Supplement: Sparse and Imperceptible Adversarial Attack via a Homotopy Algorithm

In this appendix, we describe the parameter settings used in our algorithm for both pixel-wise sparsity and group-wise sparsity in Section 1 and 2 respectively. Moreover, we present more visualizations of pixel-wise sparsity and group-wise sparsity on targeted attack by our algorithm with different ℓ_∞ constraint ϵ in Section 4. For pixel-wise sparsity results, we also include the perturbations generated by GreedyFool (Dong et al., 2020) under the same ℓ_∞ constraint to establish a direct comparison between our algorithm and recent state-of-the-art.

1. Parameter Setting for Pixel-wise Sparsity

For the parameters of nmAPG (Li & Lin, 2015) used in our algorithm, eta, $delta$, and rho are set to 0.9, 0.3, and 0.8 respectively, and MaxIter is set to 100 when used in Algorithm 2, in all settings of our experiments.

In Algorithm 1 **Lambda_Search**, the decreasing factor in line 8 is set to 0.99 in all settings of our algorithm. For both targeted and nontargeted attacks on the ImageNet dataset, β is set to 10^{-6}. For both attacks on the CIFAR-10 dataset, β is set to 10^{-3}. While c is set to 3 for targeted attack and 7 for nontargeted attack on both datasets.

In Algorithm 2 **The Homotopy Attack Algorithm**, the upper bound v of the maximum ℓ_0 increases per outer iteration in the homotopy algorithm is set to 200 for ImageNet, and 20 for CIFAR-10 datasets for targeted attack, respectively. While for nontargeted attack, v is set to 50 on ImageNet, and 10 on CIFAR-10 datasets. Further, for both the targeted and nontargeted attacks, γ is set to 0.8 on ImageNet and 0.96 on CIFAR-10; the small number of v before entering the post attack stage in line 8 is set to 10 for ImageNet and 1 for CIFAR-10; the decreasing factor of λ in line 11 is set to 0.98 for ImageNet and 0.90 for CIFAR-10 datasets, respectively.

In the optional post attack stage of Algorithm 2, i.e., line 9, p is set to ∞ in all of our experiments. Moreover, for both the attacks, set $(w_1, w_2) = (10^{-2}, 10^{-4})$ on ImageNet, and set $(w_1, w_2) = (10^{-3}, 10^{-5})$ on CIFAR-10. Furthermore, given the ℓ_0 of the current perturbation, for both the attacks on ImageNet, the number of iterations of gradient descent to optimize Equation (13) is initialized with 200, with an end 1200, and is increased by 200 per 500 increase of the ℓ_0; while on CIFAR-10, it is initialized with 50, with an end 300, and is increased by 50 per 100 increase of the ℓ_0.

2. Parameter Setting for Group-wise Sparsity

We use SLIC Superpixels (Achanta et al., 2012) to segment groups for input images, we segment about 500 groups for every input image of ImageNet and 100 groups for every image of CIFAR-10 (the number of groups is not static because of the SLIC algorithm).

All parameters for group-wise sparsity on both the two datasets are set the same as those for pixel-wise sparsity, respectively, except that: on the CIFAR-10 dataset, β is set to 10^{-1}, c is set to 10^{-1}, and v is set to 1 all the time. While on the ImageNet, β is set to 10^{-2}, c is set to 10^{-2}, v is set to 1 all the time. Moreover, in the post attack for ImageNet, $(w_1, w_2) = (10^{-1}, 10^{-3})$, and the number of iterations of gradient descent to optimize Equation (13) is set to increase 500 with a maximum bound of 6000, per 1000 increase of the l_0 of the current perturbation.

To note, v in the group-wise sparsity case denotes the maximum number of groups added per outer iteration in the homotopy.

3. More Additional Details

Datasets The CIFAR-10 and ImageNet datasets can be obtained at `https://www.cs.toronto.edu/~kriz/cifar.html` and `http://image-net.org/download-images` respectively.

Computing infrastructures All our experiments are conducted on NVIDIA RTX 3080 GPU.

4. More Visualization Results

In this section we present more visual results of our pixel-wise sparsity and group-wise sparsity targeted attack in Figure 1, 2, 3, and 4.

In Figure 1 and 3, we present results of pixel-wise sparsity targeted attack on the same input images in Figure 1 of our main text with a ℓ_∞ threshold $\epsilon = 0.02$, and $\epsilon = 0.01$ respectively. We also compare our results with the state-of-the-art method GreedyFool (Dong et al., 2020). Their original parameter settings are used and the ℓ_∞ threshold is set the same as ours.

More results of our method's pixel-wise targeted attack with a ℓ_∞ threshold $\epsilon = 0.05$ are shown in Figure 4.

We also present results of group-wise sparsity targeted attack on the same input images in Section 3.7 of main text with a ℓ_∞ threshold $\epsilon = 0.02$, and visualize them in Fig. 2.

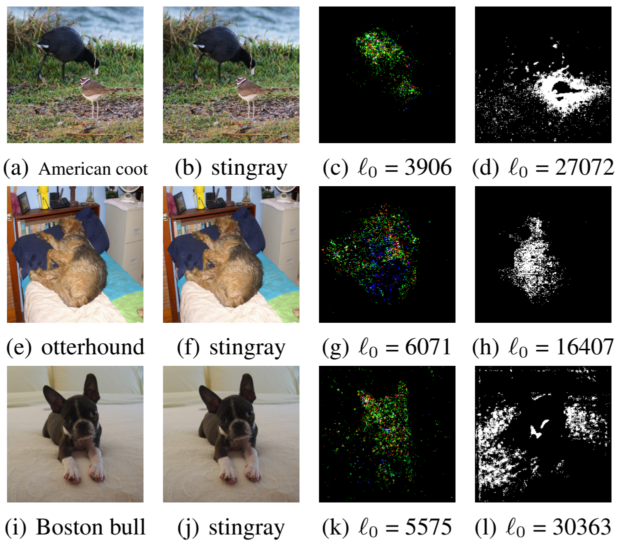

Figure 1. Visualization of pixel-wise sparsity of targeted attack when enforcing a ℓ_∞ constraint of 0.02. The images in each row, from the left to right are benign image, our adversarial example, our perturbation position, GreedyFool's perturbation position.

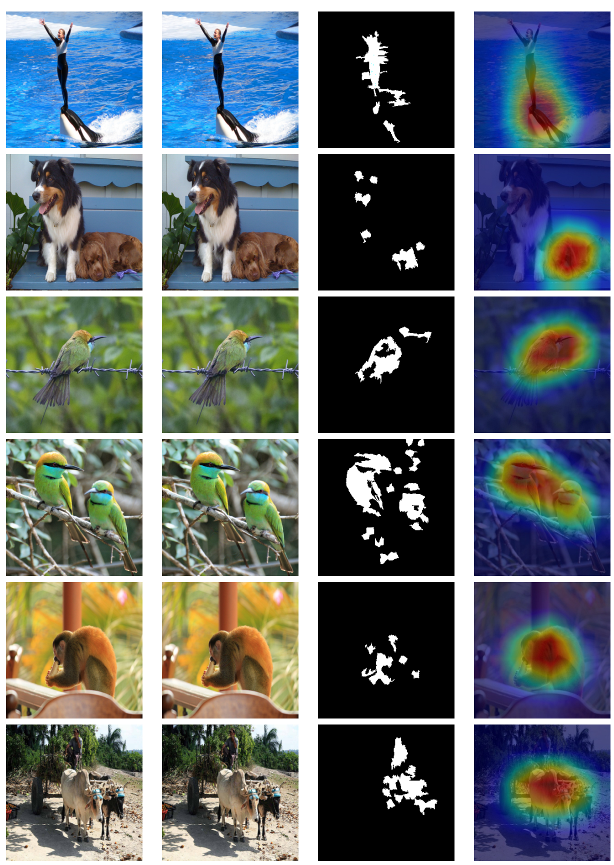

Figure 2. Visualization of targeted attack with group-wise sparsity threshold 0.02. Each row from the left to right represents benign image, adversarial example, perturbation positions, and CAM of the original label.

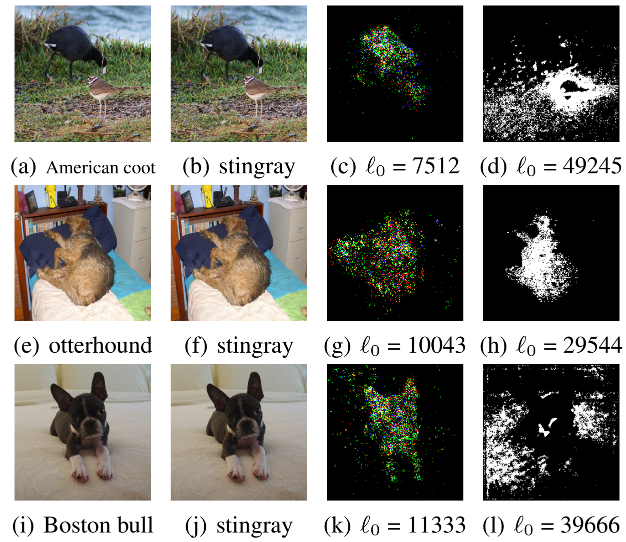

Figure 3. Visualization of pixel-wise sparsity of targeted attack when enforcing a ℓ_∞ constraint of 0.01. The images in each row, from the left to right are benign image, our adversarial example, our perturbation position, GreedyFool's perturbation position.

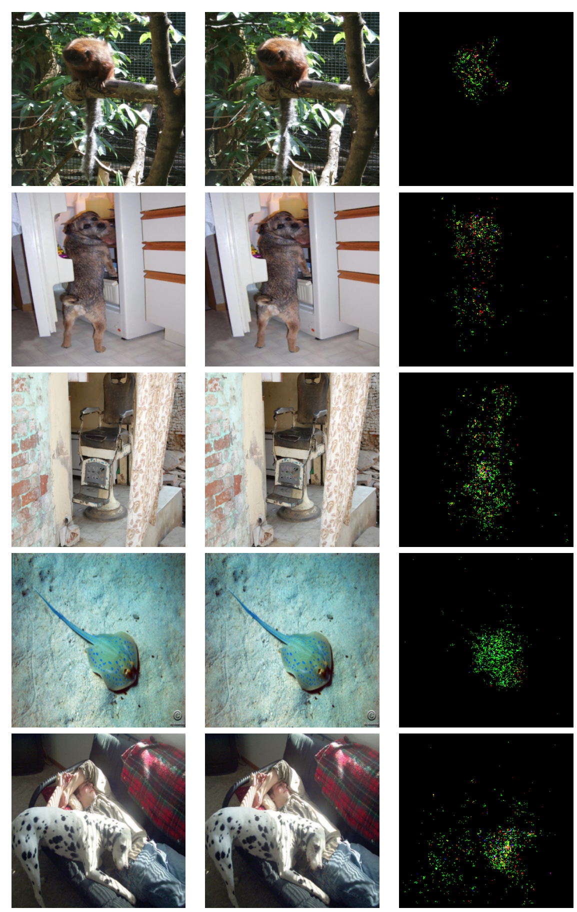

Figure 4. Visualization of pixel-wise sparsity of targeted attack when enforcing a ℓ_∞ constraint of 0.05. The images in each row, from the left to right are benign image, our adversarial example, our perturbation position.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

Dong, X., Chen, D., Bao, J., Qin, C., Yuan, L., Zhang, W., Yu, N., and Chen, D. Greedyfool: Distortion-aware sparse adversarial attack. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4724–4732, 2020.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 379–387, 2015.