Structure Invariant Transformation for better Adversarial Transferability

Xiaosen Wang[1], Zeliang Zhang[2], Jianping Zhang[3]
[1]Huawei Singularity Security Lab & [2]HUST & [3]Chinese University of Hong Kong

ICCV23 PARIS

Paper, code, and data are available:
https://github.com/xiaosen-wang/SIT

Motivation

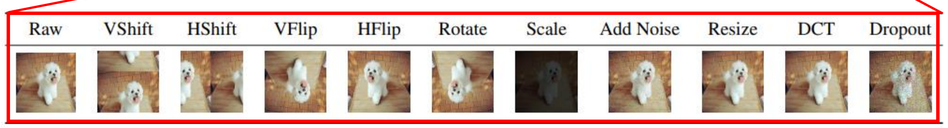
Assumption: Without harming the semantic information, the more diverse the transformed image is, the better transferability the adversarial examples have.

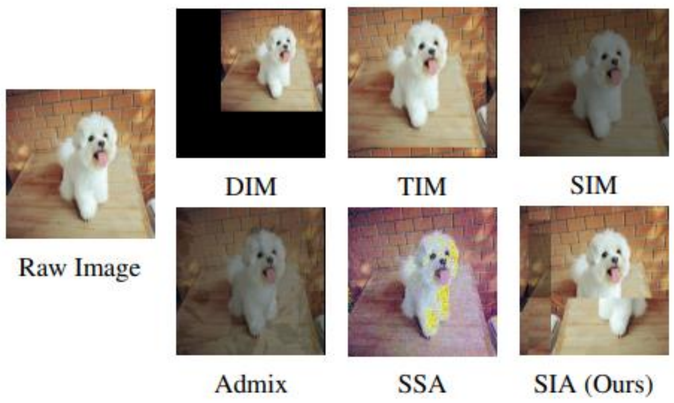
	TIM	DIM	SIM	SSA	*Admix*
Transferability	57.4	77.6	79.3	80.6	83.6
LPIPS	0.25	0.43	0.48	0.54	0.73

$$\text{LPIPS}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{H \times W} \sum_{l} \sum_{h,w} \|\boldsymbol{z}^l_{h,w} - \hat{\boldsymbol{z}}^l_{h,w}\|_2$$

The Struture of Image

Defination: Given an image x, which is randomly split into *s* x *s* blocks, the relative relation between each anchor point is the structure of image, where the anchor point is the center of the image block.





Structure Invariant Attack

Algorithm 1: Structure Invariant Attack

Input: Classifier $\boldsymbol{f}(\cdot)$ with the loss function J; The benign sample $\boldsymbol{x}$ with ground-truth label y; The maximum perturbation ϵ, number of iterations T and decay factor μ; Splitting number s; Number of transformed images N

Output: An adversarial example.

$\alpha = \epsilon/T,\ \boldsymbol{g}_0 = 0,\ \boldsymbol{x}^{adv}_0 = \boldsymbol{x}$

for $t = 0 \to T-1$ **do**

Constructing a set $\mathcal{X}$ of N transformed images using SIT

Calculating the average gradient on $\mathcal{X}$:

$$\bar{\boldsymbol{g}}_{t+1} = \frac{1}{N} \sum_{\boldsymbol{x}_i \in \mathcal{X}} \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_i, y) \quad (2)$$

Updating the momentum:

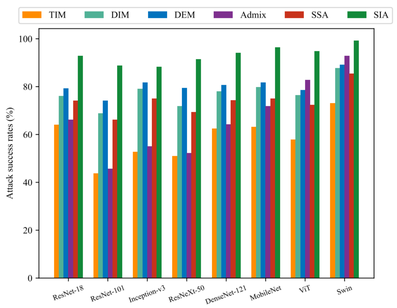
$$\boldsymbol{g}_{t+1} = \mu \boldsymbol{g}_t + \frac{\bar{\boldsymbol{g}}_{t+1}}{\|\bar{\boldsymbol{g}}_{t+1}\|_1} \quad (3)$$

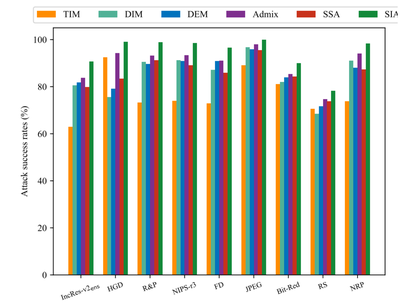
Updating the adversarial example:

$$\boldsymbol{x}^{adv}_{t+1} = \text{Clip}(\boldsymbol{x}^{adv}_t + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1}), 0, 1) \quad (4)$$

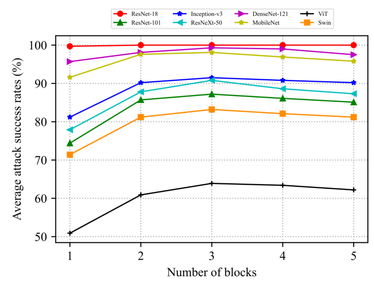
return $\boldsymbol{x}^{adv}_T$

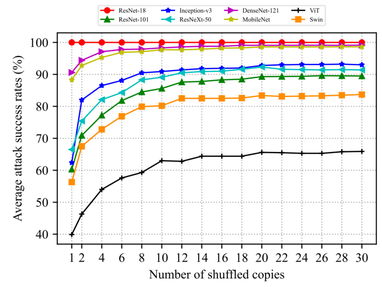
Attacking Ensemble Models and Defense Methods





Ablation study





SIA	-VShift	-HShift	-VFlip	-HFlip	-Rotate	-Sclae	-Add Noise	-Resize	-DCT	-Dropout
92.1	89.7	90.1	90.1	88.4	88.3	90.1	90.6	90.2	90.1	90.7

Attacking a Single Model