Hardly Perceptible Trojan Attack against Neural Networks with Bit Flips (Appendix)

Jiawang Bai[1], Kuofeng Gao[1], Dihong Gong[2], Shu-Tao Xia[1,3,✉], Zhifeng Li[2,✉], and Wei Liu[2,✉]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Data Platform, Tencent
[3] Research Center of Artificial Intelligence, Peng Cheng Laboratory
{bjw19,gkf21}@mails.tsinghua.edu.cn; gongdihong@gmail.com;
xiast@sz.tsinghua.edu.cn; michaelzfli@tencent.com; wl2223@columbia.edu

A Parameter Settings

On CIFAR-10 and SVHN, γ and b are set to 1000 and 10 for ResNet-18, and 100 and 20 for VGG-16. On ImageNet, γ and b are set to 10^5 and 10. For the optimization in Algorithm 1, we update $\boldsymbol{\delta}^{[k+1]}$ and $\boldsymbol{f}^{[k+1]}$, and $\hat{\boldsymbol{\theta}}^{[k+1]}$ for 5 gradient steps during each iteration with the learning rate 10^{-5}, 10^{-5}, and 10^{-4}, respectively. The ρ is initialized as 10^{-4} and increased by $\rho \leftarrow \max(100, 1.01 \times \rho)$ after each iteration. When $\max(||\hat{\boldsymbol{\theta}} - \boldsymbol{z}_1||_2^2, ||\hat{\boldsymbol{\theta}} - \boldsymbol{z}_2||_2^2)$ is smaller than a preset threshold (10^{-4} on CIFAR-10 and SVHN, and 10^{-6} on ImageNet) or the maximum number of iterations ($3,000$ on CIFAR-10 and SVHN, and $5,000$ on ImageNet) is reached, the optimization halts.

B Attacking 4-bit Quantized DNNs

Table I: Performance of HPT in attacking 4-bit Quantized DNN. The target class t is set as 0.

Dataset	**Model**	**TA** (%)	**PA-TA** (%)	**ASR** (%)	$\mathbf{N_{flip}}$
CIFAR-10	ResNet-18	94.2	94.1	94.6	10
	VGG-16	92.6	82.2	93.8	9
SVHN	VGG-16	96.3	95.3	74.3	20
ImageNet	ResNet-18	66.3	64.9	94.6	12

In this section, we present the results of HPT in attacking 4-bit quantized DNNs in Table I. It shows that HPT can obtain promising attack performance in all cases. For example, in attacking ResNet-18 on CIFAR-10, HPT achieves a 94.6% ASR with 10 bit flips, while only 0.1% accuracy degradation of original images. Overall, the results of attacking 4-bit quantized DNN are consistent with these of attacking 8-bit quantized DNN.

C More Details of Human Perceptual Study

Table II: Criteria of scoring the Trojan images.

Score	Criterion
1	This image is very easy to be distinguished from the original one.
2	This image is easy to be distinguished from the original one.
3	This image can be distinguished from the original one by the carefully inspection.
4	This image is hard to be distinguished from the original one.
5	This image is very hard to be distinguished from the original one.

We randomly select 10 clean images from each dataset and generate the corresponding Trojan images for these 30 images. In our study, all original and Trojan images are shown to 15 participants. These participants are asked to give a score $\in \{1, 2, 3, 4, 5\}$ for each Trojan image, where a higher score corresponds to less perceptible Trojan images. The participants are asked to score each Trojan image using the criteria listed in Table II. In total, we collect 2,250 scores and analyze them to evaluate the perceptibility of Trojan images crafted by different methods.

D Quantitative Results for Perceptibility of Trojan Images

Besides the human perceptual study, we use the mean square error (MSE) to measure the perceptibility of Trojan images. We calculate MSE distances between the original images and their Trojan images (in the range $[0, 255]$) for five attack methods on 500 randomly selected clean images, as shown in Table III. The MSE results of our HPT are much lower than all other methods. These quantitative results further verify that HPT is hardly perceptible.

E Complexity Analysis

The pipeline of HPT consists of two parts: Trojan injection and inference. Their complexity is analyzed as below.

Trojan Injection. We first discuss the cost of the optimization algorithm for the proposed HPT. Since the updates of $\boldsymbol{z}_1^{[k+1]}$, $\boldsymbol{z}_2^{[k+1]}$, $z_3^{[k+1]}$, $\boldsymbol{\lambda}_1^{[k+1]}$, $\boldsymbol{\lambda}_2^{[k+1]}$, and $\lambda_3^{[k+1]}$ are very simple, the costs for these variables are omitted here. During per iteration, the main cost is from the forward and backward pass. The complexity of the forward pass depends on the attacked model g. Since we only optimize the parameters of the last layer and fix the others, for the Q-bit quantized DNN,

Table III: Perceptibility of Trojan Images crafted by five attack methods (measured by MSE distances). Lower values are better.

Dataset	**TrojanNN**	**TrojanNN** ($Trans.$)	**TBT**	**TBT** ($Trans.$)	**HPT**
CIFAR-10	1305.0	117.4	1264.4	113.8	**88.8**
SVHN	1381.2	124.3	1317.6	118.6	**94.5**
ImageNet	1645.8	146.3	1574.8	140.7	**108.1**

Table IV: Running time of Trojan injection (without considering physical bit flips in the memory) and inference ($time/image$) for the proposed HPT.

Trojan Injection	Inference
$2302.4 \pm 198.9\ s$	$14.4 \pm 4.4\ \mu s$

the computation cost of updating $\hat{\boldsymbol{\theta}}^{[k+1]}$ is $\mathcal{O}(2MFKQ)$ with M clean images and M Trojan images, where F is the dimension of the last layer's input and K is the number of classes. In terms of $\boldsymbol{\delta}^{[k+1]}$ and $\boldsymbol{f}^{[k+1]}$, the costs of their updates are the same as that of the standard backpropagation for the model g. Note that we update $\hat{\boldsymbol{\theta}}^{[k+1]}$, $\boldsymbol{\delta}^{[k+1]}$ and, $\boldsymbol{f}^{[k+1]}$ using a same backward pass. Due to the fast convergence of our optimization in practice and a small set of clean images (128 for CIFAR-10 and SVHN, and 256 for ImageNet), the running time of our optimization is acceptable. After identifying the critical bits, HPT changes these bits in the memory. Due to a small set of bit flips, our attack performs efficiently according to [1].

Inference. After the Trojan injection, the main difference between HPT and normal inference is to craft the Trojan images based on the clean images. Since we apply the same $\boldsymbol{\delta}$ and $\boldsymbol{f}$ on all images, this process is very efficient.

We run HPT on CIFAR-10 with ResNet-18 architecture for 5 times and report the results in Table IV. The experiments are conducted on one GeForce RTX 2080Ti GPU.

F More Visualization

We provide more visualization examples on CIFAR-10, SVHN, and ImageNet in Figure I, II, and III, respectively. As can be observed, compared to other methods, the triggers of Trojan images generated by HPT are most natural and unnoticeable in most cases.

Fig. I: Visualization of clean and Trojan images generated by different methods on CIFAR-10.

Fig. II: Visualization of clean and Trojan images generated by different methods on SVHN.

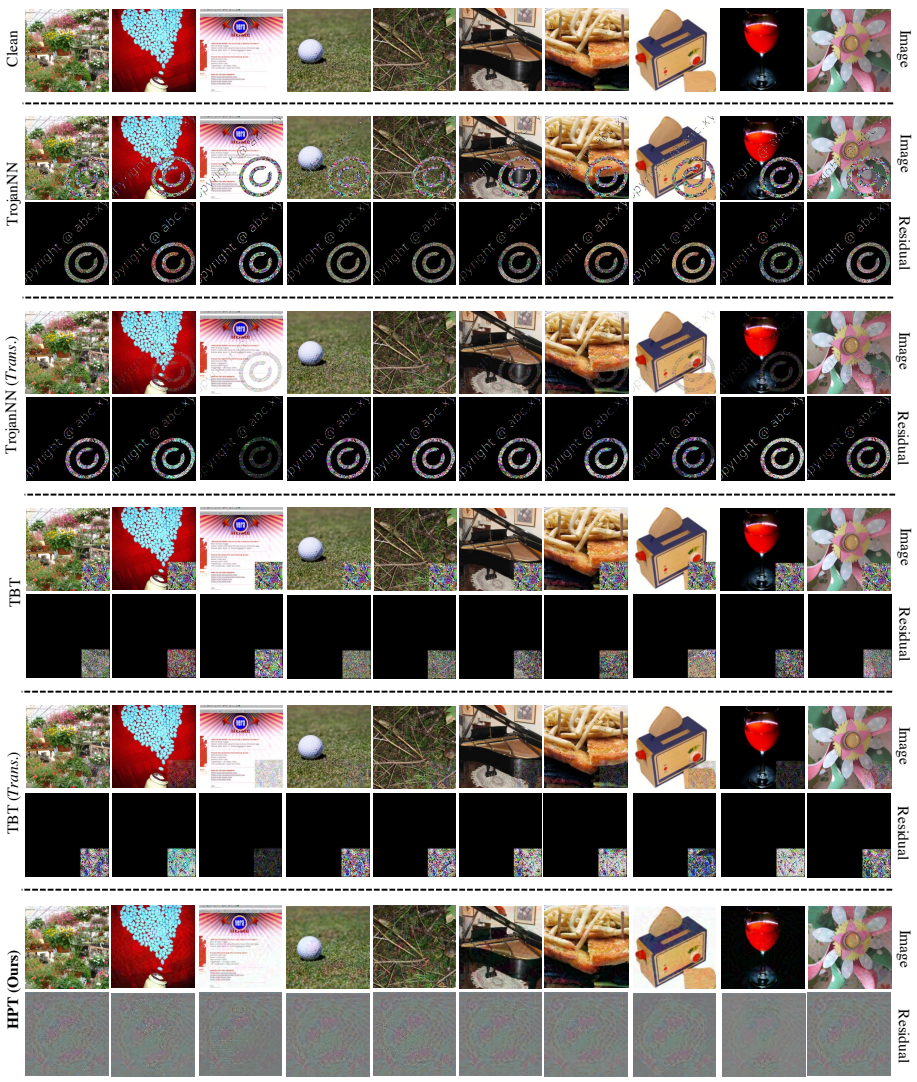

Fig. III: Visualization of clean and Trojan images generated by different methods on ImageNet.

References

1. Yao, F., Rakin, A.S., Fan, D.: Deephammer: Depleting the intelligence of deep neural networks through targeted chain of bit flips. In: USENIX Security Symposium (2020)