Paper ID 2190

Blockwise Temporal-Spatial pathway network

SeulGi Hong and MinKook Choi
Hutom, Seoul, South Korea

Overview

We propose a 3D-CNN-based action recognition model, called the blockwise temporal-spatial path-way network (**BTSNet**), which can **adjust the temporal and spatial receptive fields by multiple pathways**. We designed a novel model inspired by an adaptive kernel selection-based model, which chooses spatial receptive fields for image recognition. Expanding this approach [1] to the temporal domain, **our model extracts temporal and channel-wise attention and fuses information on various candidate operations.** We confirm that proposed TSP block supports a better representation for 3D convolutional blocks based on our visualization.

Receptive Fields

Both the SlowFast [2] and TSP blocks have a widened view along the temporal axis. However, the TSP block has more generalized RFs and can control the contributions of the RFs at each timestep. Two hyper-parameters to handle the RF in our pathway block: **the number of pathways M** and the **RF option** (O1 and O2).

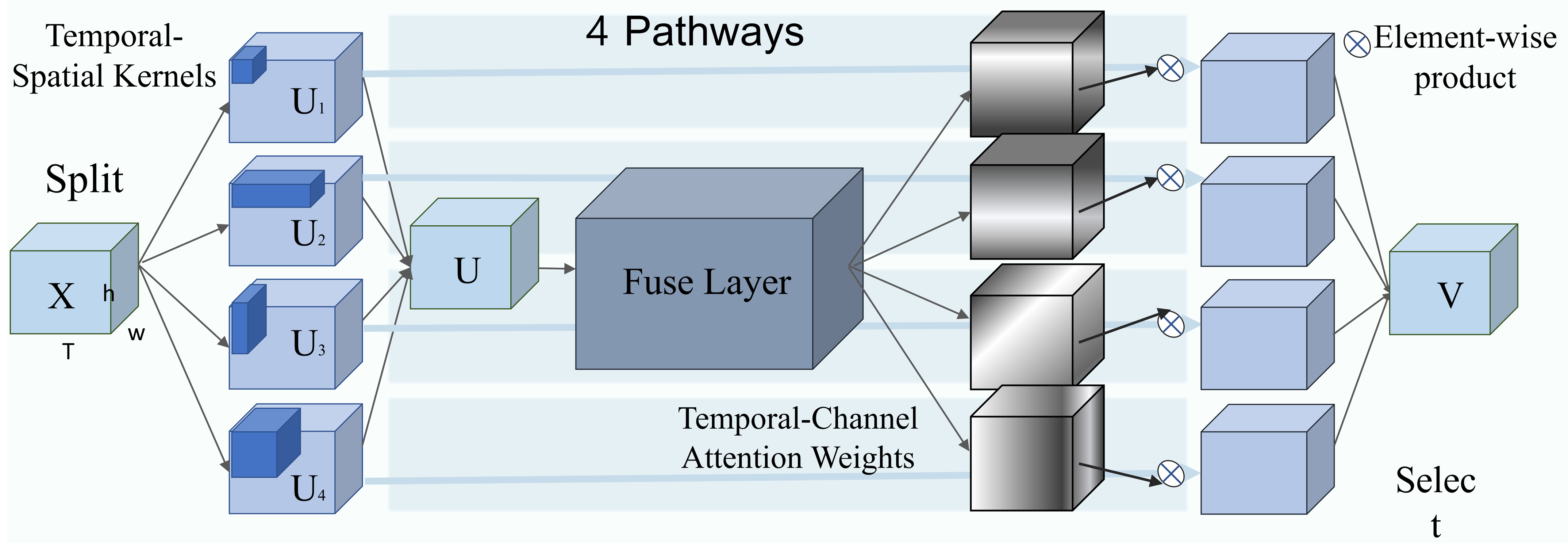
- **O1**: cube-like RFs. For dilation parameters (T, H, W), $D = \{D_1, D_2, \cdots, D_M\}, D_i = (i, i, i)$
- **O2**: customized dilation parameters eg. for M=4, $\{(1,1,1),(4,4,4),(1,4,4),(4,1,1)\}$

Temporal-Spatial Pathway Block

Split. For any given feature $X \in \mathbb{R}^{C' \times T' \times H' \times W'}$, transformation functions $F_{1,2,..,m}$ are applied first.
Our pathway blocks can be considered as slow, fast, or spatially enlarged pathways. $F_m : X \rightarrow U_m \in \mathbb{R}^{C \times T \times H \times W}$

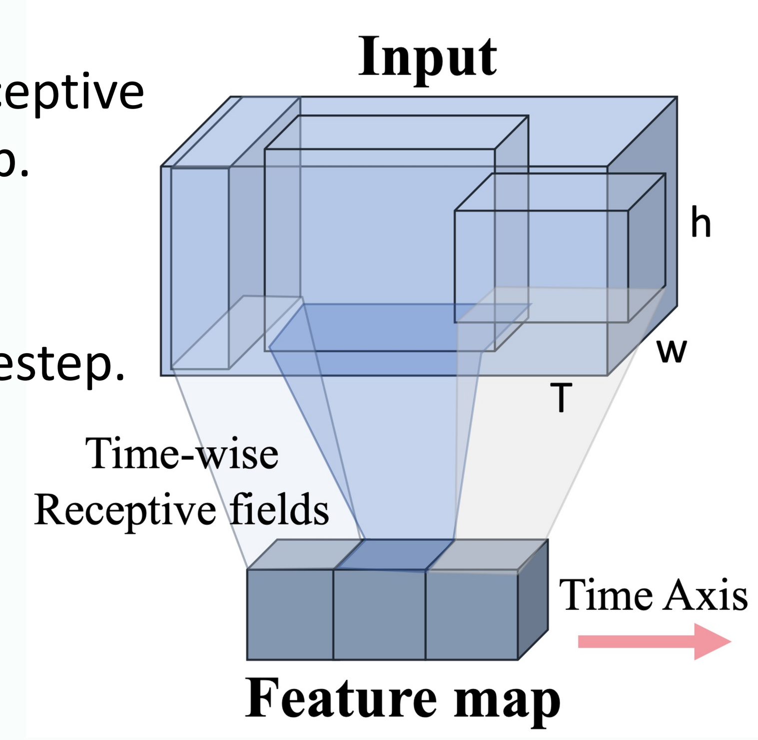
Fuse Layer. To fuse information from multiple receptive fields, we combine the previous features by adding all U_m.
Then, global average pooling (GAP) is applied to compress features U.
There are two options for this layer: **temporal-channel attention (TC)** and **channel-wise attention (C)**.
A compact feature $Z \in \mathbb{R}^{d \times T}$ or $Z \in \mathbb{R}^{d}$ can be attained by a set of operations. The set is composed of a convolution with a (1,1,1) kernel, batch normalization, and ReLU.
Z should be resized to $Z' \in \mathbb{R}^{M \times C}$ or $Z' \in \mathbb{R}^{M \times C \times T}$ to attain the attention vectors, therefore convolution with kernel size 1 is applied.
Select. To highlight the information among multiple pathways, we use temporal-channel or channel-wise attention mechanisms in this procedure. $Attn = softmax(Z')$ These attention weights emphasize each pathway along the temporal-channel axis, which has a different RFs. The final output V of the block is: $V = \sum_{m=1}^{M} Attn_m * U_m.$

Left: Procedure of TSP block of BTSNet.



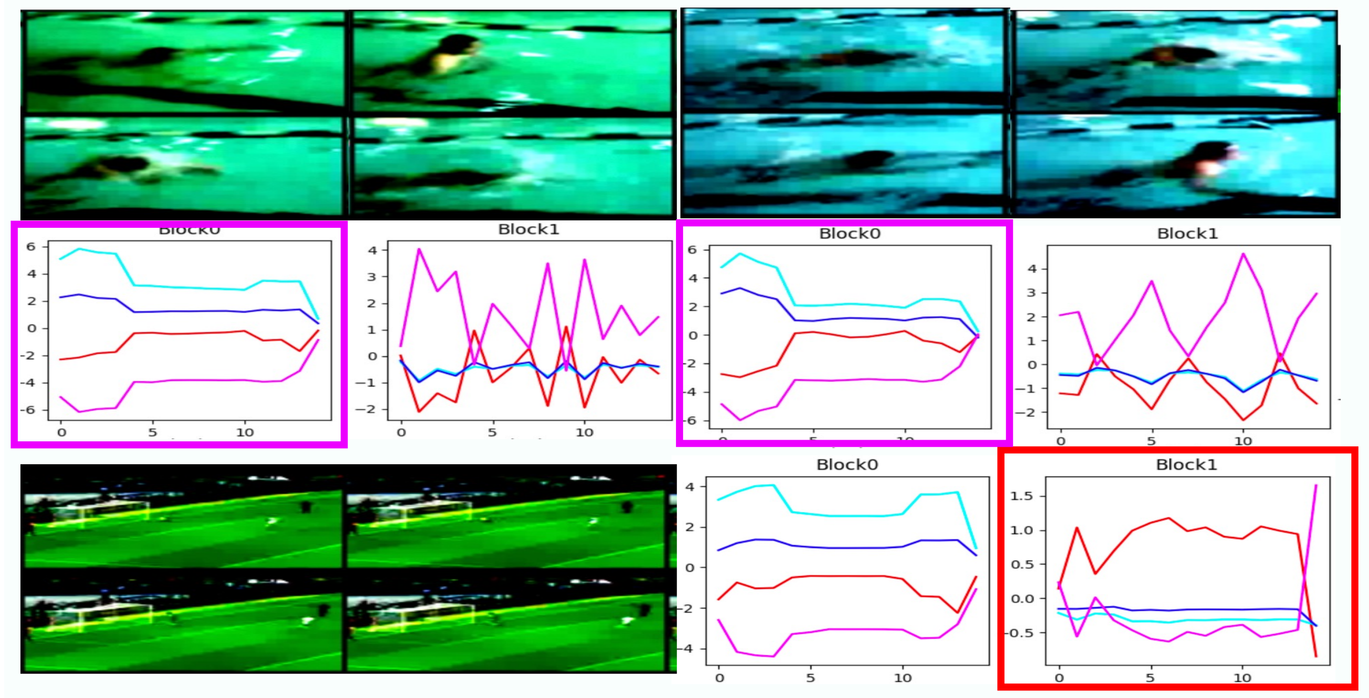
Right. Corresponding receptive fields for the feature map.

Pathways have top-1 contribution at each timestep.



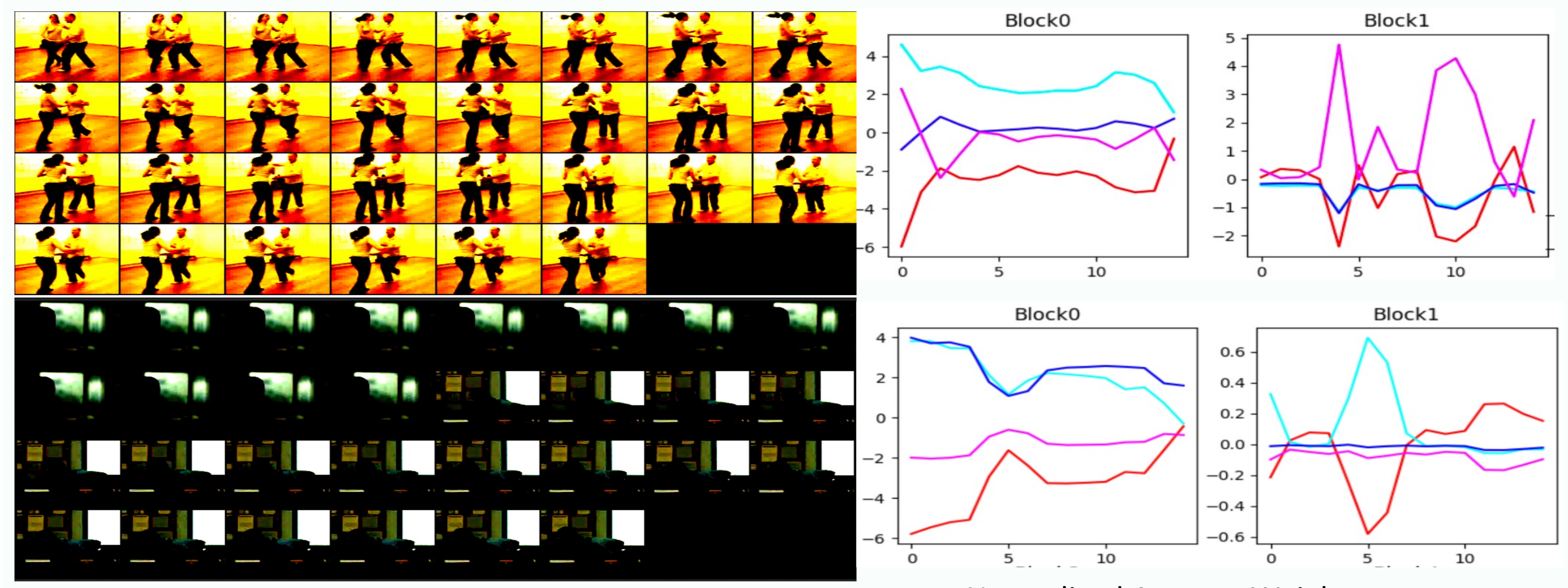
Results

To summarize, **our model attain informative areas** when there are enough pathways with temporal-channel attention. In visualization, receptive fields [T,H,W] are notated as red [1,1,1], cyan[4,4,4], blue [1,4,4], magenta[4,1,1].



(a)
- Same Action
- Dynamic
- Large Variance

(b)
- Different Action
- Stable
- Small variance



[Ablation Study]

Row 1. fuse layer.

Ablation	M3-O2-26	M4-O2-26	M3-O1-26	M3-O1-50	M3-O1-101
C	57.507	59.913	58.974	58.459	58.565
TC	60.283	60.058	61.829	59.120	60.005
TC-C	**2.776**	**0.145**	**2.855**	**0.661**	**1.440**

Row 2. the number of pathways M.

Ablation	TC-O2-50
M=2	55.855
M=3	58.895
M=4	58.274
M=7	58.551
Max Gap	3.04

Ablation	TC-O1-26	TC-O1-50	TC-O1-101
M=2	60.019	58.842	56.133
M=3	**61.829**	**59.120**	**60.005**
Ablation	C-O2-26	C-O2-50	C-O1-26
M=3	57.507	57.454	58.974
M=4	**59.913**	**58.261**	**59.252**

Row 3. RF option

Ablation	C-M4-26	C-M4-50	TC-M3-50	TC-M4-50
O1	59.252	48.678	59.120	58.961
O2	59.913	58.261	58.895	58.274
O2-O1	**0.661**	**9.583**	-0.225	-0.687

References

[1] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang,"Selective kernel networks," in2019 IEEE/CVF Con-ference on Computer Vision and Pattern Recognition(CVPR), 2019, pp. 510–51
[2] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slow-fast networks for video recognition," in2019 IEEE/CVFInternational Conference on Computer Vision (ICCV),2019, pp. 6201–621

Acknowledgements. This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202012A02-02)