Trade-off between Robustness and Accuracy of Vision Transformers (Supplementary Material)

Yanxi Li, Chang Xu
School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
yali0722@uni.sydney.edu.au, c.xu@sydney.edu.au

A. Different Backbones Sizes

Models	Clean	FGSM	PGD	A	R	C(↓)
RVT-Ti	79.2	42.7	18.9	14.4	43.9	57.0
Ours Ti/16 (0.1)	**82.1**	38.5	13.8	**33.2**	**51.4**	**42.1**
Ours Ti/16 (0.5)	80.9	44.4	29.5	26.8	49.0	43.5
Ours Ti/16 (0.9)	77.5	**63.3**	**47.9**	10.2	48.1	51.4
RVT-S	81.9	51.8	28.2	25.7	47.7	49.4
Ours S/16 (0.1)	**83.6**	48.1	22.5	**45.1**	**54.0**	**35.1**
Ours S/16 (0.5)	82.1	54.2	37.8	37.4	52.7	38.8
Ours S/16 (0.9)	78.6	**73.3**	**56.6**	20.5	50.9	43.3

Table A. The results of using backbones of various sizes.

In Table A, we present the results of including ViT-Ti/16 and -S/16 [1] as backbones in our evaluation. Among previous SOTA methods, RVT [2] also utilizes these backbones and adjusts their architectures to improve robustness. To evaluate the effectiveness of our approach, we compared our method with RVT-Ti and -S. Despite introducing different sizes of backbones, our enhancement still holds, indicating that our proposed approach is scalable and adaptable to various backbone sizes.

B. Evaluation of Various Softmax Functions

	Clean	FGSM	PGD	A	R	C(↓)
Left	77.9	45.8	23.9	20.5	33.8	62.1
Mid	81.6	53.3	37.1	28.3	52.0	41.1
Right	**83.7**	**54.7**	**38.0**	**39.2**	**56.3**	**34.4**

Table B. The results of using different methods for applying the softmax function.

Table B presents a comparison of our method for applying the softmax function, which is situated on the right side of Fig. 2 (b), with the other two existing methods located in the middle and on the left side of the figure. The results consistently demonstrate the superiority of our method over the alternative methods.

C. The Number of Blocks with Adapters

# Blocks	Clean	A	R	C(↓)
12	83.7	39.2	56.3	34.4
6	82.8	32.2	51.5	37.4
1	80.5	24.8	45.2	45.8

Table C. The results of inserting our modules into various numbers of blocks.

In our paper, we use ViT-B/16 as our backbone, which consists of a total of 12 blocks. We inserted our proposed adapters and gated fusion modules into all 12 blocks. To explore the impact of the number of blocks on the performance of our approach, we conducted an ablation study, where we evaluated the effectiveness of inserting adapters into only the last 6 blocks or the last 1 block of the backbone. In Table C, the results indicate that using adapters in fewer blocks leads to a decline in performance. We use $\lambda = 0.5$ in this experiment.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022. 1