Supplementary Material: Sparse Adversarial Attack via Perturbation Factorization

Yanbo Fan[1*], Baoyuan Wu[1*], Tuanhui Li[1], Yong Zhang[1], Mingyang Li[2], Zhifeng Li[1], Yujiu Yang[2]

[1] Tencent AI Lab
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University

1 Results of Attacking Adversarially Trained Model

We evaluate the sparse attack performance of SAPF on an adversarially trained model on CIFAR-10. The model structure is the same as that in Section 4.1 of the main manuscript. To this end, we first train this model using Madry's adversarial training method [4] with PGD-10 and $\epsilon = 8/255$. The trained model correctly classifies 752 images out of 1000 images that used in Section 4.1 of the main manuscript. We then conduct the targeted sparse attack to this model on these 752 benign images.

Method	Average case				
	ASR	ℓ_0	ℓ_1	ℓ_2	ℓ_∞
StrAttack [9]	51.73	1767	359.572	8.141	0.441
C&W-ℓ_0 [1]	**100**	607	77.153	4.748	0.731
SAPF (ours)	**100**	**563**	**36.493**	**2.033**	**0.384**

Table 1: Results of the sparse attack to an adversarially trained model on CIFAR-10. The best results are shown in bold.

The results are shown in Table 1. For better understanding, we also tabulate the performance of the C&W-ℓ_0 and the StrAttack that achieve 100% ASR on the model trained on benign images (see Table 1 of the main manuscript). The StrAttack fails to generate 100% ASR for this robust model. In comparison, our SAPF method successfully attacks all images with the minimal perturbations in terms of ℓ_p-norm ($p = 0, 1, 2, \infty$). Compared to the attack performance to the model trained on benign images (see Table 1 of the main manuscript), the adversarial training significantly increases the norm of perturbations for successful attack. However, the 100% ASR values of C&W-ℓ_0 and our method tell that the above adversarial training is still not robust enough to defend the sparse attack.

[*] indicates equal contribution. Correspondence to: Baoyuan Wu (wubaoyuan1987@gmail.com).

	One-Pixel [8]	CornerSearch [2]	PGD $\ell_0 + \ell_\infty$ [2]	SparseFool [5]	C&W-ℓ_0 [1]	StrAttack [9]	SAPF (ours)
CIFAR-10	3.33	1.23	0.19	0.54	11.33	79.71	63.19
ImageNet	217.22	464.41	1.46	16.56	318.89	1443.12	1138.89

Table 2: Average running time (seconds).

2 Comparison of Running Time

The average running time of attacking one image on CIFAR-10 and ImageNet is given in Table 2. On the CIFAR-10, the One-Pixel, the CornerSearch, the PGD $\ell_0 + \ell_\infty$, and the SparseFool are more time efficient than the other three methods. However, they fail to generate 100% attack success rate, and their ℓ_1-norm and ℓ_2-norm are much larger. The C&W-ℓ_0, the StrAttack, and our SAPF all achieve 100% ASR. Our SAPF method is slower than the C&W-ℓ_0, but faster than the StrAttack. Besides, the perturbations of the SAPF are the lowest in terms of ℓ_p-norm among these three methods. On the ImageNet, the running time of the One-Pixel and the CornerSearch increases a lot, yet they still fail to generate success attack in most cases. The SparseFool and the PGD $\ell_0 + \ell_\infty$ are more time efficient than our SAPF, however, they fail to generate 100% ASR and their perturbations are much larger than ours (see Table 1 of the main paper). Compared to the C&W-ℓ_0 and the StrAttack that achieve 100% ASR, our SAPF is slower than C&W-ℓ_0, but faster than StrAttack. And, our SAPF achieves 100% ASR with the lowest ℓ_p-norm. Besides, we would like to emphasize that the perspective of perturbation factorization of our SAPF brings multiple benefits: 1) it enables the more convenient control of the degree of sparsity; 2) it provides the extra flexibility to impose different constraints on perturbation magnitudes or selection factor. Both are helpful for the analysis of the sparse adversarial attacks.

3 Discussions

The Values of Sparse Adversarial Attack. Although several works on sparse adversarial attack have been published, it seems that the impact or value of sparse attack to the literature of adversarial examples has not been well studied. According to the open reviews for other works and our previous experiences, one typical question is that since the dense perturbations are also imperceptible to human eyes, what is the additional benefit of sparse perturbations? The work [9] has provided a good explanation via thorough experiments and analysis that the sparse attack provides a better interpretation of adversarial attack, *i.e.*, the clear correspondence between the perturbed regions and the discriminative regions in one image. This point is also verified in our experiments, as shown in Figs. 2 and 4 of the main manuscript. However, we think that the value of sparse adversarial attack is more than interpretation. It can serve as a powerful tool to separate the robust and discriminative features. For example, the sparse perturbation could evaluate the robustness of different regions, while the heatmap could evaluate

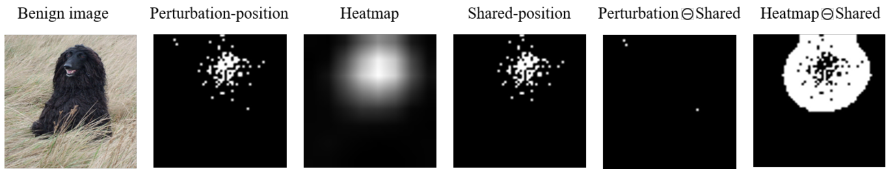

Fig. 1: One example of analyzing robust and discriminative regions. The perturbation-position (generated by the SAPF method) indicates the non-robust regions; The heatmap (generated by the Grad-CAM method [7]) tells the discriminative regions; The images of **4th - 6th columns** respectively indicate: the discriminative but non-robust regions; the non-robust and non-discriminative regions; the discriminative and robust regions.

the discrimination of different regions. Their combination could separate robust and discriminative regions. As shown in Fig. 1, the discriminative and robust positions locate at the surrounding region of the head of dog, where also includes non-robust positions, especially around the face of dog. The similar separation was studied in [3], utilizing the standard model and the adversarially trained model. Here the sparse attack provides another approach to implement such a separation. It will be thoroughly studied in our future work.

The Most Important Contribution of This Work. Finally, we would like to emphasize that the simple perspective of perturbation factorization[3] (see Eq. (2) of the main manuscript) is the most valuable and interesting contribution of this work to the literature of adversarial examples. Other contributions claimed at the end of Section 1 of the main manuscript, including the formulation to the MIP problem and the efficient continuous optimization algorithm, are derived from this perspective. Although this work mainly focuses on the sparsity of perturbations, this new perspective is not specially designed for enforcing sparsity. It could provide more potential benefits, such as incorporating the visual imperceptibility (studied in Section 3.4 of the main manuscript). More potential values of this simple perspective will be explored in our future work.

Other Extensions. The flexibility due to the perturbation factorization enables to enforce more constraints onto Problem (3) of the main manuscript, other than the group-wise sparsity and the visual imperceptibility studied in Section 3.4. For example, the attacker may require that some regions should not be perturbed. Although one can manually enforce the desired regions via a mask during each iteration of the conventional adversarial attack, our formulation provides a more elegant mechanism via the linear constraint (*i.e.*, $\mathbf{A}^\top \mathbf{G} = k$ with $\mathbf{A} \in \{0, 1\}$) in optimization. Another possible requirement is that the number

[3] Note that we factorize the perturbation on each pixel individually to the magnitude and the location, while [6] factorizes the perturbations on all pixels together to the magnitude and the direction, *i.e.*, $\boldsymbol{\epsilon} = \|\boldsymbol{\epsilon}\| \cdot \frac{\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|}$. They are intrinsically different and have different usages.

of perturbed pixels of one region should be more than that of another region, such as the background v.s. the foreground. It could be naturally embedded as an linear inequality constraint (*i.e.*, $\mathbf{A}^\top \mathbf{G} > 0$ with $\mathbf{A} \in \{-1, 0, 1\}$). In contrast, it is difficult to incorporate such non-accurate requirements in the conventional formulation of adversarial attack (see Eq. (1) of the main manuscript). Similar constraints could also be added onto perturbation magnitudes. More extensions of Problem (3) will be studied in our future work.

References

1. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
2. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: ICCV. pp. 4724–4732 (2019)
3. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems. pp. 125–136 (2019)
4. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
5. Modas, A., Moosavi-Dezfooli, S.M., Frossard, P.: Sparsefool: a few pixels make a big difference. In: CVPR. pp. 9087–9096 (2019)
6. Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E.: Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4322–4330 (2019)
7. 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: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
8. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation (2019)
9. Xu, K., Liu, S., Zhao, P., Chen, P.Y., Zhang, H., Erdogmus, D., Wang, Y., Lin, X.: Structured adversarial attack: Towards general implementation and better interpretability. ICLR (2019)