Supplementary Material for "RIBAC: Towards Robust and Imperceptible Backdoor Attack against Compact DNN"

Huy Phan[1], Cong Shi[1], Yi Xie[1], Tianfang Zhang[1], Zhuohang Li[2], Tianming Zhao[3], Jian Liu[2], Yan Wang[3], Yingying Chen[1], and Bo Yuan[1]

[1] Rutgers University, New Jersey, USA
[2] The University of Tennessee, Tennessee, USA
[3] Temple University, Pennsylvania, USA

1 Ablation study on Trigger Stealthiness.

We examine the effect of varying the maximum allowed perturbation ϵ on compression performance and attack performance. It is seen from Figure A1 that smaller values of ϵ (from 1/255 to 3/255), while can offer better stealthiness, suffer from degraded compression performance and attack performance. Higher value of ϵ (5/255) does not offer additional performance in both metrics. Hence, we believe that our default value of $\epsilon = 4/255$ gives a good balance between stealthiness, compression performance and attack performance.

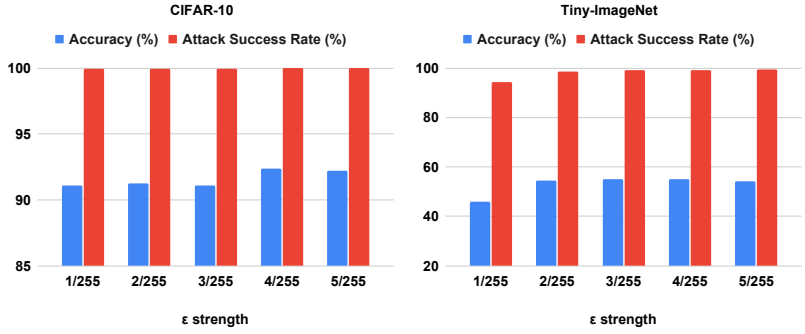

Fig. A1. Performance of RIBAC with varying trigger stealthiness (ϵ) on CIFAR-10 and Tiny-ImageNet dataset.

2 Ablation study on Number of Training Epochs.

We study the effect of changing the number of training epochs of **Step − 1** in Eq. (6) and **Step − 2** in Eq. (7). We can observe from Figure A2 that we can already

achieve very high attack performance only by using a small number of training epochs. However, to recover the clean accuracy of the pre-trained models, more training epochs are needed. Since the clean accuracy stop increase after 60^{th} epoch, it is seen that our default value of using 60 training epochs for RIBAC offer a good balance between efficiency and performance.

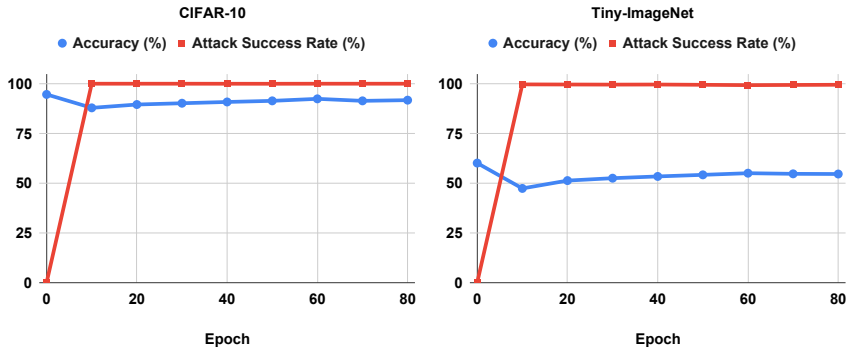

Fig. A2. Performance of RIBAC with varying number of training epochs on CIFAR-10 and Tiny-ImageNet dataset.

3 Visual results of RIBAC backdoor images and triggers.

To demonstrate the stealthiness of RIBAC backdoor images using different datasets, we show the clean images, backdoor images, and amplified triggers in Figure A3, Figure A4, Figure A5. It is seen that RIBAC backdoor images are visually indistinguishable from the clean images.

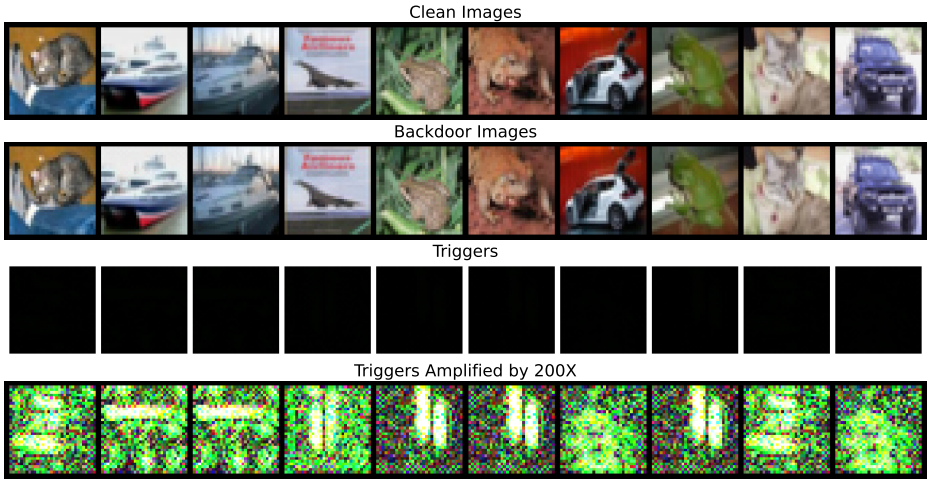

Fig. A3. Clean images, RIBAC backdoor images, and RIBAC amplified triggers on CIFAR-10 dataset.

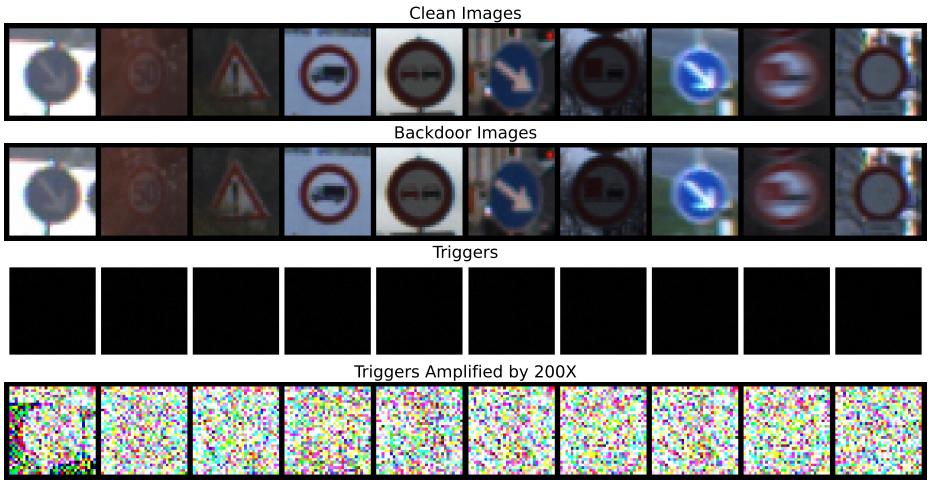

Fig. A4. Clean images, RIBAC backdoor images, and RIBAC amplified triggers on GTSRB dataset.

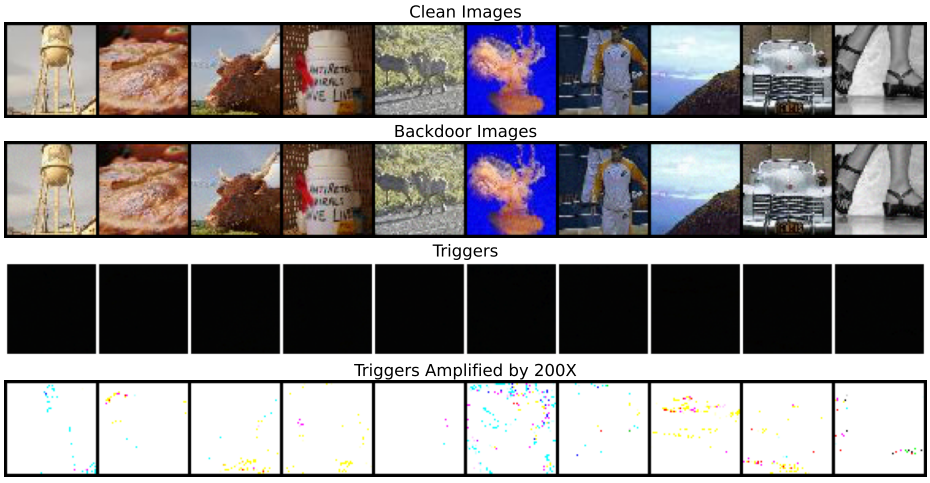

Fig. A5. Clean images, RIBAC backdoor images, and RIBAC amplified triggers on Tiny-ImageNet dataset.