Spatiotemporal Self-attention Modeling with Temporal Patch Shift for Action Recognition - Supplementary Material -

Wangmeng Xiang[1,2⋆], Chao Li[2], Biao Wang[2], Xihan Wei[2], Xian-Sheng Hua[2],
and Lei Zhang[1⋆⋆]

[1] The Hong Kong Polytechnic University, Hong Kong SAR, China
{cswxiang,cslzhang}@comp.polyu.edu.hk
[2] DAMO Academy, Alibaba, Hangzhou, China
{lllcho.lc,wb.wangbiao,xihan.wxh,xiansheng.hxs}@alibaba-inc.com

1 More details on the implementation of TPS

We present the "PyTorch style" pseudo-code of a special case of TPS, where the patches from neighboring frames are shifted using a "Bayer filter" pattern. The patch shift module is placed before the self-attention calculation, and then a shift back operation is employed after the self-attention calculation.

```
# Patch shift with pattern B
def PatchShift(x, shift_back):
   B, C, T, H, W = x.size()
   direct = 1
   if shift_back:
       direct = -1
   x[:,:,:,0::2,1::2] = roll(x[:,:,:,0::2,1::2], shifts= direct, dims=2)
   x[:,:,:,1::2,0::2] = roll(x[:,:,:,1::2,0::2], shifts=-direct, dims=2)
   return x
```

The simple patch shift function described above transfers a vanilla spatial self-attention to a spatiotemporal self-attention. Other patterns can be implemented in a similar paradigm. As we use Swin Transformer [5] as our backbone, the coordinates of patches are shifted alongside the patches.

2 More details on the shift patterns

We show the Patterns A, B, C and D in Figure 1 here. Pattern B is obtained via replacing 1/4 index-0 frame patches in pattern A by index-2 frame patches. Pattern C is obtained by placing patches from 9 consecutive frames in a 3 × 3 grid. Pattern D is obtained by placing patches from 16 consecutive frames in a 4 × 4 grid.

[⋆] Work done during an internship at Alibaba.

[⋆⋆] Corresponding author.

0	1	0	1
-1	0	-1	0
0	1	0	1
-1	0	-1	0

(a) Pattern A

0	1	0	1
-1	2	-1	2
0	1	0	1
-1	2	-1	2

(b) Pattern B

-4	1	2	-4
-1	0	3	-1
-2	-3	4	-2
-4	1	2	-4

(c) Pattern C

0	1	2	3
-1	-7	4	5
-2	-4	7	6
-3	-5	-6	8

(d) Pattern D

Fig. 1. Pattern A, B, C and D in Table 2(b) of the main paper.

3 Additional results on DeiT backbone

TPS is a plug-and-play module and it can be embedded into many existing transformers to increase their spatiotemporal modeling capabilities. To illustrate the effectiveness of TPS, we further test our method on another popular backbone DeiT [7]. We use DeiT-S for our experiment, which has 14 × 14 image patches at every layer and an additional class token. We insert a TPS module in every two blocks of DeiT and operate directly on image patches. We densely sample 16 frames as input for training. Other training and testing configurations are the same as PST-T.

We compare our proposed DeiT-S-TPS with DeiT-S-2D and DeiT-S-TSM. DeiT-S-2D uses 2D DeiT-S for frame feature extraction and average pooling for temporal modeling. DeiT-S-TSM applies temporal channel shift [4] on both image patches and class token, and the channel shift ratio is set the same as [4]. The results are show in Table 1. DeiT-S-TPS improves 2.3% top-1 accuracy on Kinetics400 over DeiT-S-2D. DeiT-S-TPS also outperforms DeiT-S-TSM by 0.5% thanks to the efficient spatio-temporal self-attention modeling.

Table 1. More backbones experiments on Kinetics400.

Model	Pretrain	Crops × Clips	FLOPs	Params	Top-1	Top-5
DeiT-S-2D [7]	IN-1K	3 × 4	74G	22M	73.0	90.7
DeiT-S-TSM	IN-1K	3 × 4	74G	22M	74.8	91.6
DeiT-S-TPS	IN-1K	3 × 4	74G	22M	**75.3**	**91.8**

4 Visualization

We use GradCAM [6] for visualization of the last stage feature maps. In Fig. 3, we present three examples of consecutive frames in Something-something-V2 videos in the first three columns, the learned feature activation maps by applying average pooling, channel shift and PST are presented in fourth, fifth and sixth columns, respectively. Our results show that PST learns spatiotemporal relation by associating relevant regions, which clearly indicates the motion of actions. Channel shift can also learn motion at some extent, but the activation map

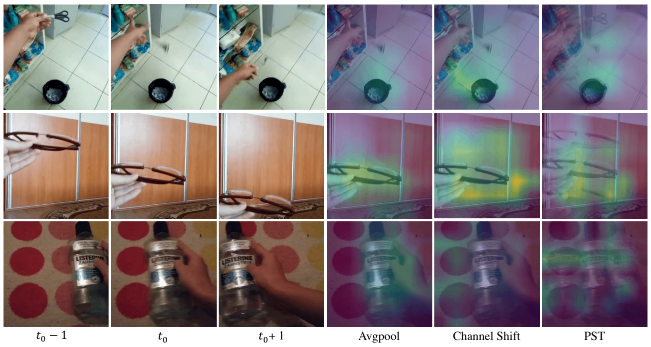

Fig. 2. Visualization by GradCAM on Something-something V2. Our PST learns to focus on the moving of objects.

indicates that the receptive field is not as broad as PST. While in feature maps of average pooling, the motion can not be easily learned.

4.1 More visualization results

We also present three visualization examples of PST from Kinetics400 [2], Diving48 [3] and Something-something V1 [1], respectively, in Fig. 3. The consecutive frames of each video are shown in the first three columns, and the learned feature maps by PST are presented in the last column. It can be seen that PST can learn to associate relevant regions, and the feature maps can indicate the motion of actions in the video.

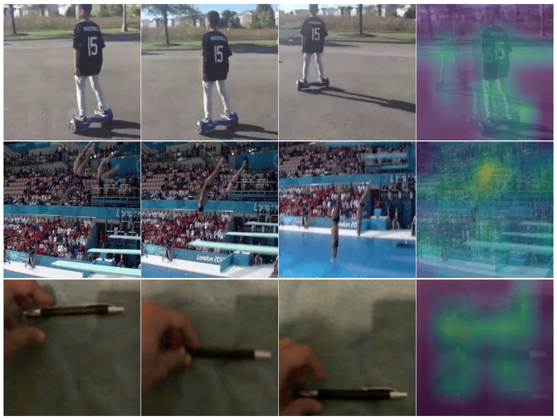

Fig. 3. Visualization by GradCAM on Kinetics400 (first row), Diving48 (second row) and Something-something V1 (third row. PST can learn to focus on the motion of objects.

References

1. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Int. Conf. Comput. Vis. pp. 5842–5850 (2017)
2. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
3. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–528 (2018)
4. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Int. Conf. Comput. Vis. pp. 7083–7093 (2019)
5. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
6. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
7. Touvron, H., Cord, M., Matthijs, D., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: ICML 2021: 38th International Conference on Machine Learning (2021)