MFI: Multi-range Feature Interchange for Video Action Recognition

Sikai Bai (presenter), Qi Wang and Xuelong Li

School of Computer Science and Center for OPTical IMagery Analysis and Learning
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

Self-introduction

Sikai Bai, received the B.E. degree in software engineering from the China University Of Geosciences, Wuhan, 730074, Hubei, P. R. China, in 2018. He is currently pursuing the Master degree from school of Computer Science and Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

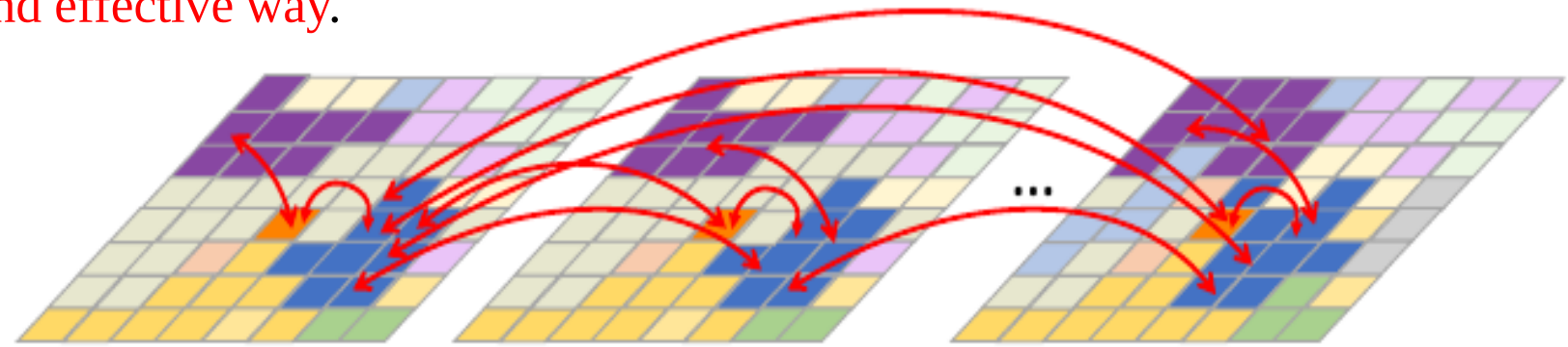
Pipeline

- **Motivations**
- **Our approach**
- **Experiments**

Motivation

- Short-range motion features and long-range dependencies are two complementary and vital cues.

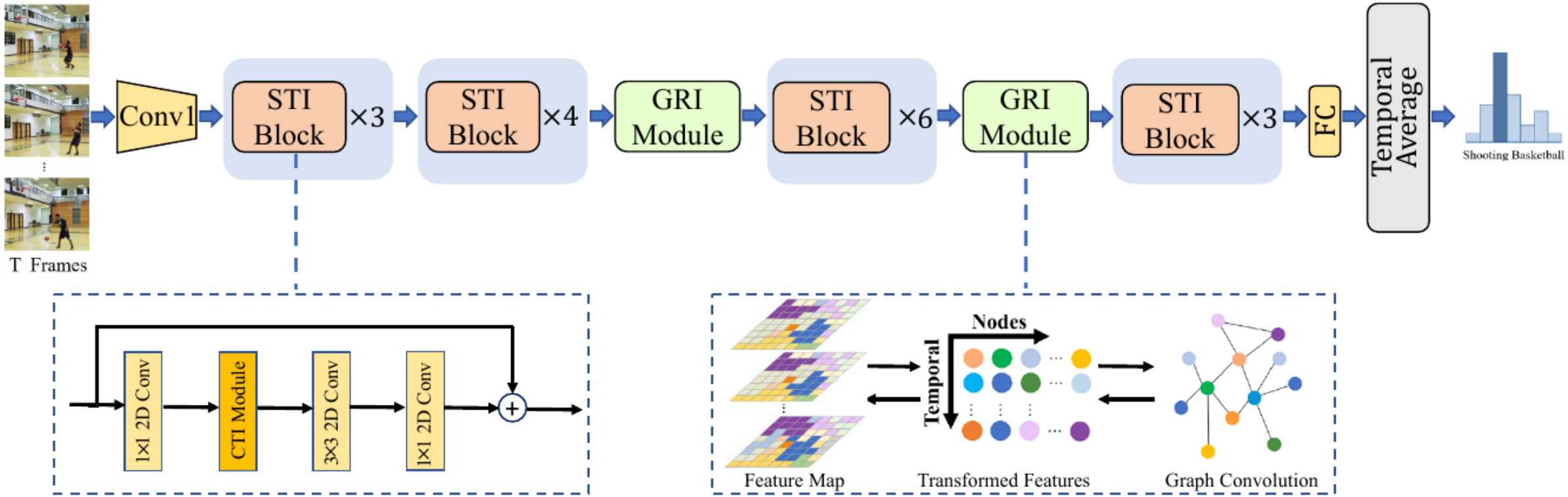
- It is still unclear how to capture temporal information on multiple ranges using an efficient and effective way.



- Regard short-range motion encoding and long-range dependency learning as the interchange between features in multiple ranges.

Our approach

Multi-range Feature Interchange for Video Action Recognition



Our approach

- Channel-wise Temporal Interchange (CTI)
  - Temporal difference.

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

  - Temporal interchange operation.
- Graph-based Regional Interchange (GRI)
  - Feature Transformation.

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

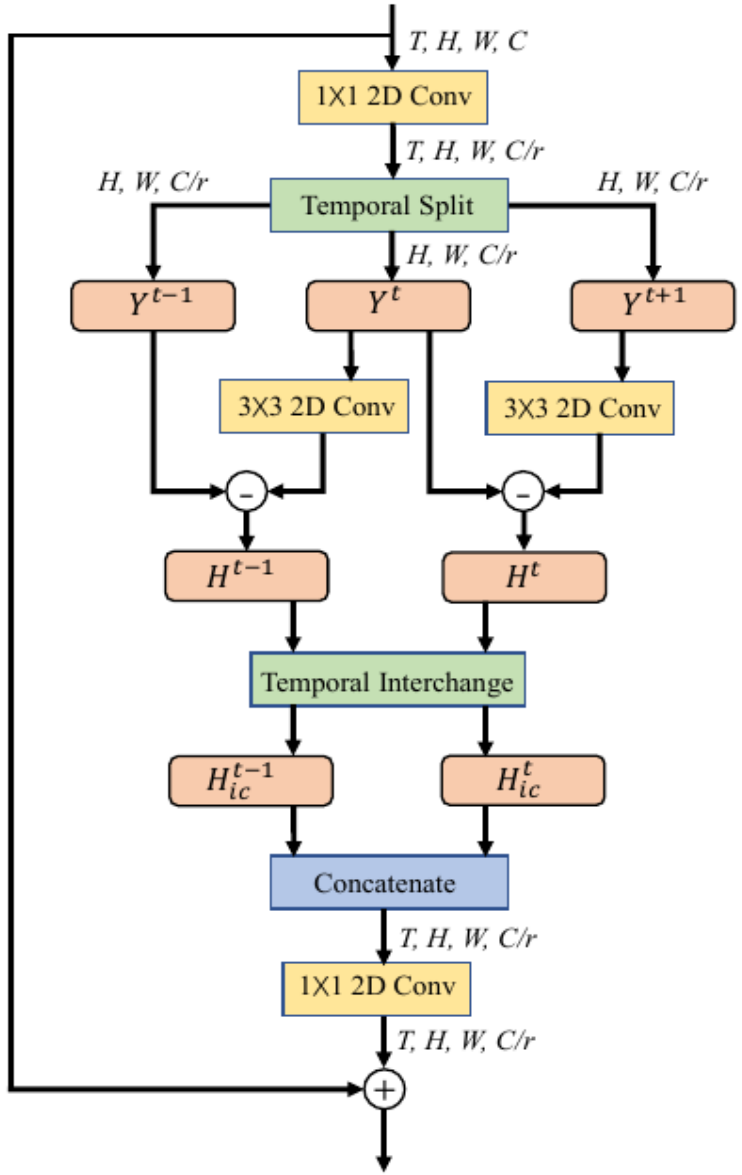
$$V_t = W_t * \Phi_r(X), \qquad V_t \in R^{N \times C}.$$

  - Graph Convolution.

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

  - Feature Reverse.

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



Experiments

- Benchmark Comparison

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	$153G \times 2$	41.6	72.2
Non-Local-I3D	3D ResNet50	32×2	$168G \times 2$	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

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

Experiments

- Ablation Study

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

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

Thank you for listening!

Sikai Bai
Email: whitesk1973@gmail.com
2020-12-07