MSG-Transformer: Exchanging Local Spatial Information by Manipulating Messenger Tokens
- Supplementary Materials -

Jiemin Fang[1,2], Lingxi Xie[3], Xinggang Wang[2†], Xiaopeng Zhang[3], Wenyu Liu[2], Qi Tian[3]
[1]Institute of Artificial Intelligence, Huazhong University of Science & Technology
[2]School of EIC, Huazhong University of Science & Technology [3]Huawei Inc.
{jaminfong, xgwang, liuwy}@hust.edu.cn
{198808xc, zxphistory}@gmail.com tian.qi1@huawei.com

Table 1. Image classification accuracy on ImageNet-1K comparing with concurrent hierarchical Transformers.

w/ dw-conv?	Method	Params	FLOPs	Top-1 (%)
✗	Swin-T [4]	28M	4.5G	81.3
	MSG-T	25M	3.8G	**82.4**
✓	GG-T [5]	28M	4.5G	82.0
	Shuffle-T [3]	29M	4.6G	82.5
	CSWin-T [2]	23M	4.3G	82.7
	MSG-T$_{\text{dwc}}$	25M	3.9G	**83.0**

A. Appendix

A.1. Comparisons with Concurrent Hierarchical Transformers

Some related concurrent works [2,3,5] also focus on improving attention computing patterns with different manners based on a hierarchical architecture and achieve remarkable performance. These works introduce additional depth-wise convolutions [1] into Transformer blocks, which improve recognition accuracy with low FLOPs increase. Our MSG-Transformers in the main text do not include depth-wise convolutions to make the designed model a purer Transformer. We further equip MSG-T with depth-wise convolutions, resulting in a variant named MSG-T$_{\text{dwc}}$. As in Tab. 1, MSG-T$_{\text{dwc}}$ shows promising performance with low FLOPs. We believe these newly proposed attention computing patterns will facilitate future vision Transformer research in various manners and scenarios.

A.2. Analysis about Advantages of MSG Tokens

We take Swin- [4] and MSG-Transformer for comparison, and analyze their behaviors from two aspects as follows.

Table 2. Ablation study about MSG token manipulation.

Manip. Op.	Shuffle	Average	Shift
ImageNet Top-1 (%)	**81.1**	80.8	80.6

Receptive fields. Let the window size be W. For Swin, the window is shifted by $\frac{W}{2}$ in every two Transformer blocks and the receptive field is $\left(\frac{3W}{2}\right)^2$ after two attention computations. For MSG, assuming the shuffle size is $S \geq 2$, a larger receptive field of $(SW)^2$ is obtained with two attention computations.

Information exchange. In Swin, each patch token obtains information from other patch tokens in different windows, where valuable information is extracted by interacting with many other patch tokens. In MSG, information from one window is summarized by a MSG token and directly delivered to patch tokens in other windows. This manner eases the difficulty of patch tokens obtaining information from other locations and promotes the efficiency.

A.3. Study about Manipulations on MSG Tokens

As claimed in the main text, how to manipulate MSG tokens is not limited to the adopted shuffle operation. We study two additional manipulations, namely, the 'average' (MSG tokens are averaged for the next-round attention) and 'shift' (MSG tokens are spatially shifted). As in Tab. 2, 'shuffle' works the best, and we conjecture that 'average' lacks discrimination for different windows, and 'shift' requires more stages to deliver information to all the other windows. We believe explorations on manipulation types carry great potential and will continue this as an important future work.

References

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 1

[2] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv:2107.00652*, 2021. 1

[3] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv:2106.03650*, 2021. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[5] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan Yuille, and Wei Shen. Glance-and-gaze vision transformer. In *NeurIPS*, 2021. 1