Supplementary Material for Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses

Yingwei Li[1], Song Bai[2], Cihang Xie[1], Zhenyu Liao[3], Xiaohui Shen[4], and Alan Yuille[1]

[1] Johns Hopkins University
[2] University of Oxford
[3] Kuaishou Technology
[4] ByteDance Research

The document contains the supplementary material for "Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses". The primary goal of this document is to provide an ablation study on K, the number of region partitions. Besides, we present additional quantitative results about training the gradient transformer module with different numbers of images, which further support our claim that the gradient transformer module becomes quasi-input-independent when training with a large number of images.

1 Number of Regions

In this section, we study the effect of K, the number of region partitions. Specifically, we split the image into 1196, 598, 299, 150, 100, 75, 50, 38, 25, 17, 8 or 4 regions. We show the learned universal perturbations in Figure 1. Due to the limitation of GPU memory, we study at most 1196 regions. The performance comparison is presented in Table 1. We observe that for stronger defenses (*e.g.*, PGD [4] and FD [7]), the optimal value of K is relatively small. We explain that these strong defenses have stronger ability to denoise (Some work [3, 7] interprets the defense procedure as denoising), while the perturbations with small K are less like a noise, thereby serving as strong perturbations for the strong defenses.

2 Quantitative Universal Results

Besides the quantitative results included in the main manuscript, we show some qualitative results in Figure 2. Specifically, we show and compare the perturbations generated by universal inference (use a fixed zero as the input of the gradient transformer module) and image-dependent (use the loss gradient as the input) inference via using the gradient transformer module trained with different numbers of images. We arrive at the same conclusion as the main manuscript, *i.e.*, the gradient transformer module becomes quasi-input-independent when training with a large number of images (*e.g.*, 5k images or more).

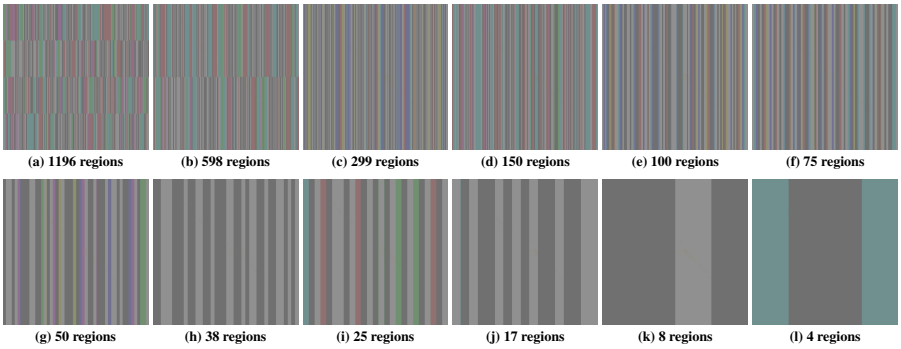

Fig. 1: Regionally homogeneous universal perturbations with different number of region partitions

Table 1: The increase of error rates (%) after attacking. The adversarial examples are generated with IncV3. In each row, we show the performance when splitting the images into a different number of regions

#regions	TVM [1]	HGD [3]	R&P [6]	Inc_{ens3} [5]	Inc_{ens4} [5]	$\text{IncRes}_{\text{ens}}$ [5]	PGD [4]	ALP [2]	FD [7]
1196	32.9	24.6	21.4	30.4	29.6	21.5	1.88	19.0	1.68
598	34.0	**27.2**	23.1	32.1	**32.0**	24.4	1.86	19.2	1.86
299	33.0	26.8	**23.3**	**32.5**	31.6	**24.6**	2.40	17.8	2.38
150	37.1	25.5	**23.3**	31.0	30.6	24.0	2.06	**20.3**	1.84
100	37.2	23.9	20.8	26.2	26.7	22.3	2.10	18.8	2.50
75	**39.0**	25.3	20.9	26.5	26.6	24.3	2.66	19.8	2.84
50	33.5	19.0	19.0	22.2	24.7	20.1	3.26	17.1	3.06
38	26.4	11.0	11.9	14.9	14.2	11.7	**3.88**	14.6	**3.62**
25	28.3	15.6	16.4	17.3	20.2	18.1	3.32	19.1	3.04
17	21.0	7.82	8.86	9.02	9.32	9.26	3.20	11.8	3.16
8	4.90	0.66	1.28	1.52	1.26	0.76	1.34	3.14	0.88
4	5.92	0.80	1.18	0.12	0.66	1.26	1.34	7.26	1.06

References

1. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: ICLR (2018)
2. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. In: NIPS (2018)
3. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: CVPR (2018)
4. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
5. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: ICLR (2018)
6. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: ICLR (2018)
7. Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: CVPR (2019)

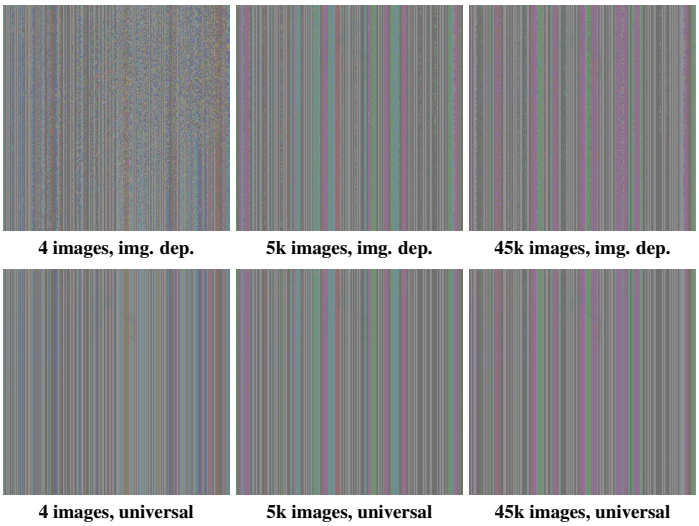

Fig. 2: A comparison of perturbations between image dependent inference (img. dep.) and universal inference (universal) when training with 4, 5k or 45k images. In the case of "4 images", the difference between image dependent inference and universal inference is large, while that is small when training with more images