Supplementary Materials for LIRA: Learnable, Imperceptible and Robust Backdoor Attacks

Khoa Doan, Yingjie Lao, Weijie Zhao, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{khoadoan106, laoyingjie, zhaoweijie12, pingli98}@gmail.com

This document provides additional details, analysis, and experimental results to support the main submission. We begin by providing additional discussion on the related works in Section 1. Next, we discuss the detailed experimental setup and implementation of the methods in Section 2. Finally, we provide additional attack and defense experiments, as well as sensitivity analysis of the proposed algorithm in Section 3.

1. Related Methods

In this section, we provide an additional detailed discussion on other related methods [12, 2]. LIRA's algorithm is based on alternating updates between the two functions f and T, which has crucial differences to these related adversarial and backdoor approaches.

The clean-label backdoor work of [12] assumes a different threat model than LIRA and WaNet. Specifically, [12] focuses on poisoning the training data, where the classifier is trained by the user. Furthermore, [12] uses GAN to interpolate an image toward the target class's images, but still manually design some less visible patch-based trigger that is superimposed on the perturbed images. This is because their objective is to make a triggered image look natural while having the label consistent with its content. However, to achieve higher success rates, the perturbed images must be interpolated very far away from the original image and closer to the target images. This makes the interpolated images visually unnatural.

The work in [2] proposes an adversarial framework, DeepConfuse, to learn a function to generate noise-perturbed images. The goal of DeepConfuse is to prevent the released data from being illegally used to train a model without the data owner's permission. For such a reason, their objective is to ensure that training any classifier on the released, noise-perturbed data will fail; i.e., the classifier's accuracy will be very low (compared to the classifier trained on clean data). This is very different from our objective, where we aim to inject a backdoor while preserving the classifier's performance on the clean data. As the authors indicated in the paper of DeepConfuse, the backdoor attack is a different and more difficult adversarial attack than DeepConfuse.

2. Detailed Experimental Setup

2.1. Datasets

As we described in the main paper, we use four datasets, MNIST, CIFAR10, GTSRB, and T-ImageNet, to evaluate our method. Note that MNIST, CIFAR10, and GTSRB have been widely used in the literature of backdoor attacks on DNN. On the other hand, the use of a more complex dataset, T-ImageNet, enables better evaluation for multiple-target backdoor attacks such as all-to-all, thanks to the diversity of images in T-ImageNet and its large number of classes.

- **MNIST** [6] is a subset of the larger dataset available from the National Institute of Technology. This dataset (found here[1]) consists of 70,000 grayscale, 28×28 images, divided into a training set of 60,000 images and a test set of 10,000 images. We applied random cropping and random rotation as data augmentation for the training process. During the evaluation stage, no augmentation is applied.
- **CIFAR-10** is first introduced by [5]. It is a labeled subset of the 80-millions-tiny-images dataset, collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, consists of 60,000 color images at the resolution of 32×32, out of which 10,000 images are randomly selected as the query set, and the remaining images used as the retrieval set. The data set is public[2].
- **GTSRB** (German Traffic Sign Recognition Benchmark [10]) is used as an official dataset for the challenge

[1] http://yann.lecun.com/exdb/mnist
[2] https://www.cs.toronto.edu/ ~kriz/cifar.html

Layer	Filters	Filter Size	Stride	Padding	Activation
Conv2D	16	3×3	3	1	BatchNorm2D+ReLU
MaxPool2d	-	2×2	2	0	-
Conv2D	64	3×3	2	1	BatchNorm2D+ReLU
MaxPool2d	-	2×2	2	0	-
ConvTranspose2D	128	3×3	2	-	BatchNorm2D+ReLU
ConvTranspose2D	64	5×5	3	1	BatchNorm2D+ReLU
ConvTranspose2D	1	2×2	2	1	BatchNorm2D+Tanh

Table 1: Autoencoder-based generator network used in this paper.

	MNIST	CIFAR10	GTSRB	T-ImageNet
OTHERS				
Optimizer	SGD	SGD	SGD	SGD
Batch Size	128	128	128	128
Learning Rate	0.01	0.01	0.01	0.01
Learning Rate Schedule	100,200,300,400	100,200,300,400	100,200,300,400	100,200,300,400
Learning Rate Decay	0.1	0.1	0.1	0.1
Training Epochs	1000 epochs	1000 epochs	1000 epochs	1000 epochs
LIRA Only				
ϵ	0.005	0.005	0.005	0.005
α	0.5	0.5	0.5	0.5
β	0.5	0.5	0.5	0.5
k	1 epoch	1 epoch	1 epoch	1 epoch
m	50 epochs	50 epochs	50 epochs	50 epochs
T's Optimizer	SGD	SGD	SGD	SGD
T's Learning Rate	0.001	0.001	0.001	0.001
Clean Accuracy	0.99	0.94	0.99	0.57

Table 2: Experiment setup and parameters for the datasets used in this paper.

held at the International Joint Conference on Neural Network (IJCNN) 2011. GTSRB[3] consists of 60,000 images, divided in 43 classes, with resolutions varying from 32×32 to 250×250. The training set contains 39,209 images, while the test set has 12,630. In our experiments, GTSRB input images are all resized into 32×32 pixels, then applied random crop and random rotation in training. In the evaluation stage, no augmentation is used.

- **Tiny-ImageNet (T-ImageNet)** is a smaller subset of the large-scale ImageNet dataset [1] and is introduced in [14]. This dataset consists of 200 images classes. The training set has 500 images per class, resulting in 100,000 images, while the test set has 50 images per class,resulting in 10,000 images. T-ImageNet input images are all resized into 64×64 resolution. Random crop and random rotation are applied in the training stage. No augmentation is used in the evaluation stage.

2.2. Noise Generator Models

For MNIST, we use a self-defined autoencoder, which is detailed in Table 1. For the other datasets, we employ the UNet architecture [9]. We observe only a slight performance difference between the simpler autoencoder and the complex UNet on these datasets.

2.3. Training Hyperparameters

Table 2 provides additional details to Section 5.1 in the main paper.

2.4. Classification Models

In this work, we use a simple CNN classifier, which is previously used in WaNet [8], for MNIST. For convenience, we include the detailed architecture in Table 3. For CIFAR10 and GTSRB datasets, we use PreActResnet18 [4]. For T-ImageNet, we use Resnet18 [4].

Layer	Filters	Filter Size	Stride	Padding	Activation
Conv2D	32	3×3	2	1	ReLU
Conv2D	64	3×3	2	0	ReLU
Conv2D	64	3×3	2	0	ReLU
Linear	512	-	-	-	ReLU
Conv2D	10	-	-	-	Softmax

Table 3: CNN model architecture for MNIST.

[3] http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset

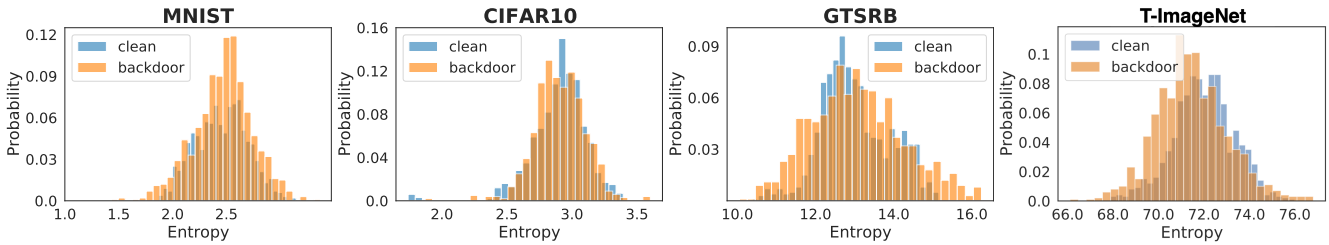


Figure 1: All-to-all attacks against STRIP defense.

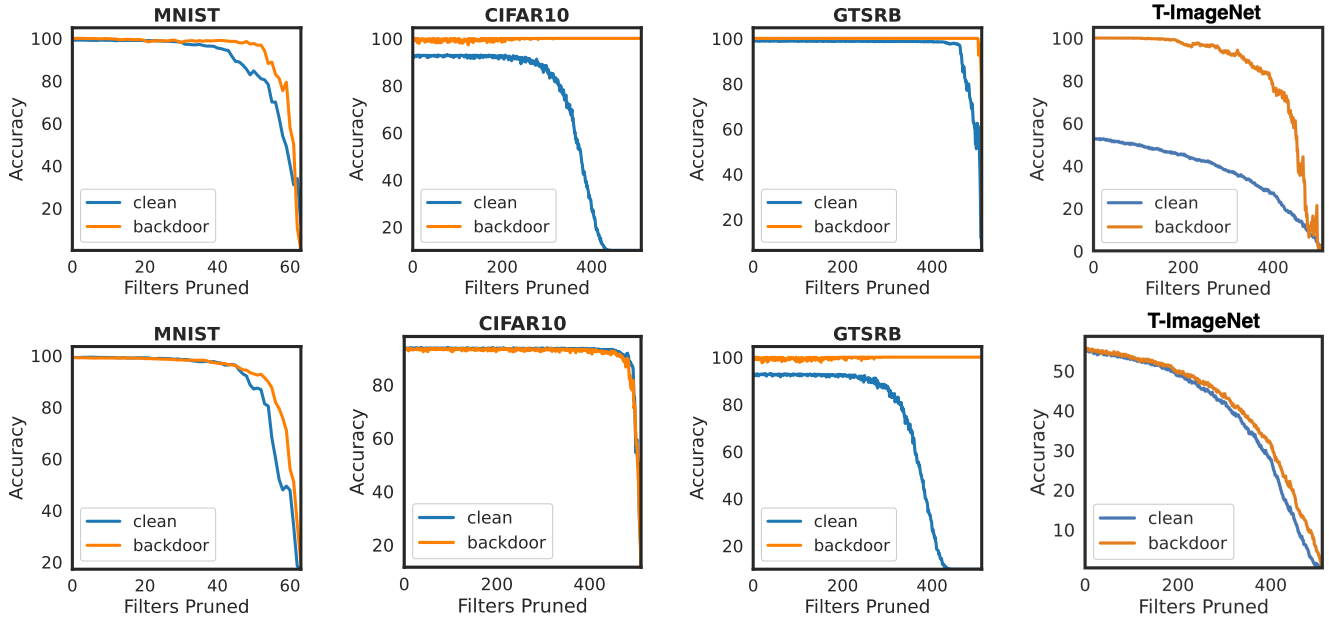


Figure 2: All-to-one (top row) and all-to-all (bottom row) attacks against fine-pruning defense experiments.

3. Additional Results

3.1. Defense Experiments

3.1.1 STRIP

Figure 1 presents the all-to-all results against STRIP [3] defense experiments, which demonstrates that LIRA has a very high degree of stealthiness against STRIP.

3.1.2 Neural Analysis Defense: Fine-pruning

Fine-pruning [7] focuses on neuron analyses. Given a specific layer in the model, it analyzes the neuron responses on a set of clean images and detects the dormant neurons, assuming they are more likely to tie to the backdoor. These neurons are then gradually pruned to mitigate the backdoor. We tested Fine-Pruning on our models and plotting the network accuracy, either clean or attack, with respect to the number of neurons pruned in Figure 2. On all datasets, at no point does the backdoor accuracy drop considerably higher than the clean accuracy, making backdoor mitigation without destroying the classifier impossible.

3.1.3 Neural Cleanse

Figure 3 presents the all-to-all results against Neural Cleanse [13] defense experiments, which shows superior performance than WaNet.

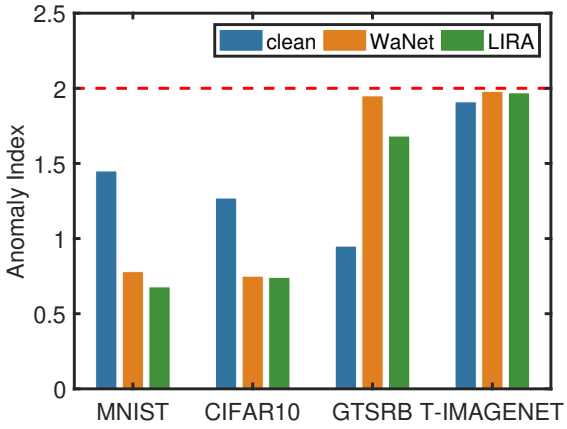


Figure 3: All-to-all attacks against Neural Cleanse defense.

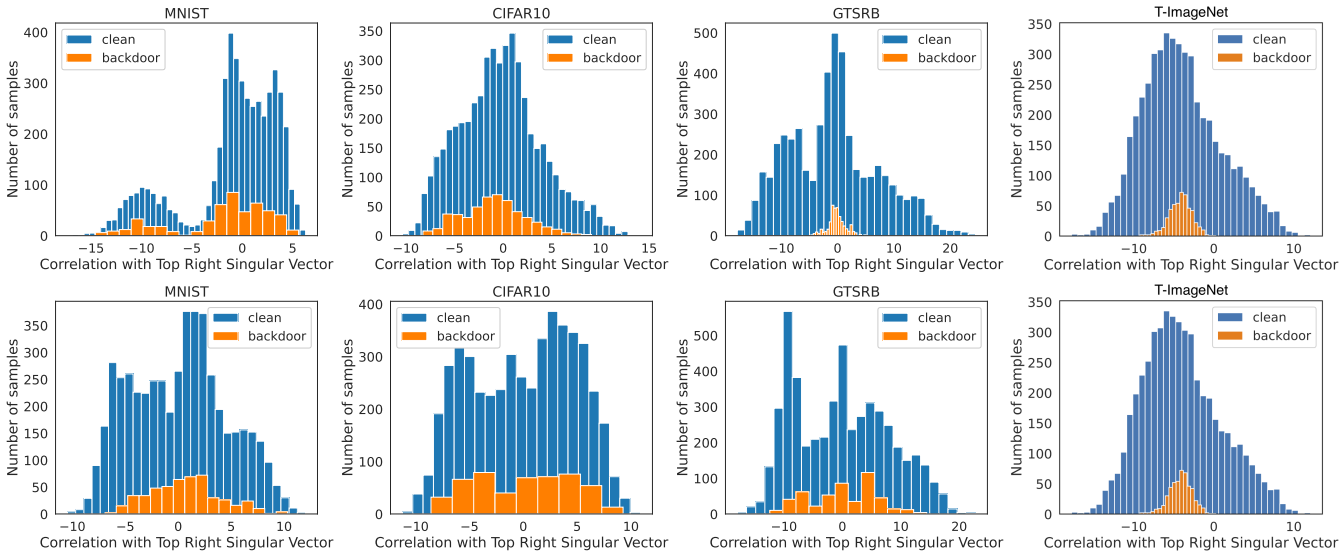


Figure 4: Defense experiments against Spectral Signature. Top row shows all-to-one experiments. Bottom row shows all-to-all experiments. The correlations of the clean and backdoor samples with the top singular vector of the covariance matrix *in the feature space are not separable*.

3.1.4 Latent Space Defense: Spectral Signature

The authors of [11] proposed a defense method based on "spectral signature" of backdoor images. The "spectral signature" is the correlation w.r.t the clean data's top singular value of the covariance matrix of the latent features of the clean data. Similar to WaNet, this defense configuration does not match our threat model. But we find it useful to verify if our backdoor data have the spectral signature. Following the same experiments in [11], we first select 5,000 clean samples and 500 backdoor samples for each dataset. Then we plot the histograms of the correlations between these samples' learned representations and their covariance matrix's top right singular vectors, as shown in Figure 4. It can be seen that the histograms of the two populations are completely inseparable. Consequently, the backdoor training samples could not be removed from the training dataset using this "spectral signature" method.

3.2. Sensitivity Analysis

In this section, we conduct the sensitivity analysis of LIRA. Figure 5 shows the effect of different the mixing parameters α & β (Top) and values of the perturbed noise ϵ (Bottom) on the attack performance of the classifier.

3.2.1 Sensitivity of α and β

We perform experiments on the mixing parameters where $\alpha + \beta = 1$, as shown in Figure 5 (Top). Similar to our discussion in the main paper, LIRA is robust against the variations in these two parameters, which can effectively achieve the near-optimal clean and attack performances in general. We use 0.5 for both α and β in our experiments.

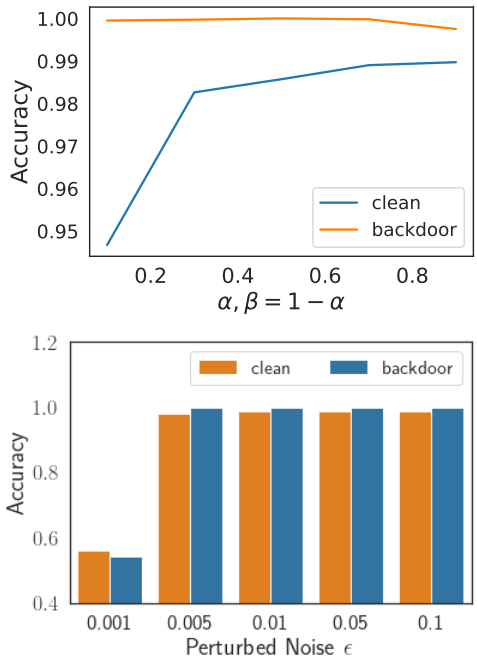


Figure 5: Accuracy under different values of mixing parameter α (we set $\beta = 1 - \alpha$) and the perturbation noise ϵ.

3.2.2 Sensitivity of ϵ

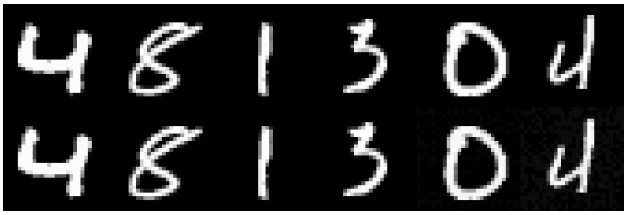
Figure 5 (Bottom) shows the effect of the perturbation parameter ϵ. In general, a larger value of ϵ makes it easier for the algorithm to learn the optimal attack. However, as can be seen, even at a very small noise of 0.002, we still achieve a nearly-optimal backdoor attack.

Additionally, we can see the different backdoor images for different values of ϵ in Figure 6. For low-resolution images, we can observe that a smaller ϵ than 0.01 is adequate, while for larger-resolution inputs, even the perturbed noise of 0.1 still renders visually indifferent backdoor images. This experiment suggests that a value of 0.005 can be used in most cases. In fact, this is the value we use to evaluate LIRA in this paper.

3.3. Visual Inspection Experiments

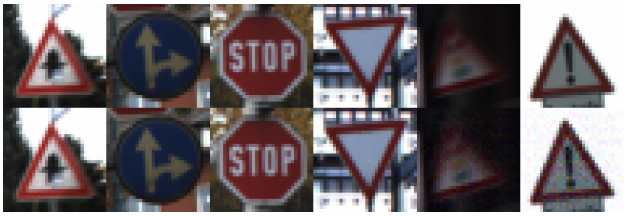
Figure 7 presents additional visual comparisons between different methods. As can be observed, previous perturbation-based attacks (Patched, Blended, SIG, and ReFool) can be completely mitigated under human inspection because of their apparent visual triggers. While the warping-based WaNet is more difficult to be detected than the previous perturbation-based attacks, we still find a considerable amount of "difficult" cases where WaNet's attacks can fail under human inspection, as quantitatively demonstrated in the human inspection tests in Table 1 of the main paper. For example, in Figure 7, the edges of the rhombus or triangle traffic sign are not visually straight, or the circle sign is not round. LIRA's attacks, on the other hand, are extremely difficult to be detected because of its stealthy infinitesimal noise pattern, which is blended perfectly into the contents of the images.

In Figure 8, we provide clean and backdoor images, as well as the corresponding amplified residual, for randomly selected high-resolution images (larger than 196×196) from the GTSRB datasets. LIRA's backdoor images have indistinguishable visual differences from the clean images.



(a) MNIST attack on resolution 32 × 32

(b) CIFAR10 attack on resolution 32 × 32



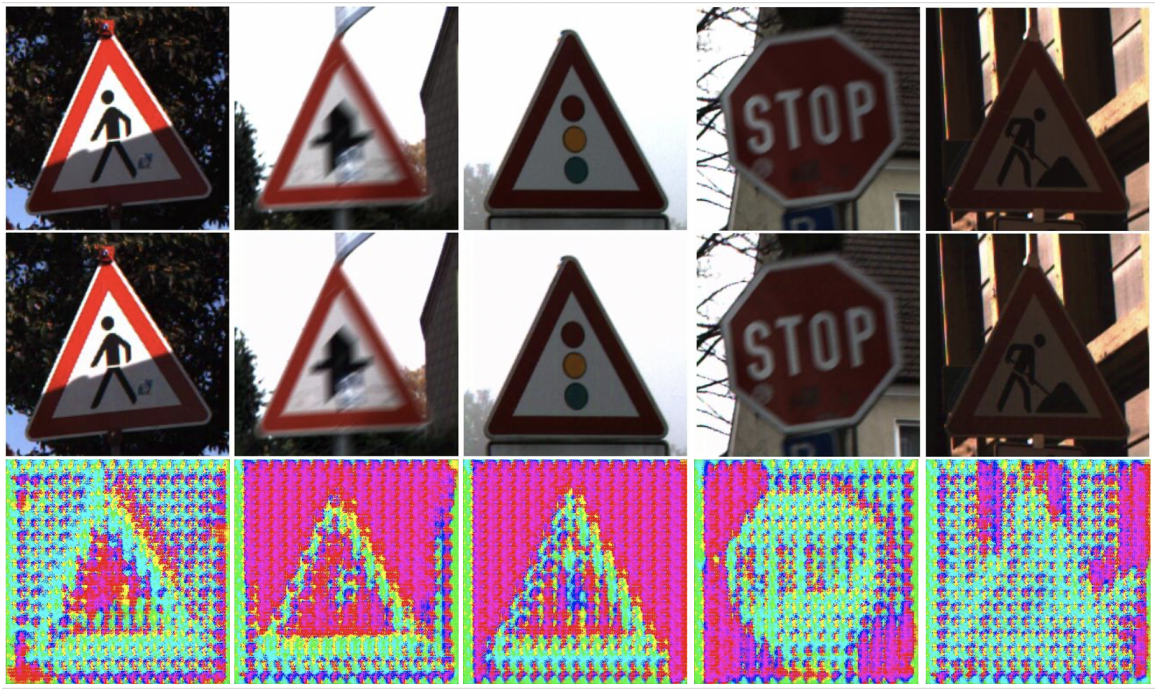
(c) GTSRB attack on resolution 32 × 32

(d) GTSRB attack on resolution 64 × 64

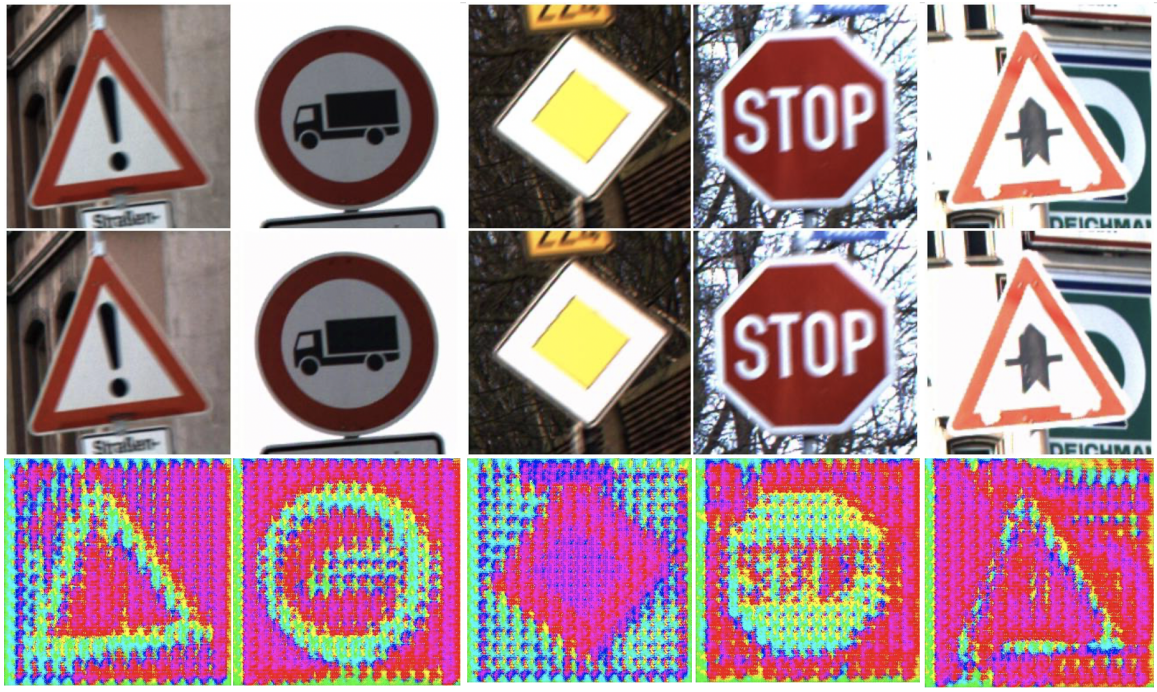
(e) GTSRB attack on resolution 128 × 128

Figure 6: LIRA with different perturbed noises selected from $\epsilon \in \{0, 0.001, 0.005, 0.01, 0.05, 0.1\}$. The top images are clean.

Figure 7: Backdoor images created from different backdoor methods. In WaNet, edges from common shapes such as a circle, rhombus or triangle are deformed (e.g. circle is not entirely round, or edges from rhombus or rectangles are not straight), thus the backdoor can be detected with closer inspection.



(a) Top: Original. Middle: LIRA's backdoor images. Bottom: Amplified residual.



(b) Top: Original. Middle: LIRA's backdoor images. Bottom: Amplified residual.

Figure 8: Randomly selected high-resolution (higher than 196×196) clean images and backdoor images (generated by LIRA) in the GTSRB dataset. The perturbed noise ϵ is 0.01.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, 2009. 2

[2] Ji Feng, Qi-Zhi Cai, and Zhi-Hua Zhou. Learning to confuse: Generating training time adversarial data with auto-encoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11971–11981, Vancouver, Canada, 2019. 1

[3] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. STRIP: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC)*, pages 113–125, San Juan, PR, 2019. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016. 2

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[7] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, pages 273–294, Heraklion, Crete, Greece, 2018. 3

[8] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021. 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 8th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Part III*, pages 234–241, Munich, Germany, 2015. 2

[10] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 1

[11] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8011–8021, Montréal, Canada, 2018. 4

[12] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1

[13] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, San Francisco, CA, 2019. 3

[14] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny Imagenet challenge, 2017. 2