Stacked Temporal Attention: Improving First-person Action Recognition by Emphasizing Discriminative Clips Supplementary Material

Lijin Yang
yang-lj@iis.u-tokyo.ac.jp
Yifei Huang
hyf@iis.u-tokyo.ac.jp
Yusuke Sugano
sugano@iis.u-tokyo.ac.jp
Yoichi Sato
ysato@iis.u-tokyo.ac.jp

Institute of Industrial Science
The University of Tokyo
Tokyo, Japan

1 Implementation Details

We use PyTorch [8] for all the implementation. All backbones are pretrained on Kinetics dataset [1]. We train STAM together with pretrained backbone in the end-to-end manner using the Adam optimizer [5] with initial learning rate 1e-4 for 40 epochs. We decay the learning rate by a factor of 10 when the loss does not decrease for 3 consecutive epochs. As for the loss weight λ we empirically set all the λs to 1.

The hidden dim d is set as 512 in all experiments. For the number of global attention layer M, we empirically set different values to cooperate with different backbone encoders. We set $M = 2$ when using TSM [7] and R3D-50 [4] as backbone encoder. For the backbone encoder I3D [2], we utilize 3 global attention layers for better performance. Following the practice of [10], we use 9 clips with 16 frames as input for I3D backbones while 6 clips with 16 frames for R3D-50 backbones. For the 2D backbone TSM, we uniformly sample 16 frames as the input. The output temporal dimension is the same as clip number for both 2D and 3D backbones.

2 More results on initialization options of global feature

Similar to experiments in Section 4.1, here we show the comparison result of global feature initialization alternatives with different input clip number in Table 5. Note that in Table 1 of main manuscript the input clips are 9, while in Table 5 of this supplementary material the number of input clips are 6. Combining the results of Table 1 and 5, we choose self-attention

Global feature options	**Num of global att layer**		
	1	2	3
Max pooling	64.8	65.4	64.4
Avg pooling	63.8	65.3	**65.6**
Bi-GRU	64.4	64.8	64.9
1D-conv	64.1	64.4	64.7
Self-att	**65.1**	**65.6**	65.1
Light-weight CNN	64.6	65.3	64.9

Table 5: Different design options for the global feature. Experiments are conducted on the EGTEA dataset split1 using I3D as backbone and 6 clips as input.

as the global feature initialization option unless otherwise stated since this option provides better and more stable results.

3 Experiments on HMDB51 dataset

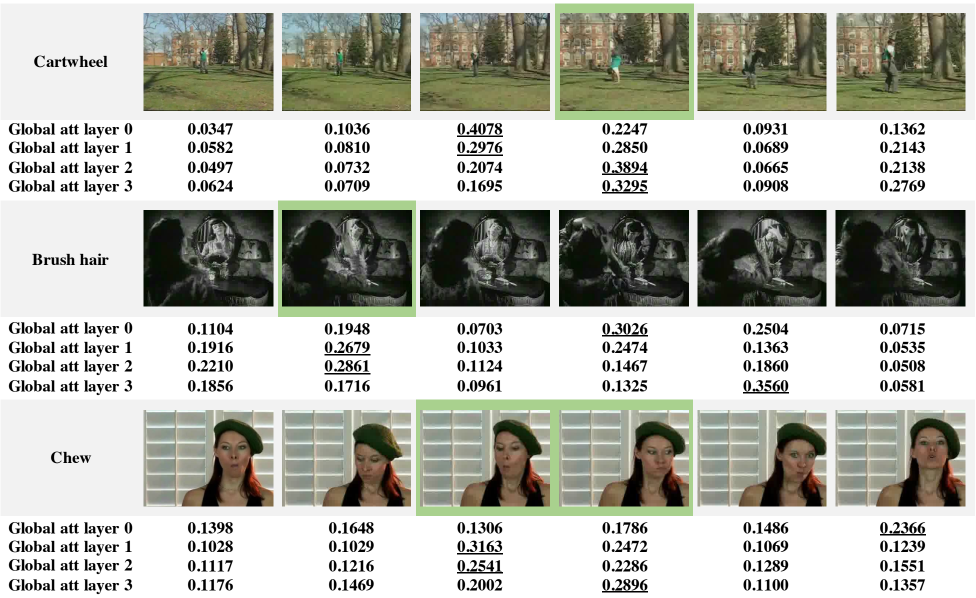
Table 6 shows the result comparison on the HMDB51 dataset [6]. Although the performance of our STAM on TSM backbone is not as obvious as that in first-person datasets, clear improvement can still be validated when using I3D and R3D-50 as backbone encoders, which proves that our STAM is generalizable to third-person dataset. Moreover, I3D with STAM on the top achieves better accuracy than CatNet [9] which also uses I3D as the backbone encoder. This performance validates that refining temporal attention by stacking global attention layers can generate more reasonable temporal attention.

Method	**Acc**
TSN* [10]	69.4
TLE* [3]	71.1
CatNet [9]	75.2
TSM [7]	69.4
I3D [2]	73.1
R3D-50 [4]	67.8
TSM + STAM	70.3
I3D + STAM	**76.1**
R3D-50 + STAM	70.5

Table 6: Results on the HMDB51 dataset. * indicate the method uses optical flow as input.

4 Visualization examples from HMDB51 dataset

Figure 4 shows three examples from HMDB51 dataset together with the temporal attention value of each clip when using 6 clips as input. We use one frame to represent each 16-frame clip in this Figure. As can be seen from the examples, the scores produced by the initialization layer tend to highlight some repeated clips, thus resulting wrong prediction. However, after adding global attention layers, the temporal attention gradually shifts from wrong clips to discriminative clips. In this figure, the temporal attention score becomes optimal when stacking 2 global attention layers, and saturates when stacking more layers.



Cartwheel						
Global att layer 0	**0.0347**	**0.1036**	**0.4078**	**0.2247**	**0.0931**	**0.1362**
Global att layer 1	**0.0582**	**0.0810**	**0.2976**	**0.2850**	**0.0689**	**0.2143**
Global att layer 2	**0.0497**	**0.0732**	**0.2074**	**0.3894**	**0.0665**	**0.2138**
Global att layer 3	**0.0624**	**0.0709**	**0.1695**	**0.3295**	**0.0908**	**0.2769**

Brush hair						
Global att layer 0	**0.1104**	**0.1948**	**0.0703**	**0.3026**	**0.2504**	**0.0715**
Global att layer 1	**0.1916**	**0.2679**	**0.1033**	**0.2474**	**0.1363**	**0.0535**
Global att layer 2	**0.2210**	**0.2861**	**0.1124**	**0.1467**	**0.1860**	**0.0508**
Global att layer 3	**0.1856**	**0.1716**	**0.0961**	**0.1325**	**0.3560**	**0.0581**

Chew						
Global att layer 0	**0.1398**	**0.1648**	**0.1306**	**0.1786**	**0.1486**	**0.2366**
Global att layer 1	**0.1028**	**0.1029**	**0.3163**	**0.2472**	**0.1069**	**0.1239**
Global att layer 2	**0.1117**	**0.1216**	**0.2541**	**0.2286**	**0.1289**	**0.1551**
Global att layer 3	**0.1176**	**0.1469**	**0.2002**	**0.2896**	**0.1100**	**0.1357**

Figure 4: Visualization of temporal attention scores of three samples from the HMDB51 dataset when stacking different numbers of global attention layers. The highest temporal attention score in each layer is underlined. In the case of global attention layer 0, initialization using the self-attention method is used. The frame with green background indicates that the backbone model can predict the correct action class with this clip alone as input.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2329–2338, 2017.

[4] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[7] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

[9] Jiaze Wang, Xiaojiang Peng, and Yu Qiao. Cascade multi-head attention networks for action recognition. *Computer Vision and Image Understanding*, 192:102898, 2020.

[10] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.