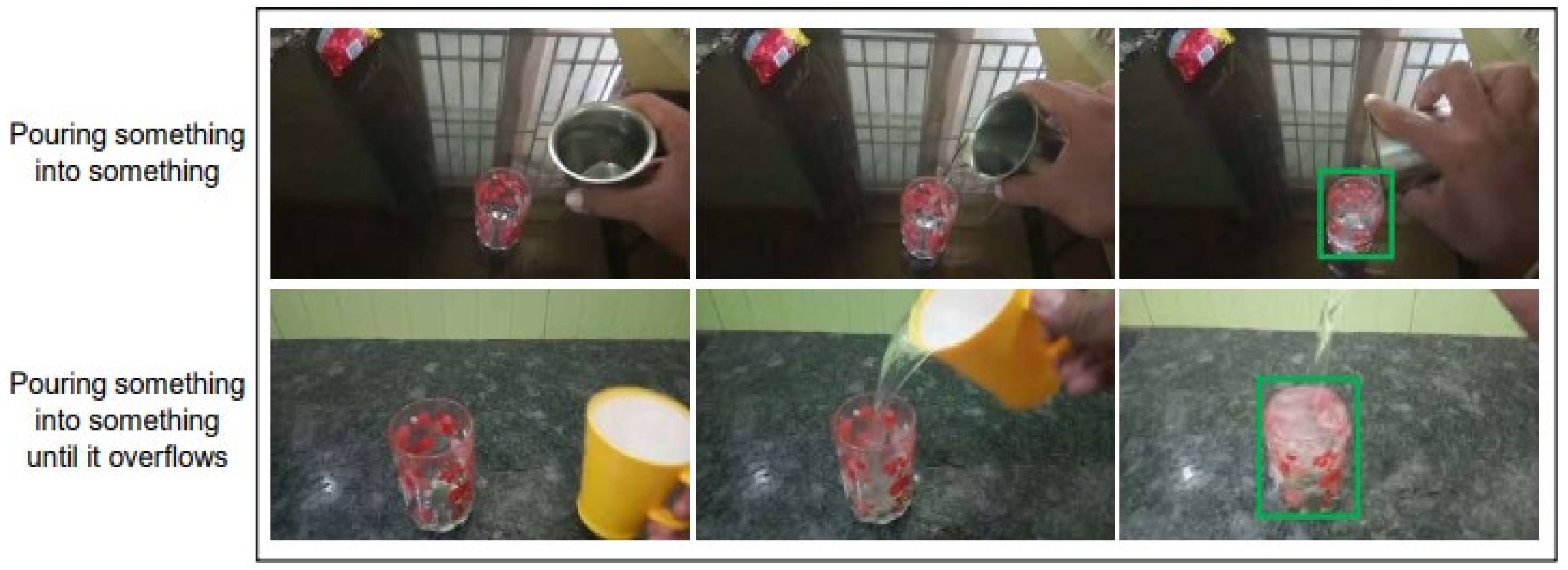
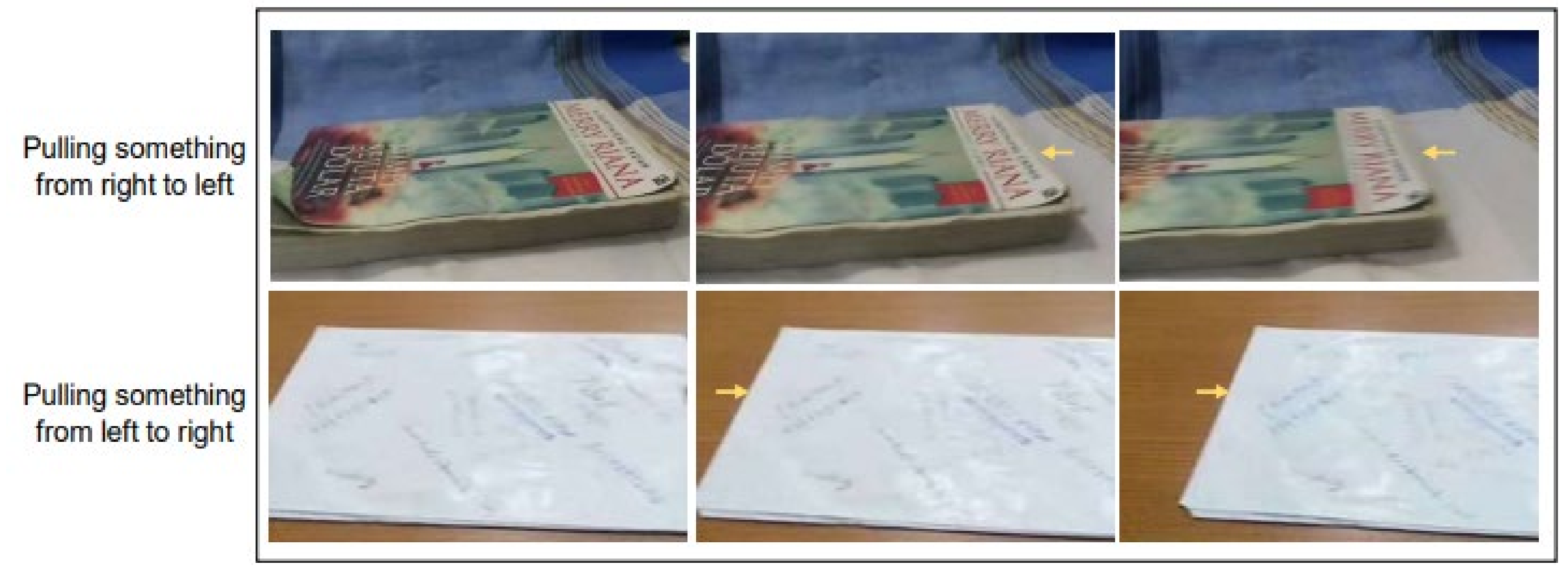
Dynamic Spatio-Temporal Specialization Learning for Fine-Grained Action Recognition

Tianjiao Li[1*], Lin Geng Foo[1*], Qiuhong Ke[2], Hossein Rahmani[3], Anran Wang[4], Jinghua Wang[5], and Jun Liu[1]
[1] Singapore University of Technology and Design, [2] University of Monash, [3] Lancaster University, [4] ByteDance Inc., [5] Harbin Institute of Technology

Introduction

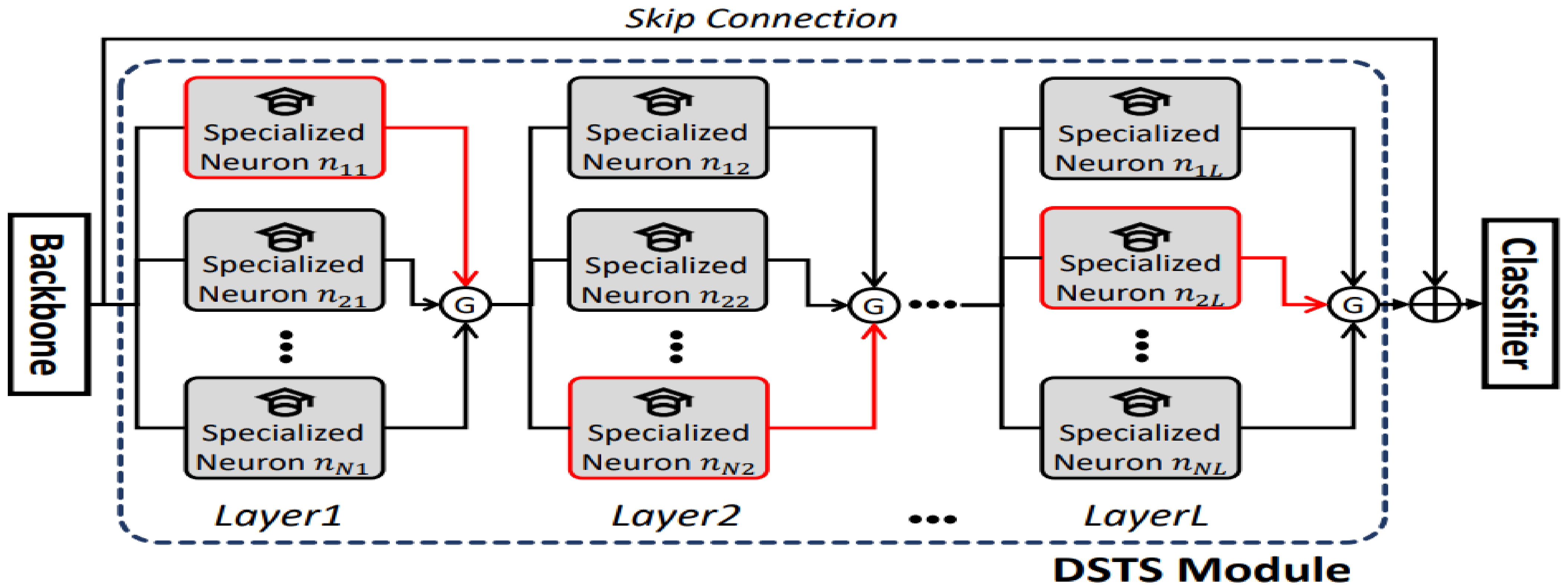

Pouring something into something
Pouring something into something until it overflows


Pulling something from right to left
Pulling something from left to right

Goal: To successfully discriminate between action categories with subtle differences.
Motivation: As shown in the figure above, the fine-grained differences lie either in spatial aspects (top) or temporal aspects (temporal). Thus, this motivates us to capture either more spatial or temporal fine-grained information, to better tackle the large range of spatio-temporal variations in the videos.

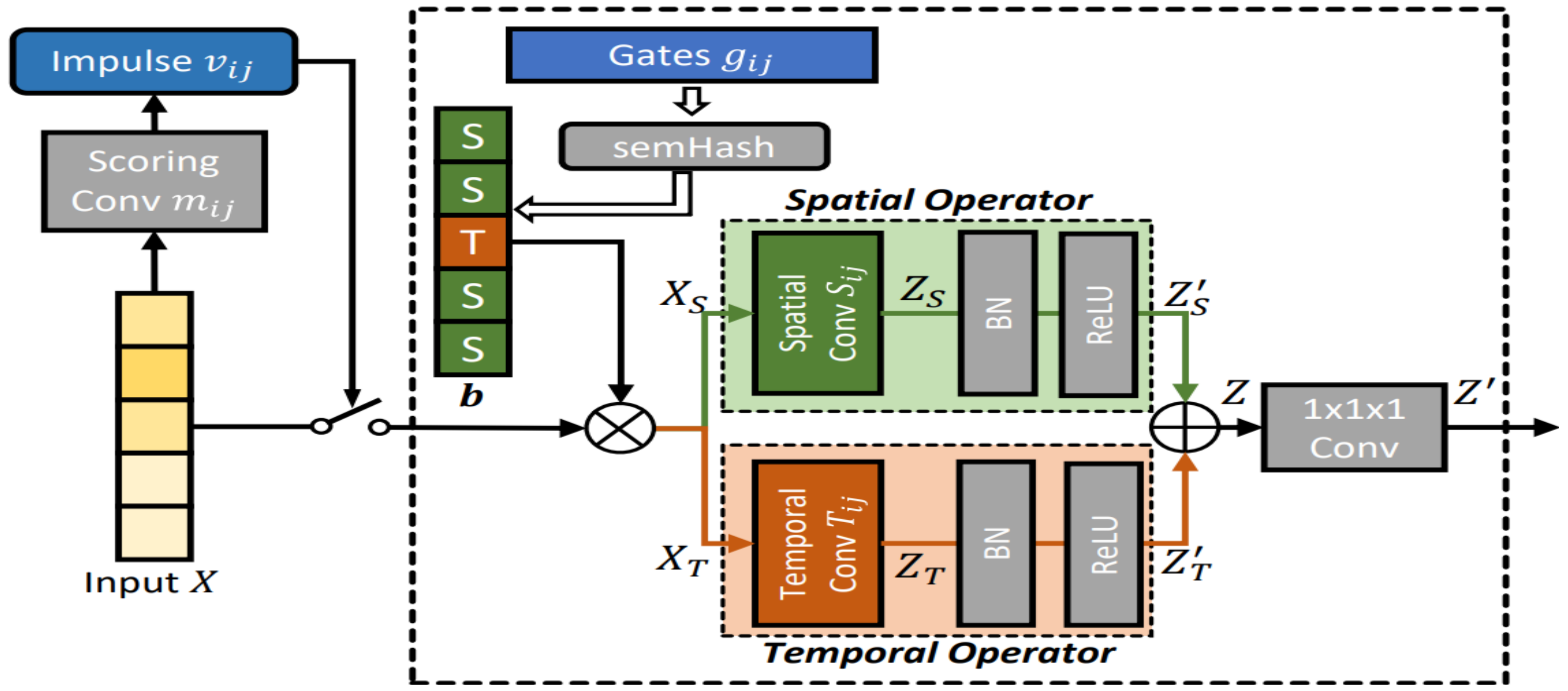
DSTS Module

Inspired by the specialization of neurons in the human brain, we improve fine-grained recognition capabilities of a neural network by using specialized parameters that are only activated on a subset of the data. More specifically, we design a Dynamic Spatio-Temporal Specialization (DSTS) module which consists of specialized neurons that are only activated when the input is within their area of specialization (either in spatial aspects or temporal aspects).


Skip Connection
Backbone
Specialized Neuron n_{11}
Specialized Neuron n_{21}
Specialized Neuron n_{N1}
G
Specialized Neuron n_{12}
Specialized Neuron n_{22}
Specialized Neuron n_{N2}
G
Specialized Neuron n_{1L}
Specialized Neuron n_{2L}
Specialized Neuron n_{NL}
G
Classifier
Layer1
Layer2
LayerL
DSTS Module

Specialized Neurons in DSTS Module

To achieve spatial or temporal specialization, we explicitly restrict the specialized neurons to choose between spatial or temporal operators for each input channel. During training, this design forces each specialized neuron to exploit fine-grained differences in each channel between similar samples in the chosen aspect, leading to better sensitivity towards these fine-grained differences.



To further improve the performance, a UDL algorithm is introduced to optimize the model parameters that make dynamic decisions (scoring kernels ***m*** and gate parameters ***g***), which we call upstream parameters. These upstream parameters that make dynamic decisions and downstream parameters (spatial and temporal operators ***S*** and ***T***) that process input, are jointly trained during our end-to-end training, which can be challenging as upstream parameters themselves also affect the training of downstream ones. Thus, we use meta-learning to optimize upstream parameters while taking their downstream effects into account:

Simulated Update Step: $\hat{d} = d - \alpha\nabla_d \cdot L(u, d; D_{train})$

Meta-Update Step: $u' = u - \alpha\nabla_u \cdot L(\hat{u}, \hat{d}; D_{val})$

Actual Update Step: $d' = d - \alpha\nabla_d \cdot L(d, u'; D_{train})$

Experimental Results

1. Results on Something-Something v2

Methods	Top-1 Acc	Top-5 Acc
SlowFast	63.1	87.6
TPN	64.7	88.1
ViViT-L	65.4	89.8
TSM	66.6	91.3
MViT-B	67.7	90.9
Swin-B	69.6	92.7
DSTS + TPN	**67.2**	**89.2**
DSTS + Swin-B	**71.8**	**93.7**

2. Results on Diving-48

Methods	Top-1 Acc	Class-wise Acc
I3D	48.3	33.2
TSM	52.5	32.7
GST	78.9	69.5
TQN	81.8	74.5
Swin-B	80.5	69.7
TPN	86.2	76.0
DSTS + TPN	**88.4**	**78.2**
DSTS + Swin-B	**83.0**	**71.5**

Conclusion

In this paper, we proposed a novel DSTS module consisting of dynamically activated specialized neurons for fine-grained action recognition. Our spatio-temporal specialization method optimizes the architectures of specialized neurons to focus more on spatial or temporal aspects. We obtain state-of-the-art fine-grained action recognition performance on two popular datasets by adding DSTS modules to baseline architectures.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022
October 23-27, 2022, Tel Aviv