MFI: Multi-range Feature Interchange for Video Action Recognition

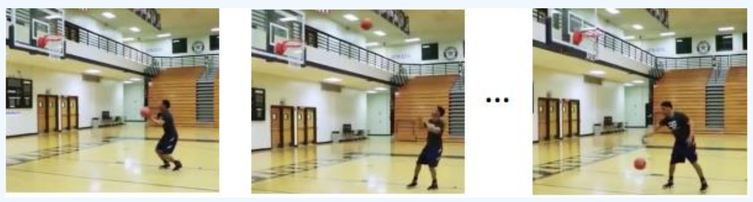
Sikai Bai, Qi Wang and Xuelong Li
Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

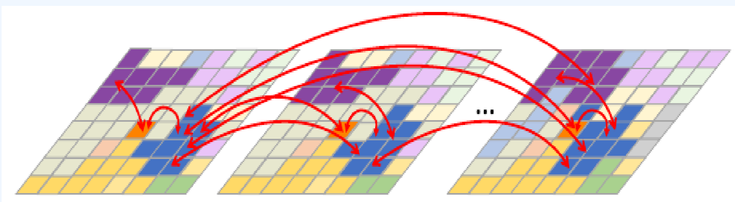
Introduction

Motivation:

- Video action recognition is a fundamental yet challenging task in the field of computer vision.



- Short-range motion features and long-range dependencies are two complementary and vital cues for action recognition in videos.
- It is still unclear how to capture temporal information with complex evolution on multiple ranges using an efficient and effective way.



- Feature interchange: the features from the colored regions bi-directionally shift in the feature map of video models.

Contribution:

- Perform channel-wise temporal interchange (CTI) along the temporal dimension to effectively encode short-range motion features.
- Construct graph-based regional interchange (GRI) module to learn efficiently long-range dependencies using graph convolution.
- Propose a novel multi-range feature interchange (MFI) network to integrate the proposed two modules. Achieves competitive results by using very limited computing cost.

Network Architecture

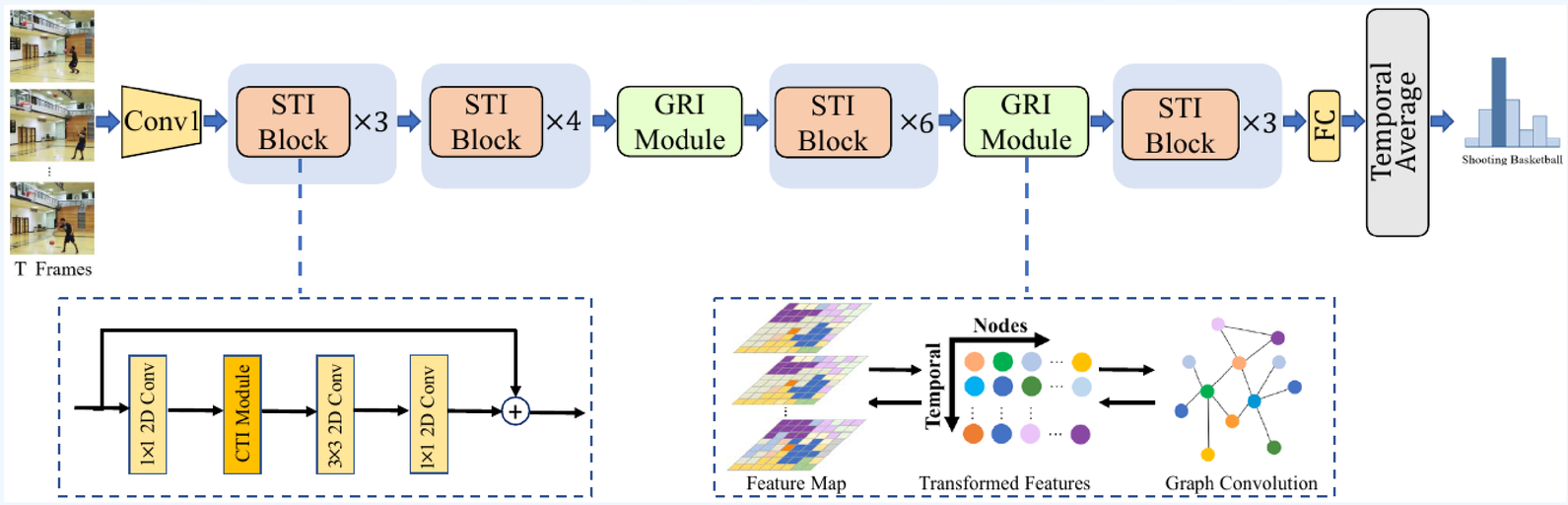


Fig. 2: The overview architecture of Multi-range Feature Interchange Network for video action recognition.

Following [1], T sampled frames are obtained from a video as the input of the network. 2D ResNet-50 is utilized as the backbone, and all original bottleneck blocks are replaced by the proposed STI blocks. We also insert two GRI modules between middle and top STI blocks. The global temporal pooling is applied to average action predictions for all of the sampled frames.

Architecture Details

- Channel-wise Temporal Interchange (CTI) Module
  - The temporal difference can be obtained by calculating the difference between the features of two consecutive frames.

$$H_c^T = Conv_{trans} \otimes Y_c^{t+1} - Y_c^t, \qquad t \in [1, T-1].$$

  - Temporal interchange operation.

$$H_{ic}^t[h,w,c] = H^{t+1}[h,w,c], \quad t \in [1, T-1], c \in [0, C/8r],$$
$$H_{ic}^t[h,w,c] = H^{t-1}[h,w,c], \quad t \in [2, T], c \in [C/8r, C/4r],$$
$$H_{ic}^t[h,w,c] = H^{t}[h,w,c], \quad t \in [1, T], c \in [C/4r, C/r].$$

- Channel-wise Temporal Interchange (CTI) Module
  - Transform from the features in a regular feature map to the state of nodes in a non-grid graph.

$$W_t = [Conv'_{trans} \otimes \Phi_r(X)]^T, \qquad W_t \in R^{N \times L},$$
$$V_t = W_t * \Phi_r(X), \qquad V_t \in R^{N \times C}.$$

  - Graph Convolutional Operation. The nodes propagate their state with each other.

$$V_{out} = ReLU(F(V_t, A_g, W_g) + V_t)$$

  - Reverse the output into the regular feature maps to be compatible with CNN models.

$$Y_{inv} = \varphi_r(W_t^T * V_{out})$$

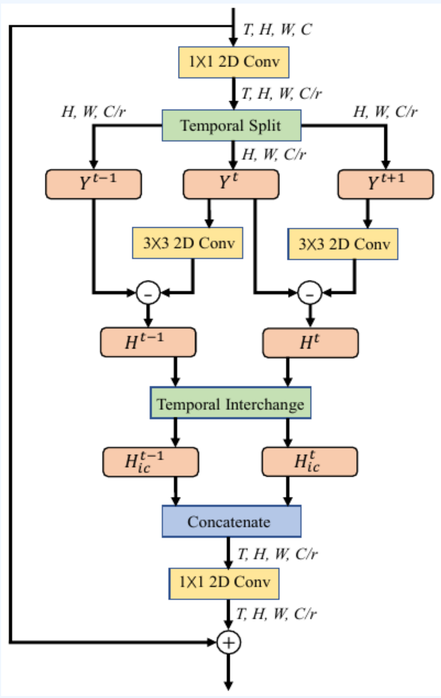


Fig. 3: The architecture of the channel-wise temporal interchange module.

References

[1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in ECCV, 2016, pp. 20–36.
[2] Y. Yuan, D. Wang, and Q. Wang, "Memory-augmented temporal dynamic learning for action recognition," arXiv preprint arXiv:1904.13080, 2019.
[3] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in CVPR, 2019, pp. 7083–7093.
[4] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Subgraph localization for temporal action detection," in CVPR, 2020, pp. 10 156–10 165.
[5] W. Zhang, J. Cen, and H. Zheng, "Temporal inception architecture for action recognition with convolutional neural networks," in ICPR, 2018, pp. 3216–3221.
[6] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in ECCV, 2018, pp. 803–818.
[7] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in ECCV, 2018, pp. 695–712.
[8] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in CVPR, 2018, pp. 6546–6555.

Contact Info

WeChat

Github

name: Sikai Bai （白思开）
Email: whitesk1973@gmail.com

Experiments

- Benchmark Comparison

Table 1: The comparison of performance on Something-Something V1.

Method	Backbone	#Frames	FLOPs	Val-Top1 (%)	Val-Top5 (%)
TSN	BNInception	8	16G	19.5	-
TSN	ResNet-50	8	33G	19.7	46.6
MultiScale TRN	BNInception	8	16G	34.4	-
TSM	ResNet-50	8	33G	43.4	73.2
TSM	ResNet-50	16	33G	44.8	74.5
ECO_{8f}	BNInception+3D ResNet18	8	32G	39.6	-
ECO_{16f}	BNInception+3D ResNet18	16	64G	41.4	-
I3D	3D ResNet50	32 × 2	153G × 2	41.6	72.2
Non-Local-I3D	3D ResNet50	32 × 2	168G × 2	44.4	76.0
MFI(Ours)	ResNet-50	8	33.6G	43.9	73.9
MFI(Ours)	ResNet-50	16	67.2G	45.5	76.0

Table 2: The comparison on UCF101 and HMDB51.

Method	#Frames	UCF101	HMDB51
Two-stream CNN	16+16	88.0	59.4
Two-stream TSN	8+8	94.2	69.6
StNet	7	93.5	-
TSM	8	94.5	70.7
ECO	92	93.6	68.0
STC-ReNeXt101	16	93.7	70.5
ARTNet	16	94.3	70.9
I3D-RGB	64	95.4	74.8
Two-steam I3D	64+64	98.0	80.7
MFI(Ours)	8	94.9	71.9
MFI(Ours)	16	95.6	73.3

- Ablation Study

Table 3: Efficiency Analysis of different methods.

Model	#Frames	FLOPs	Param.	Acc.(%)
TSN	8	33G	24.3M	19.7
	16	66G	24.3M	19.9
ECO	16	64G	47.5M	41.4
I3D	32	306G	28.0M	41.6
TSM	8	33G	24.3M	43.4
	16	36G	24.3M	44.8
MFI	8	33.6G	24.6M	43.9
	16	67.2G	24.6M	45.5

Table 4: Components effectiveness of the proposed method.

Method	Val-Top1 (%)	Val-Top5 (%)
baseline(TSN)	19.7	46.6
GRI	38.2	67.2
CTI	42.8	71.3
MFI	43.9	73.9

Fig. 4: Some prediction examples on Something-Something V1. The top 2 predictions with green text indicating a correct prediction.