M	153G×2	41.6	72.2	-	-
NL*+I3D[22]CVPR'18	3D ResNet-50	RGB	32×2clips	35.3M	168G×2	44.4	76.0	-	-
NL*+I3D+GCN[23]ECCV'18	3D ResNet-50+GCN	RGB	32×2clips	62.2M	303G×2	46.1	76.8	-	-
ECO[24]ECCV'18	BNInc*+Res3D18*	RGB	8	47.5M	32G	39.6	-	-	-
ECO[24]ECCV'18	BNInc*+Res3D18*	RGB	16	47.5M	64G	41.4	-	-	-
ECO_{En}*Lite*[24]ECCV'18	BNInc*+Res3D18*	RGB	92	150M	267G	46.4	-	-	-
TSM[11]ICCV'19	ResNet-50	RGB	8	24.3M	33G	45.6	74.2	58.7†	85.4
TSM[11]ICCV'19	ResNet-50	RGB	16	24.3M	65G	47.2	77.1	61.0†	86.8
Ours:									
MSTNet	ResNet-50	RGB	8	24.3M	33.2G	46.7	75.4	59.5	86.0
MSTNet	ResNet-50	RGB	16	24.3M	65.3G	**48.4**	**78.8**	**61.8**	**87.3**

*BNInc means BNInception, *Res3D18 means 3D Resnet 18, *NL means Non-Local[22].
†Using official released pre-trained weight and testing with one clip and center crop.

Figure 3: Comparisons with state-of-the-art methods on Something-Something v1 and Something-Something v2.

5. References

[1] R. Goyal and et al., "The "something something" video database for learning and evaluating visual common sense," in *ICCV*, 2017.

[2] A. Karpathy, G. Toderici, and et al., "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[3] W. Limin, X. Yuanjun, and et al., "Temporal segment networks for action recognition in videos," *TPAMI*, 2018.

[4] B. Zhou, A. Andonian, and et al., "Temporal relational reasoning in videos," in *ECCV*, 2018.

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

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017.

[7] D. Tran, L. D. Bourdev, and et al., "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[8] S. Xie, C. Sun, and et al., "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018.

[9] D. Tran, H. Wang, and et al., "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

[10] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017.

[11] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[12] K. He, X. Zhang, and et al., "Deep residual learning for image recognition," in *CVPR*, 2016.

[13] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset