Supplementary Material: Tracing the Origin of Adversarial Attack for Forensic Investigation and Deterrence

Han Fang[1] Jiyi Zhang[1] Yupeng Qiu[1] Jiayang Liu[1]
Ke Xu[2] Chengfang Fang[2] Ee-Chien Chang[1,*]
[1]National University of Singapore [2]Huawei International
{fanghan, ljyljy}@nus.edu.sg {jiyizhang, qiu_yupeng}@u.nus.edu
{xuke64, fang.chengfang}@huawei.com changec@comp.nus.edu.sg

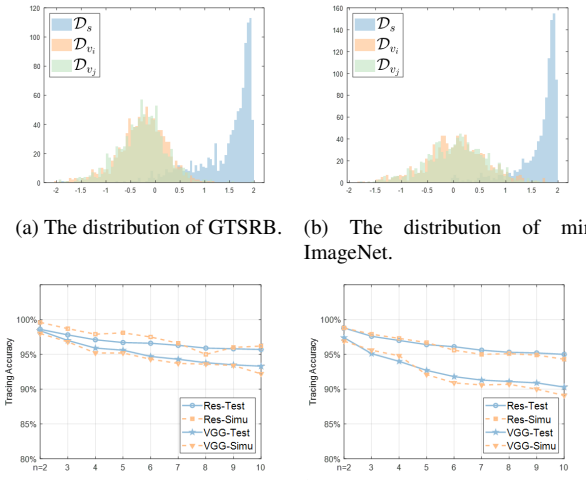
Abstract

1. Simulation results of GTSRB and mini-ImageNet

In this section, we mainly show the simulation and real tracing results of multiple copies with the datasets GTSRB and mini-ImageNet. We conduct the same experiments on GTSRB and mini-ImageNet as we introduced in Section 4.4, the results of the distribution of DOL and the tracing results are shown in Fig. 1.

From Fig. 1 we can see that the results with datasets GTSRB and mini-ImageNet are similar to the result of CIFAR10. The DOL of victim tracers $\mathcal{T}_{v_i}$s follows the same distribution, and the DOL of victim tracers $\mathcal{T}_{v_i}$ and that of the source tracers $\mathcal{T}_s$ follows the different distributions. As for the tracing results, it can be seen that the simulated results are similar to the real test results, which indicates that the tracing performance can be estimated by the distributions, such a performance is also the same as that of CIFAR10. Besides, we also can see that with "ResNet" backbone, the tracing performance of GTSRB and mini-ImageNet (10 copies) both can achieve 95%, which indicates the effectiveness of the proposed method.

*Corresponding author.



(a) The distribution of GTSRB. (b) The distribution of mini-ImageNet.

(c) The tracing results of GTSRB. (d) The tracing results of mini-ImageNet.

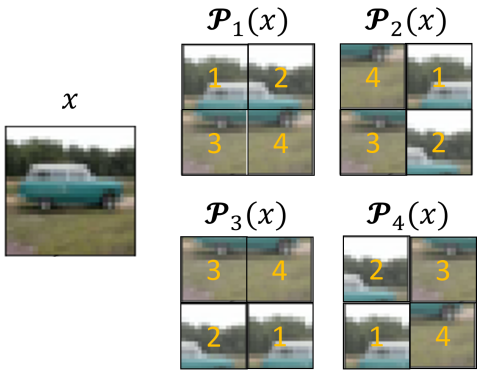
Figure 1: The distribution of DOL with HSJA and ResNet backbone and tracing performance of multiple branches.

2. Scrambling-based method

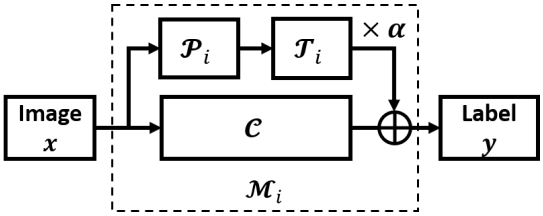
As introduced in Section 4.4, when $u = 1$, the upper bound of the valid copy numbers is $K \times (K-1)$, which is quite small when K is small. In this section, we mainly introduced a scrambling-based method which could effectively enlarge the copy numbers. The specific algorithm is introduced as follows: After we obtained $\mathcal{T}_0$, we can generate each $\mathcal{T}_i$ by concatenating a pre-permutation operation $\mathcal{P}_i$ with $\mathcal{T}_0$:

$$\mathcal{T}_i(x) = \mathcal{T}_0(\mathcal{P}_i(x)) \tag{1}$$

where $\mathcal{P}_i$ indicates the i^{th} scrambling, x indicates the input images. $\mathcal{P}$ should satisfy: no two scrambled images "overlap" more than th pixels where th is a pre-defined constant



(a) The example of the proposed scrambling algorithm.



(b) The workflow of the proposed scrambling-based method.

Figure 2: The scrambling-based method.

which controls the trade-off between the copy number and the tracing performance. That is, for $\mathcal{P}_i$ and $\mathcal{P}_j$ $(i \neq j)$:

$$\mathcal{P}_i(x) \cap \mathcal{P}_j(x) \leq th$$

An example of $\mathcal{P}_i$ is illustrated in Fig. 2a. And the workflow of the scrambling-based method is shown in Fig. 2b.

Besides, we also test the tracing performance of ***5 distributed models*** with scrambling-based methods (by dividing the image into non-overlapped 4×4 blocks and random scrambling with $th = 0$). The backbone we used is "ResNet", the attack we choose is HSJA [1] and QEBA [2]. α is fixed as 0.15. The results are shown in Table 1.

Table 1: The trace accuracy of scrambling-based algorithm with 5 copies.

Attack	CIFAR10	GTSRB	mini-ImageNet
HSJA	98.8%	97.2%	96.5%
QEBA	99.8%	98.5%	96.7%

The results in Table 1 illustrate the effectiveness of the scrambling-based algorithm, we can see that the tracing accuracy is higher than 96% for each task with 5 distributed models. Such a result is similar to the performance of the permutation-based method.

Besides, the scrambling-based method is independent and can be combined with the permutation-based method. Based on the combination, the valid copy numbers can be effectively enlarged. In order to show the independence, we also test the tracing performance with the tracer which is a combination of the scrambling-based method and permutation-based method, i.e. $\pi_i(\mathcal{T}_0(\mathcal{P}_j(x)))$. We randomly choose ***10 copies***, and the tracing performance is shown in Table 2.

Attack	Method	CIFAR10	GTSRB	mini-ImageNet
HSJA	Permutation	97.3%	95.4%	94.4%
	Scrambling	97.9%	95.8%	94.6%
	Combined	98.0%	95.4%	94.9%
QEBA	Permutation	98.8%	95.7%	95.0%
	Scrambling	99.1%	96.6%	95.6%
	Combined	98.7%	96.4%	95.4%

Table 2: The trace accuracy of different tracer generation methods with 10 copies.

Table 2 illustrated that the performance of the combined tracer is similar to the scrambling-based-only tracer and the permutation-based-only tracer, which indicates the combination will not influence the performance, and these two methods can be applied cooperatively. So applying different permutations and scramblings can effectively enlarge the valid number of the distributed models, which will be $O \times K \times (K-1)$ (O indicates the total scrambling number).

3. The importance of $\mathcal{L}_{Trap}$

The most important loss in training the tracer is $\mathcal{L}_{Trap}$, which is the key to inducing attacks. Here we evaluate the importance of $\mathcal{L}_{Trap}$. Specifically, we use 5 randomly initialized $\mathcal{T}_i$ to conduct the tracing experiment on 1000 adversarial images. The attack we choose is HSJA [1], α is fixed as 0.15. The experimental results are shown in Table 3.

Table 3: The trace accuracy of HSJA attack with/without $\mathcal{L}_{Trap}$.

Attack	CIFAR10		GTSRB		mini-ImageNet	
	ResNet18	VGG16	ResNet18	VGG16	ResNet50	VGG19
Random	32.6%	17.1%	14.9%	20.3%	19.3%	16.2%
Proposed	97.5%	90.2%	96.4%	95.2%	94.2%	90.7%

We can see that without $\mathcal{L}_{Trap}$, the tracing accuracy only achieves up to 32.6%, which is much lower than training with $\mathcal{L}_{Trap}$. This indicates that $\mathcal{L}_{Trap}$ is very important in realizing accurate tracing, only pairing a random network with the main classifier is not enough to trap the attack to result in specific features.

References

[1] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (S&P)*, pages 1277–1294. IEEE, 2020. 2

[2] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1221–1230, 2020. 2