MixTConv: Mixed Temporal Convolutional Kernels for Efficient Action Recognition

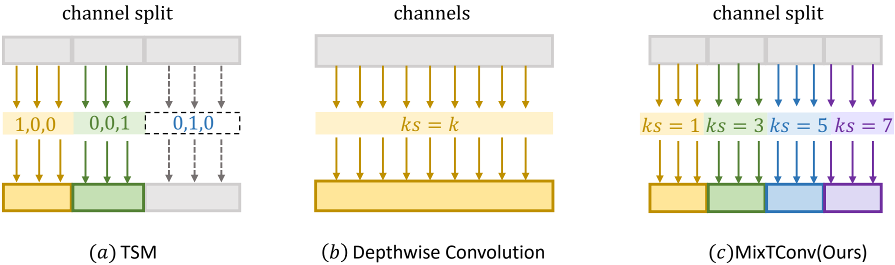
Kaiyu Shan, Yongtao Wang, Zhi Tang, Ying Chen and Yangyan Li

Wangxuan Institute of Computer Technology, Peking University
Email: {shankyle, wyt, tangzhi}@pku.edu.cn
{chenying.ailab, yangyan.lyy}@alibaba-inc.com

January 10, 2020

Outline

Different Temporal Operations



(a)

Figure: Comparison of different temporal operations. (a)*shift* temporal operation with *fixed kernel weight and kernel size.* (b) *learnable* temporal operation with the *fixed kernel size* of depthwise 1D convolution. (c) *Mixed Temporal Convolution*(MixTConv) with different kernel sizes of depthwise 1D convolution.

Outline

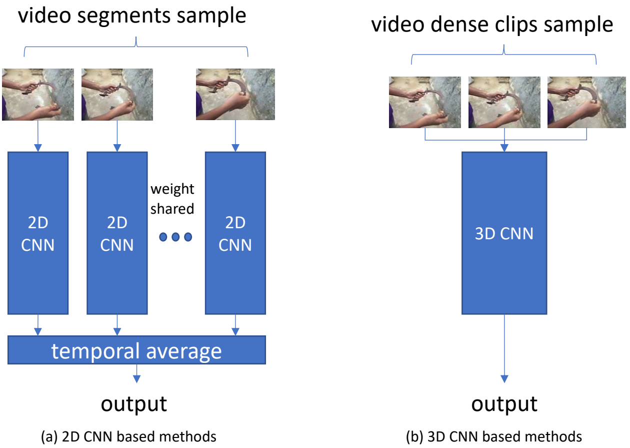
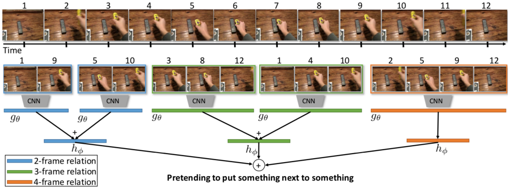
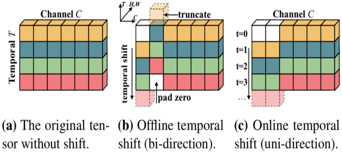


Figure: (a)2D CNN-based methods divide the video into N segments and samples one frame from each segment, then consensus the result by averaging. (b)3D CNN-based methods jointly learn spatiotemporal features in an elegant way.



(a)TRN



(b)TSM

Figure: (a)TRN adds temporal fusion after feature extraction, leading to limited improvement of performance. (b) TSM utilizes *shifting operation* which shifts a portion of the channels along the temporal dimension.

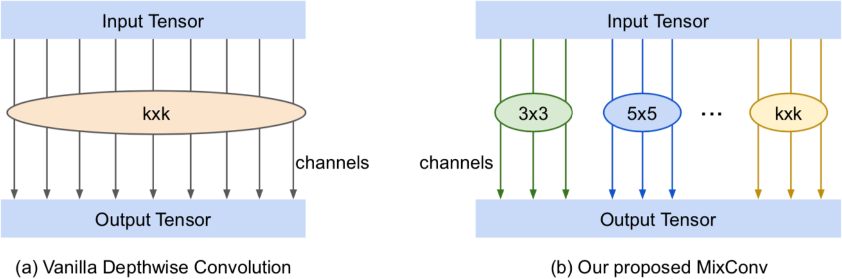


Figure: MixConv uses 2D spatial convolution filters of different kernel sizes to extract spatial features of various resolutions, for improving *image recognition* accuracy.

Outline

MixTConv: Mixed Temporal Convolution

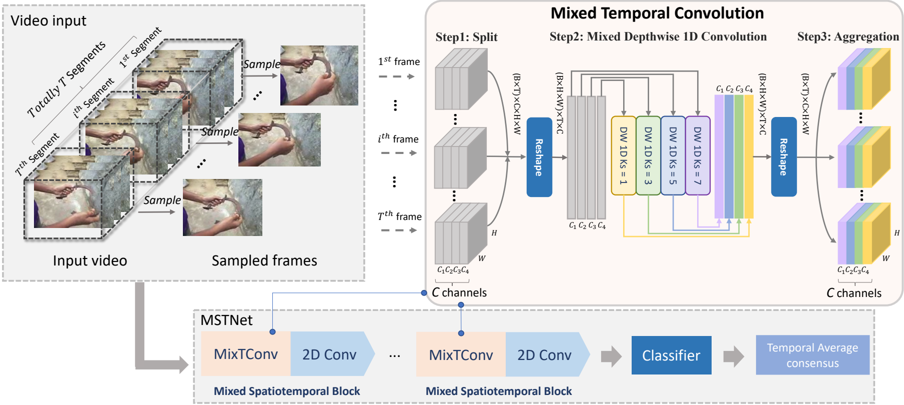


Figure: The pipeline of the proposed video action recognition network Mixed Spatiotemporal Network(*MSTNet*), based on the Mixed Temporal Convolution. "Ks" means kernel size, and "DW" means depthwise.

MixTConv: Mixed Temporal Convolution

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

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 $\{\hat{Z}^{1}, ..., \hat{Z}^{g}\}$:

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

Mixed Spatiotemporal Block

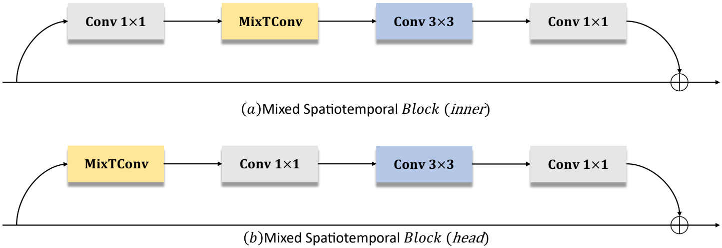


Figure: Comparision for MST Block *head* and MST Block *inner*.

Outline

Datasets

Kinetics-400[14] is a large-scale dataset with 400 classes sourced from YouTube, and is one of the most popular action recognition benchmarks.
UCF-101[15] is an action recognition dataset of realistic action videos, collected from YouTube, having 101 action categories.
Something-Something v1 and *v2* [1] are two large-scale video datasets for action recognition.
Jester[13] is a large collection of densely-labeled video clips that show humans performing pre-defined hand gestures in front of a laptop camera or webcam.

Comparision of 2D CNN Baseline

Table: Comparisons between the proposed MSTNet and 2D CNN baseline TSN.

Dataset	Model	MixTConv	Top-1	Top-5	Δ Top-1
Kinetics-400	TSN[3]	✗	68.8	88.3	+2.5
	Ours	✓	**71.3**	**89.5**	
UCF-101	TSN[3]	✗	91.5	99.2	+3.3
	Ours	✓	**94.8**	**99.6**	
Something v1	TSN[3]	✗	20.5	47.5	+27.6
	Ours	✓	**48.1**	**77.3**	
Something v2	TSN[3]	✗	30.4	61.0	+31.4
	Ours	✓	**61.8**	**87.8**	
Jester	TSN[3]	✗	83.9	99.6	+13.0
	Ours	✓	**96.9**	**99.9**	

Comparision of kernel sizes

Table: Comparisons of different temporal operations and configurations (i.e., the kernel size and the combinations of the filters) on Something-Something v1. "ks"denotes kernel size and * denotes shifting convolution.

Method	**Kernel Size**	**Dilation**	**Learnable**	**Top-1**	**FLOPS**
TSN(baseline)[3]	-	-	✘	19.7	33G
TSN+Ordinary 1D	3	1	✓	41.0	43G
TSM*[11]	3*	1	✘	45.6	33G
TSN+ks3	3	1	✓	45.9	33.13G
TSN+ks5	5	1	✓	46.3	33.23G
TSN+ks7	7	1	✓	45.8	33.32G
TSN+ks13	1,3	1	✓	45.8	33.09G
TSN+ks135	1,3,5	1	✓	46.4	33.13G
TSN+ks1357	1,3,5,7	1	✓	**46.7**	33.18G
TSN+ks357	3	1,2,3	✓	46.4	33.13G

Something-Something v1 and v2

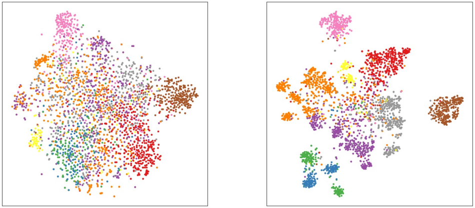
Method	Backbone	Modality	Frames	Params	FLOPs	Something-Something v1		Something-Something v2	
						Val Top-1	Val Top-5	Val Top-1	Val Top-5
TSN[3]ECCV'16	BNIception	RGB	8	10.7M	16G	19.5	-	-	-
TSN(baseline)[3]ECCV'16	ResNet-50	RGB	8	24.3M	33G	19.7	46.6	27.8	57.6
TRN Multiscale[4]ECCV'18	BNInception	RGB	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	$58.7^{\dagger}$	85.4
TSM[11]ICCV'19	ResNet-50	RGB	16	24.3M	65G	47.2	77.1	$61.0^{\dagger}$	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: Comparisons with state-of-the-art methods on Something-Something v1 and Something-Something v2.

Outline

Visualization



(a) TSN (b) MSTNet

Figure: t-SNE plots of the output layer features preceding the final fully connected layers for (a) TSN, and for MSTNet(b) on Something-Something v1.

Outline

Conclusions

In this work, we propose a lightweight and plug-and-play operation named Mixed Temporal Convolution (MixTConv) for action recognition, which partitions input channels into groups and performs depthwise 1D convolution with different kernel sizes to capture multi-scale temporal information. It can be flexibly inserted into any 2D CNN backbones to enable temporal modeling with negligible extra computational cost. We further design a Mixed Spatiotemporal Network (MSTNet) for action recognition, by plugging MixTConv into the building block of ResNet-50. Experimental results on Something-Something v1, v2 and Jester benchmarks consistently indicate the superiority of the proposed MSTNet with the MixTConv operation. Additional ablation studies further demonstrate that the designs of the proposed MixTConv operation and MSTNet are effective and reasonable.

References

- R. Goyal and et al., “The ”something something” video database for learning and evaluating visual common sense,” in *ICCV*, 2017.
- A. Karpathy, G. Toderici, and et al., “Large-scale video classification with convolutional neural networks,” in *CVPR*, 2014.
- W. Limin, X. Yuanjun, and et al., “Temporal segment networks for action recognition in videos,” *TPAMI*, 2018.
- B. Zhou, A. Andonian, and et al., “Temporal relational reasoning in videos,” in *ECCV*, 2018.
- C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *CVPR*, 2016.
- J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *CVPR*, 2017.
- D. Tran, L. D. Bourdev, and et al., “Learning spatiotemporal features with 3d convolutional networks,” in *ICCV*, 2015.