PaperID 5152

MULTI-DIRECTIONAL CONVOLUTION NETWORKS WITH SPATIAL-TEMPORAL FEATURE PYRAMID MODULE FOR ACTION RECOGNITION

Bohong Yang[1], Zijian Wang[1], Wu Ran[1], Hong Lu[1], Yi-Ping Phoebe Chen[2]

[1]Shanghai Key Laboratory of Intelligent Information, Fudan University, Shanghai, P.R.China

[2]Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia

ABSTRACT -- Recent attempts show that factorizing 3D convolutional filters into separate spatial and temporal components brings impressive improvement in action recognition. However, traditional temporal convolution operating along the temporal dimension will aggregate unrelated features, since the feature maps of fast-moving objects have shifted spatial positions. In this paper, we propose a novel and effective Multi-Directional convolution (MDConv), which extracts features along different spatial-temporal orientations. Especially, MDConv has the same FLOPs and parameters as the traditional 1D temporal convolution. Also, we propose the Spatial-Temporal Feature Pyramid Module (STFPM) to fuse spatial semantics in different scales in a light-weight way. Our extensive experiments show that the models which integrate with MDConv achieve better accuracy on several large-scale action recognition benchmarks such as Kinetics, Something-Something V1&V2 and AVA datasets.

Multi-Directional Comvolution

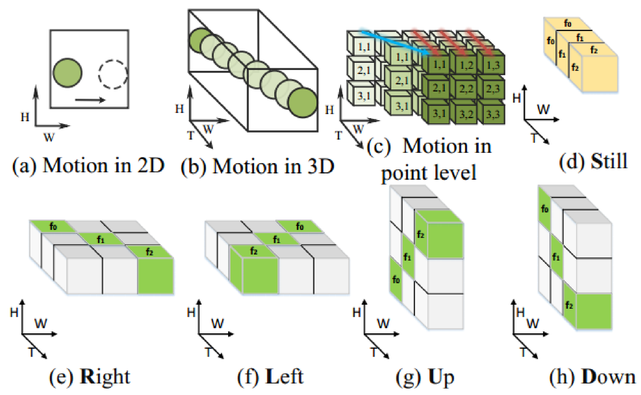

Fig. 1. An example of moving object in spatio-temporal state and our proposed Multi-Directional Convolution (MDConv). *T* , *H*, *W* represent the time dimension and the height and width dimensions of the video.
(a), (b) and (c) are the states of moving circular in spatio-temporal space. The traditional spatial-temporal convolution (d) can not recognize the trajectory of the object (the red arrows in (c)). And our proposed MDConvs (e)-(h) can recognize this motion (the blue arrow in (c)).

Our proposed Multi-Directional Convolution (MDConv) is defined as follow:

$$\mathcal{R}_{mul}\begin{cases}\mathcal{R}_{right} = \{(t_0+\Delta t, h_0, w_0+\gamma\Delta t)|\Delta t \in \mathcal{T}\} \\ \mathcal{R}_{left} = \{(t_0+\Delta t, h_0, w_0-\gamma\Delta t)|\Delta t \in \mathcal{T}\} \\ \mathcal{R}_{up} = \{(t_0+\Delta t, h_0+\gamma\Delta t, w_0)|\Delta t \in \mathcal{T}\} \\ \mathcal{R}_{down} = \{(t_0+\Delta t, h_0-\gamma\Delta t, w_0)|\Delta t \in \mathcal{T}\} \\ \mathcal{R}_{still} = \{(t_0+\Delta t, \quad h_0, \quad w_0 \quad)|\Delta t \in \mathcal{T}\}\end{cases} \quad \mathcal{R}_{Move} = \mathcal{R}_{mul} - \{\mathcal{R}_{still}\}.$$

$$y_{dc}(p) = \sum_{p_i \in \mathcal{R}_{dc}} w(p_i)\cdot x(p_i), \quad \mathcal{R}_{dc} \in \mathcal{R}_{Move}$$

$$y_{output} = [y_{right}, y_{left}, y_{up}, y_{down}]$$

ARCHITECHTURE

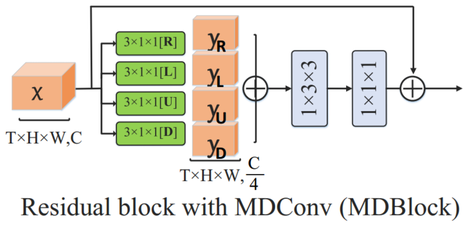

Fig. 2. Our proposed MDBlock. We replace the spatial-temporal convolution with four MDConvs, concatenate the outputs together, and obtain the feature maps with the same size of spatial-temporal convolution.

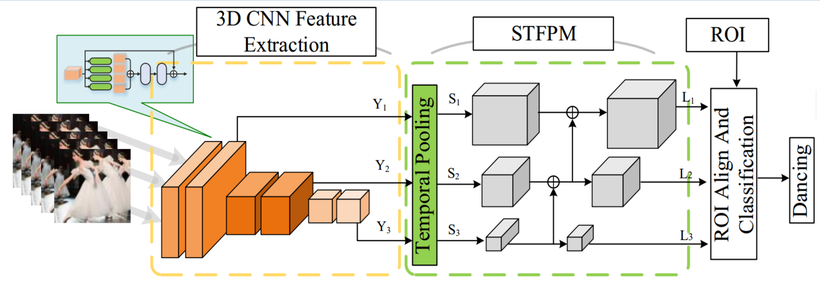

Fig. 3. The overview architecture of the proposed action detection networks. The networks consists of 3D CNN feature extraction backbone, STFPM and RoIAlign and classifier. Each block of backbone is replaced by our proposed MDBlock which is shown in Fig. 2. The STFPM aggregates multi-scale feature.

EXPERIMENT ON ACTION CLASSIFICATION

Method	Frames	MD-Conv	Something V1 Top-1 Acc	Something V1 Top-5 Acc	Something V2 Top-1 Acc	Something V2 Top-5 Acc
I3D [1]	32	×	41.6	72.2	-	-
		✓	43.4	75.1	-	-
SlowOnly 8 × 4 [3]	4	×	38.6	68.9	51.9	81.2
		✓	**40.1**	**71.0**	**53.8**	**82.5**
SlowFast 8 × 4 [3]	16+4	×	46.8	76.4	59.6	86.3
		✓	**49.3**	**79.1**	**61.3**	**88.5**

Table 1. Performance of different 3D CNN methods with/without MDConv on Kinetics-400.

Method	Backbone	Pretrain	Frames	MD-Conv	Top-1 Acc	Top-5 Acc
I3D [1]	BN-Inception	ImageNet	64	×	71.1	89.3
				✓	**72.4**	**90.9**
R(2+1)D [2]	ResNet-34	none	32	×	72.0	90.0
				✓	**73.3**	**91.0**
SlowOnly 16×4 [3]	ResNet-50	none	4	×	72.6	90.3
				✓	**72.8**	**90.4**
SlowFast 16×4 [3]	ResNet-50	none	32+4	×	75.6	92.1
				✓	**76.3**	**92.4**
SlowFast 8×8 [3]	ResNet-50	none	32+8	×	77.0	92.6
				✓	**77.6**	**92.9**

Table 2. Performance of different 3D CNN method with/without MDConv on Something-Something V1&V2. All methods are pretrained on Kinetics-400 with backbone ResNet-50.

EXPERIMENT ON ACTION DETECTION

Method	Backbone	Pretrain	MDConv	STFPM	mAP
I3D [1]	ResNet-50	ImageNet	×	×	14.5
			✓	×	15.6
			✓	✓	**16.3**
SlowFast 16×4 [3]	ResNet-50	Kinetics-400	×	×	24.9
			✓	×	25.4
			✓	✓	**26.0**
SlowFast 8×8 [3]	ResNet-50	Kinetics-400	×	×	26.3
			✓	×	26.9
			✓	✓	**27.7**

Table 3. Performance of different 3D CNN methods with/without MDConv and STFPM on AVA dataset.

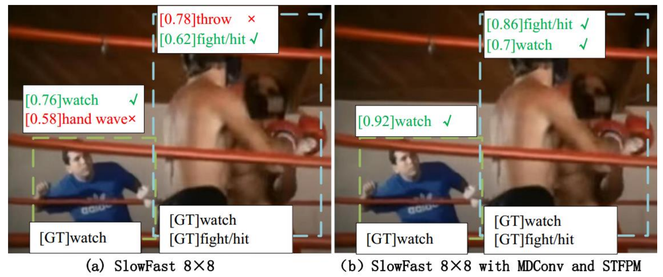

Fig. 4. Comparison of results of action localization by SlowFast models and SlowFast-MDConv-STFPM. Adding modules improve the accuracy.

CONCLUSIONS -- In this paper we propose the Multi-Directional Convolution (MDConv), which operates simultaneously on the spatial and temporal dimensions. We also propose the Spatial-Temporal Feature Pyramid Module (STFPM) to fuse spatial semantics in different scales. With MDConv and STFPM, the improved model can encode features by crossing spatialtemporal boundaries. And our extensive evaluations show that the proposed method has achieved state-of-the-art accuracy on action classification datasets and action detection datasets.

REFERENCE

[1] Joao Carreira and Andrew Zisserman, "Quo Vadis, Action ˜Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[2] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *2019 IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.