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

Abstract

Recently, machine learning models have demonstrated to be vulnerable to backdoor attacks, primarily due to the lack of transparency in black-box models such as deep neural networks. A third-party model can be poisoned such that it works adequately in normal conditions but behaves maliciously on samples with specific trigger patterns. However, the trigger injection function is manually defined in most existing backdoor attack methods, e.g., placing a small patch of pixels on an image or slightly deforming the image before poisoning the model. This results in a two-stage approach with a sub-optimal attack success rate and a lack of complete stealthiness under human inspection.

In this paper, we propose a novel and stealthy backdoor attack framework, LIRA, which jointly learns the optimal, stealthy trigger injection function and poisons the model. We formulate such an objective as a non-convex, constrained optimization problem. Under this optimization framework, the trigger generator function will learn to manipulate the input with imperceptible noise to preserve the model performance on the clean data and maximize the attack success rate on the poisoned data. Then, we solve this challenging optimization problem with an efficient, two-stage stochastic optimization procedure. Finally, the proposed attack framework achieves 100% success rates in several benchmark datasets, including MNIST, CIFAR10, GTSRB, and T-ImageNet, while simultaneously bypassing existing backdoor defense methods and human inspection.

1. Introduction

Machine learning models, especially deep neural networks (DNNs), have recently achieved state-of-the-art performance in various applications and tasks, ranging from conventional research topics such as computer vision [24, 22] and natural language processing [13, 16] to distant fields such as games [44, 6], computational advertising [57, 54], and structural biology [1, 14]. However, along with the evolution, recent works have shown DNN models are vulnerable to various categories of adversarial attacks [37, 46, 29], which might be attributed to the lack of model transparency and explainability. Among these attacks, evasion attacks such as adversarial examples [7, 34] attempt to fool a trained model by manipulating the inputs in the inference phase, while causative attacks including poisoning [35, 43, 58] and backdoor attacks [32, 30, 10, 20] seek to maliciously alter the model in the training phase.

Recently, the backdoor attack has attracted a lot of attention. The increasing complexity of model building that promoted training outsourcing and machine learning as a service (MLaaS) has also yielded security deficiencies in the supply chain [12, 55]. Existing literature on backdoor attacks [32, 30, 10, 20] has demonstrated that by injecting a backdoor trigger (usually a specific pattern such as a small square) to a small portion of the training data, the trained DNN induces misclassifications while facing inputs with the presence of this trigger. In addition, the model behaves normally on clean inputs, which makes this type of attack hard to detect. As this field of research has evolved, the strength and capabilities of these attacks have increased, leading to methodologies that work with stealthier triggers [51] or compromise extended scenarios [4, 53].

Aligning with the research direction of adversarial examples [23, 33, 52], one property of interest for the backdoor attack is also to improve the fidelity of poisoned images that are used to inject the backdoor and hence reduce the perceptual detectability by human observers. To this end, several works have indeed adopted adversarial example generation in crafting poisoned images that have improved visual quality or indistinguishability from vanilla training data [47, 26]. Blended and other novel trigger patterns have also been investigated [30, 5, 31]. Still, although the attacker carefully crafts the backdoor triggers in these works before poisoning the model, their trigger patterns can be detected by visual inspection. A very recent work, WaNet [36], creates stealthier backdoor images with manually designed warping transformation triggers and achieves state-of-the-art results in both attack success

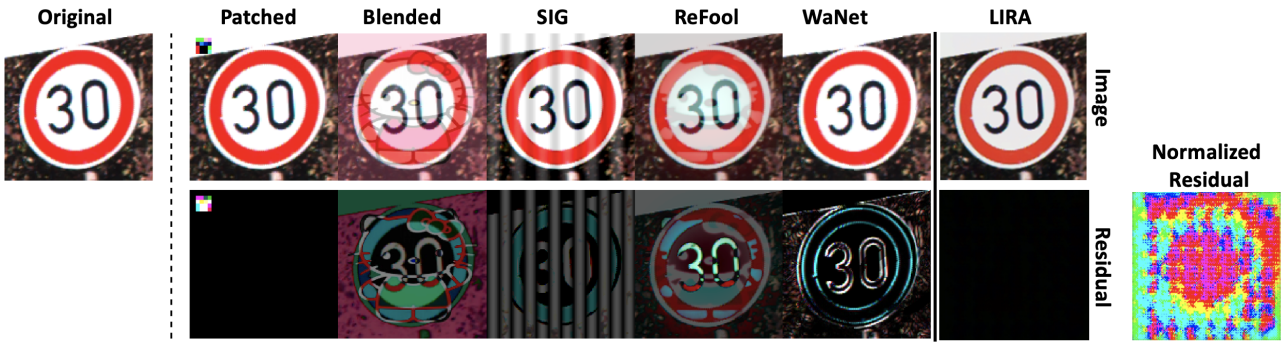

Figure 1: Visualization of backdoor images from different methods. We use the examples from [36]. Images on top from left to right: the original image, images generated by patch based BadNets [21], blended backdoor [10], sinusoidal strips based backdoor (SIG) [5], reflection backdoor (ReFool) [31], warping based backdoor (WaNet) [36], and the proposed LIRA. Bottom images are residual maps that are amplified by $2\times$. It is clear that the images generated by our method is natural and undetectable, as seen in the residual. For further illustration, we present the residual that is amplified by $500\times$ (rightmost).

rate and stealthiness against defense.

In this paper, we propose a novel framework that simultaneously "learns" to generate the perfect yet invisible trigger and poisons the classifier. We first formulate the process of finding the optimal trigger and the optimal classifier in a constrained optimization problem. Then, we propose an effective yet simple alternating stochastic optimization process to solve such a problem. The algorithm allows us to learn to generate the optimal trigger while successfully poisoning the classifier whose performance on the clean data is unchanged (compared to a vanilla classifier trained only on clean data). Paired with a visually imperceptible noise generation trigger function, the trigger patterns generated by our method are extremely difficult to detect, which can successfully pass both the conducted human inspection test (with a significant improvement over the existing backdoor attacks) and several backdoor defense mechanisms. We showcase our backdoor images in Figure 1. We call our method **L**earnable **I**mperceptible and **R**obust Backdoor **A**ttack (**LIRA**).

Our technical **contributions** are summarized below:

- We propose a novel non-convex, constrained optimization problem, which unifies the process of generating the trigger patterns and poisoning the classifier. To solve this problem, we propose an efficient stochastic optimization algorithm that first alternates between finding the optimal trigger function and the optimal poisoned classifier in the highly non-linear parameter space, then fine-tunes only the poisoned classifier.

- We propose a stealthy conditional trigger generation function (also called the transformation function in this work), which can generate remarkably stealthy backdoor images whose residuals with respect to their clean versions are only 1/1000-1/200x of the inputs. As a result, our backdoor attack is visually imperceptible. One example is shown in Figure 1.

- Finally, we achieve state-of-the-art attack performance and stealthiness against both human inspection and existing defense mechanisms.

2. Background

2.1. Backdoor Attacks

Previous works of backdoor injection have understood the attack as the process of introducing malicious modifications to a model, $f(\cdot)$, trained to classify the dataset $\mathcal{S}$. These changes force an association with specific input triggers, denoted as $T(x)$, to the desired model output, namely the target class, y_t [21, 30, 3, 56]. The main methodologies used to inject this functionality into the model are contaminating the training data [10, 30], altering the training algorithm [3], or overwriting/retraining the model parameters [25, 17]. The trigger is either built on a perturbation on the clean image [40] or warping-based [36] to activate the backdoor. Note that patch-based triggers [21, 30] are a special case of the perturbation-based approach, where a small patch is superimposed on the images.

This work follows the direction of perturbation-based methods to enhance the visual quality of poisoned images and stealthiness against human inspection. To this end, the trigger can be defined as

$$T(x) = x + g(x), \tag{1}$$

where g is an imperceptible noise generative model. In essence, the adversarial goal is to force predictions of $T(x)$ to targeted behaviors while minimizing the model's benign loss function with respect to x.

Most patch-based triggers in the literature are perceptually visible such that the corresponding backdoor images can be easily detected under human inspection. Several

techniques have been proposed by prior works to improve the stealthiness of backdoor attacks, including blended [10] and dynamic [41] triggers, which are able to reduce the efficacy of backdoor detection mechanisms. The concept of clean-label backdoor (i.e., poisoned images have labels that are consistent with the model prediction) [49, 40, 5, 31] has also been studied to bypass data sanitization or label inspection. To further improve the stealthiness against human visual inspection, techniques from adversarial example generation have also been adopted in crafting poisoned images [47, 26]. However, while enhancing the stealthiness, these techniques suffer from a degraded success rate for backdoor injection. For example, on MNIST, the success rates of the attack in [26] are less than 75% for digit '3' and less than 70% for digit '2'. A recent work, WaNet, proposes to use a small and smooth warping field in generating backdoor images, making the modification unnoticeable [36], which we consider as the state-of-the-art in this direction of research. While warping-based triggers are more difficult to identify, they can still be visually detected in some cases, given the still relatively large residual as shown Figure 1.

2.2. Backdoor Detection

Several categories of defensive mechanisms have been developed in recent literature to counter backdoor attacks, including detection [8, 48, 18, 45], input mitigation [32, 27], model mitigation [28, 11, 50, 9, 38] approaches.

Detection-based methods aim at detecting malicious training samples crafted to inject backdoor into the models by analyzing the model behavior. For example, activation values in the latent space [8] and predictions of perturbed images [18] have been shown to help detect potential backdoor. Input mitigation methods attempt to remove the trigger of inputs by altering or filtering the image so that the model would still behave normally even the model is injected with a backdoor (i.e., the backdoor would not be activated) [32, 27]. In contrast to the above methods that target a deployed model, the objective of the model mitigation methods is to alleviate the threat from backdoor attacks before deployment. For example, fine-pruning [28] utilizes DNN model pruning to eliminate redundant weights or neurons based on the vanilla training set in the hope of mitigating the potentially injected backdoor, while Neural Cleanse [50] detects whether a trained model has been injected backdoor by searching for possible trigger patches.

3. Threat Model

We consider the same threat model as in prior studies [10, 49, 40, 5] including the state-of-the-art WaNet [36], which assumes the backdoor injection is performed at training and the adversary can have full access to the victim model, including both structures and parameters. Then, the poisoned model will be delivered to a victim customer or deployed by the victim user. A successful backdoor attack over an image classification task should produce malicious behavior on images with the trigger, while otherwise working normally on clean images. However, in typical backdoor attacks, the poisoned images are visually inconsistent with natural images, which can be easily identified by human observers. Besides, it is also desirable to hide the trigger pattern during the backdoor injection, i.e., the poisoned images do not reveal the trigger.

To this end, we propose a stronger backdoor attack where the poisoned images are crafted from clean images with unnoticeable modifications. We advance the state-of-the-art by further enhancing the imperceptibility and robustness of the backdoor attack.

4. Methodology

4.1. Problem Formulation

Consider the standard supervised classification task where one hopes to learn a mapping function $f_\theta : \mathcal{X} \longrightarrow \mathcal{C}$ where $\mathcal{X}$ is the input domain and $\mathcal{C}$ is the set of target classes. The task is to learn the parameters θ from the training dataset $\mathcal{S} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{C}, i = 1, .., N\}$.

Following the standard training scheme of backdoor attacks, the classifier is trained with the combination of the clean and poisoned subsets of S. To create a poisoned sample, a clean training sample (x, y) is transformed into a backdoor sample $(T(x), \eta(y))$, where T is a backdoor injection function (also called the transformation function) and η is the target label function. When training f with the clean and poison samples, we alter the behavior of f so that:

$$f(x) = y, \quad f(T(x)) = \eta(y), \tag{2}$$

for any pair of clean data $x \in \mathcal{X}$ and its corresponding label $y \in \mathcal{C}$. There are two commonly studied backdoor attack settings: all-to-one and all-to-all. In all-to-one attack, the label is changed to a constant target, i.e. $\eta(y) = c$; for all-to-all attack, the true label is one-shifted, i.e. $\eta(y) = (y + 1) \ mod \ |\mathcal{C}|$. In existing works, the transformation function T is selected before training f and fixed during the training process of f.

The main focus of this paper is to simultaneously learn the transformation T, parameterized by ξ, along with the poisoned classifier f, parameterized by θ. T is modeled as a conditional image transformation function; after training, T transforms a clean image x into a backdoor image $T_{\xi^*}(x)$ which are imperceptibly different from x but forces f to make an incorrect classification toward the target class $\eta(y)$. On the other hand, the trained classifier f_{θ^*} has similar performance on the clean data as that of its vanilla version (i.e., the same classifier trained only on the clean data), but its prediction is modified toward the target class whenever it is under attack by the learned transformation function

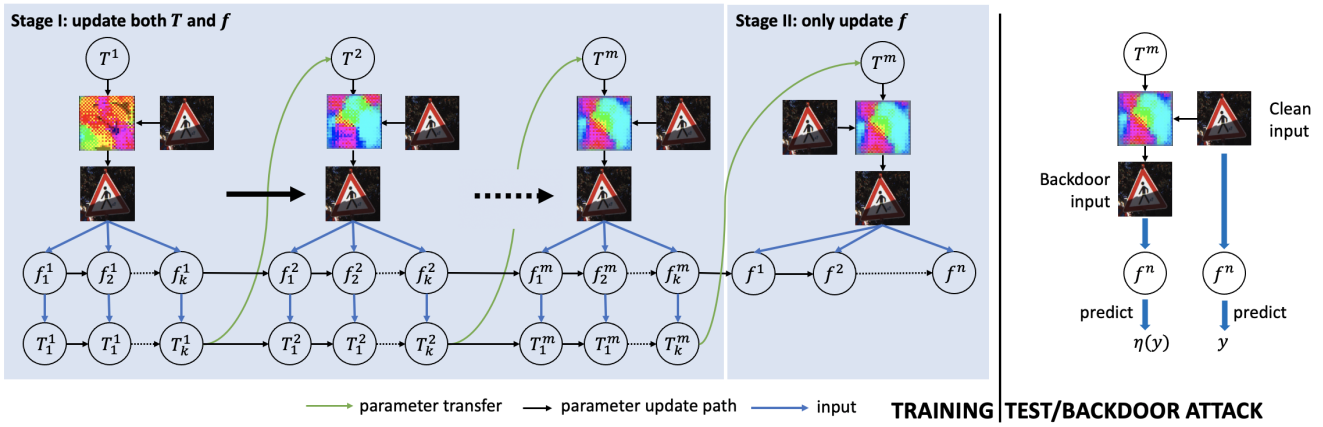

Figure 2: Training: in Stage I, we train both f and T; in Stage II, we only fine-tune f using a learned T in Stage I. Backdoor Attack: correctly classify clean samples while incorrectly classify the backdoor samples toward the specified target class.

$T_{\xi^*}(x)$. The primary advantage of simultaneously learning such a conditional transformation function is that the superimposed trigger patterns are different for different images, thus making the backdoor very difficult to be detected while being optimal for the attack. In the next section, we propose a constrained, non-linear optimization problem that achieves these objectives. An illustration of the proposed backdoor image generation scheme, along with the details of the backdoor attack in LIRA, is described in Figure 2.

4.2. Learning to Backdoor

Consider the empirical risk minimization setting where one hopes to minimize the following loss function on the training data:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i).$$

Our goal is to learn a transformation function $T_\xi : \mathcal{X} \longrightarrow \mathcal{X}$ and a classification model f_θ with the following constraints: 1) the clean image x and its corresponding backdoor image $T(x)$ are imperceptibly different; 2) the classifier simultaneously performs indifferently on x compared to the classifier's vanilla version, but changes its prediction on the backdoor image to the target class $\eta(y)$.

Learning f and T simultaneously has some advantages. First, the backdoor performance of T can be maximized on a specific classifier f. Second, we can model T as a conditional generative function where the generated trigger patterns vary from image to image, thus making the backdoor images difficult to be detected. Finally, the task of selecting the optimal trigger becomes automated, making the backdoor design process more efficient in real-world settings.

We can now formalize the above task. Given the mixing, scalar parameters α and β, a distance function d that measures the visual difference between two images, and a constant scalar threshold value ϵ that is selected to ensure $T_\xi(x)$ is imperceptibly different from x, we solve the following constrained optimization problem:

$$\min_{\theta} \sum_{i=1}^{N} \alpha\mathcal{L}(f_\theta(x_i), y_i) + \beta\mathcal{L}(f_\theta(T_{\xi^*(\theta)}(x_i)), \eta(y_i)) \quad (3)$$

$$s.t. \quad (i) \quad \xi^* = \arg\min_{\xi} \sum_{i=1}^{N} \mathcal{L}(f_\theta(T_\xi(x_i)), \eta(y_i))$$

$$(ii) \quad d(T(x), x) \leq \epsilon$$

In the above problem, a learned classification model with a specific parameter configuration θ is associated with an optimal, stealthy backdoor transformation function, which is trained to fool the model. The goal is to find the paired optimal poisoned classification model $f_{\theta^*(\xi^*)}$ and the optimal transformation function T_{ξ^*} such that $f_{\theta^*(\xi^*)}$ makes a correct prediction on clean data x but an incorrect prediction toward the specified target class $\eta(y)$ on the poison data $T_{\xi^*}(x)$. The parameters α and β control the mixing strengths of the loss signals from the clean and backdoor data when training the classifier. In our experiments, we observe that if α is larger than β, the classifier's performance on the clean data quickly converges to the optimal performance of the vanilla classifier. Conversely, when β is larger than α, the classifier's performance on the backdoor data reaches the optimal value quickly. However, the backdoor classifier still converges to the same optimal performances on both the clean and backdoor samples in both cases. For such reasons, we assume $\alpha = 0.5$ and $\beta = 0.5$ in the remaining part of the paper.

The non-convex, constrained optimization in Equation (3) is challenging because of its non-linear constraint. Fortunately, we can observe that the goal of T is to fool f. Consider the decision boundary of the classifier f at some

specific point (θ,ξ) in the parameter space. In a stochastic optimization algorithm where T and f are neural networks, updating T is equivalent to injecting "more difficult" poisoned data points into the training set to train the classifier. These new backdoor samples cause the decision boundary of the classifier to slightly shift to ensure that the classifier works well on the backdoor data. Consequently, this may worsen its previous performance on the clean data. We can see that this adversarial game between f and T is similar to the training of generative adversarial networks (GANs) [19, 2, 15]. Similar to training GANs, we can update f on both clean and adversarial data while updating T only on adversarial data. However, also similar to GANs, such rapid alternating-update scheme where an update of T causes an update of f can either lead to a longer convergence or cause the training process get stuck in a bad local minima (e.g., when the backdoor attack loss quickly goes to zero, it causes slow or saturated update of the classifier on the clean data because it is hard for the classifier to improve clean-data loss), especially if f or T is stronger than the other.

To stabilize this training process, we first propose to update the current backdoor data that is used to train f only after a certain number of iterations k. Specifically, we update f on clean data and poisoned data generated by the current transformation function while collecting its update trajectories, as shown in Figure 2. The update trajectories are then used to update the T after a fixed number of iterations k. After that, we repeat this trial for m number of times. Under this new training scheme, we find that the classifier can still take a large number of steps to converge to a good performance on both the clean data and backdoor; for example, while on MNIST, the poisoned classifier can reach the optimal clean-data performance of the vanilla classifier after several epochs, on other datasets, it fails to converge to a good clean-data performance as that of the corresponding vanilla classifier. This could be explained by the fact that even training the vanilla classifier (i.e., clean-data training only) already takes a significantly longer number of epochs to reach the optimal performance; e.g., 2 to 3 epochs to reach the optimal performance on MNIST but several hundreds of epochs on the other datasets. Therefore, we propose to use a two-stage training scheme: in Stage I, we train f and T with the proposed alternating scheme for a fixed number of trials; then in Stage II, we fine-tune only the classifier f with both clean and backdoor data generated by the learned transformation T in Stage I. The detailed training procedure is illustrated in Algorithm 1.

4.3. Stealthy Trigger Generator

Inspired by adversarial examples, we model the transformation as a perturbation on the input, as follows:

$$T_\xi(x) = x + g_\xi(x), \quad ||g_\xi(x)||_\infty \leq \epsilon \quad \forall x \tag{4}$$

Algorithm 1 LIRA Backdoor Attack Algorithm

Input:
(1) training samples $S = \{(x_i, y_i), i = 1, ..., N\}$
(2) number of iterations for training the classifier k
(3) number of trials m
(4) number of fine-tuning iterations n
(5) learning rate to train the classifier γ_f
(6) learning rate to train the transformation function γ_T
(7) batch size b
(8) LIRA parameters α and β

Output:
(1) learned parameters of transformation function ξ^*
(2) learned parameters of poisoned classifier θ^*

Initialize θ and ξ.
// Stage I: Update both f and T.
$\hat{\xi} \leftarrow \xi$, $i \leftarrow 0$
repeat
  $j \leftarrow 0$
  repeat
    Sample minibatch (x, y) from S
    $\hat{\theta} \leftarrow \theta_j^i - \gamma_f \nabla_{\theta_j^i}(\alpha \mathcal{L}(f_{\theta_j^i}(x), y) + \beta \mathcal{L}(f_{\theta_j^i}(T_{\hat{\xi}}(x)), \eta(y)))$
    $\hat{\xi} \leftarrow \hat{\xi} - \gamma_T \nabla_{\hat{\xi}} \mathcal{L}(f_{\hat{\theta}}(T_{\hat{\xi}}(x)), \eta(y))$
    $\theta_{j+1}^i \leftarrow \theta_j^i - \gamma_f \nabla_{\theta_j^i}(\alpha \mathcal{L}(f_{\theta_j^i}(x), y) + \beta \mathcal{L}(f_{\theta_j^i}(T_{\xi}(x)), \eta(y)))$
    $j \leftarrow j + 1$
  until $j = k$
  $\xi \leftarrow \hat{\xi}$, $i \leftarrow i + 1$
until $i = m$
// Stage II: Fine-tuning f.
$i \leftarrow 0$, $\theta_0 \leftarrow \theta_k^m$
repeat
  Sample minibatch (x, y) from S
  $\theta_{i+1} \leftarrow \theta_i - \gamma_f \nabla_{\theta_i}(\alpha \mathcal{L}(f_{\theta_i}(x), y) + \beta \mathcal{L}(f_{\theta_i}(T_{\xi}(x)), \eta(y)))$
  $i \leftarrow i + 1$
until $i = n$

The generator function g_ξ takes an input x and generates an artificially imperceptible noise on the same input space, which guarantees the stealthiness of the backdoor attack. We can design such generator function as an autoencoder or the more complex U-Net architecture [39]. However, by training the generator function and the classifier with the proposed training algorithm, we observe that there is not a significant performance difference between a simple autoencoder and U-Net.

Given the proposed generator function, ϵ controls the stealthiness of the trigger-generating function. In practical settings, if ϵ is smaller than 0.01, there is typically no

visible difference between the clean and perturbed images, even on the gray-scale MNIST dataset. This formulation of the transformation function formally makes our attack a perturbation-based backdoor approach. Note that, under this transformation function, the distance d is ℓ_∞-norm on the image-pixel space.

5. Experimental Results

5.1. Experimental Setup

We choose four widely-used datasets for backdoor poisoning attack study: **MNIST**, **CIFAR10**, **GTSRB** and Tiny ImageNet (**T-ImageNet**). For the classifier f, we follow the setting of WaNet [36] and consider a mixed of popular models: Pre-activation Resnet-18 [22] for CIFAR10 and GTSRB datasets and Resnet-18 for T-ImageNet. For the gray-scale MNIST dataset, we also employ the same CNN model that is used by WaNet [36].

For the attack experiments, we compare our methods against the state-of-the-art backdoor attack method, WaNet [36], since its generated backdoor is significantly more stealthy and its attack success rates are significantly better than those of prior perturbation-based attacks.

For the baseline WaNet, we train the networks using the SGD optimizers. The initial learning rate is set to 0.01 with a learning rate decay of 0.1 after every 100 epochs. For other hyperparameters, we use the same values for all the datasets, as suggested in [36]. For LIRA, we use the same optimizer, learning rate, and number of epochs. We train the classifier and the transformation function using the proposed alternating update algorithm for 50 epochs where k is the number of iterations in one epoch (Stage I), i.e., we update the backdoor data which is used to train f after each epoch. Then we continue to fine-tune only the classifiers as described in Stage II for the remaining epochs. Note that this setup results in LIRA's training process with almost the same training time as that of WaNet for each experiment. We choose ϵ as small as 0.005 on all datasets to maintain the stealthiness. Typically, the larger the value of ϵ is, the more successful the backdoor attack; however, we observe that there is no significant performance difference when we perform a grid search on values of ϵ in the range of 0.001 to 0.1. Our implementation was based on the PaddlePaddle deep learning platform.

5.2. Human Inspection Test

To evaluate the stealthiness of the backdoor attacks in real-world settings, we perform a similar human inspection test as proposed in [36]. Specifically, a human is trained with the knowledge of the mechanism and characteristics of the attack and acts as a backdoor defender. We randomly select 25 images from the GSTRB dataset and create their corresponding backdoor images for each backdoor method; this results in a dataset of 50 images. Finally, we have a cohort of 40 human defenders to classify which images are genuine. Effectively, there are 2,000 responses per method.

Images	Patched	Blended	ReFool	WaNet	LIRA
Backdoor	8.7	1.4	2.3	38.6	60.8
Clean	6.1	10.1	13.1	17.4	40.0
Both	7.4	5.7	7.7	28.0	50.4

Table 1: Human Inspection Tests: Success Fooling Rates (%) of Each Method.

We present the percentages of incorrect answers as the success fooling rates in Table 1. As can be observed, LIRA has significantly higher success fooling rates than all other perturbation-based attacks (Patched [20], Blended [10], ReFool [31]) and the currently most stealthy warping-based attack, WaNet, in the backdoor inputs. Furthermore, LIRA's stealthiness causes increasing confusion between the testers when deciding whether an image is genuine for the clean inputs. Essentially, deciding whether an input is a backdoor becomes a random guess, as seen in the averaged fooling rate of 50.4% (the case of "both" backdoor and clean images). This can be explained by the fact that even though the defenders are trained with the knowledge of how LIRA works, LIRA's perturbed noise is so small that there is no visual difference between the clean and backdoor images, as can be seen in Figure 3. Furthermore, LIRA's perturbed noise is conditionally generated, thus varies from image to image. These two properties of LIRA make its backdoor images extremely difficult to be detected. In other backdoor methods, there exist subtle properties that trained defenders can detect. For example, in WaNet, a circle traffic sign is not entirely round.

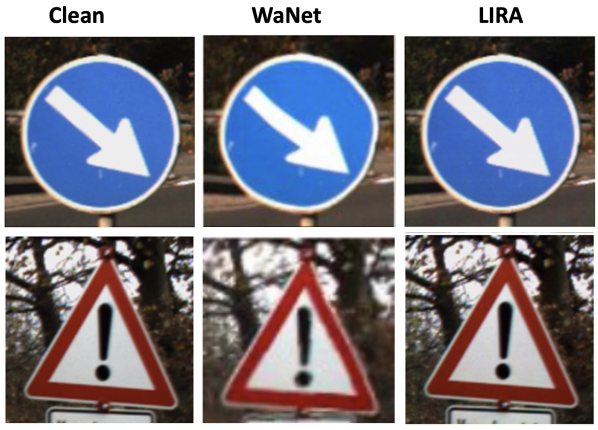

Figure 3: Distinguishable cases.

5.3. Attack Experiments

In this experiment, we first poison the classifier for each compared backdoor attack method and calculate its accura-

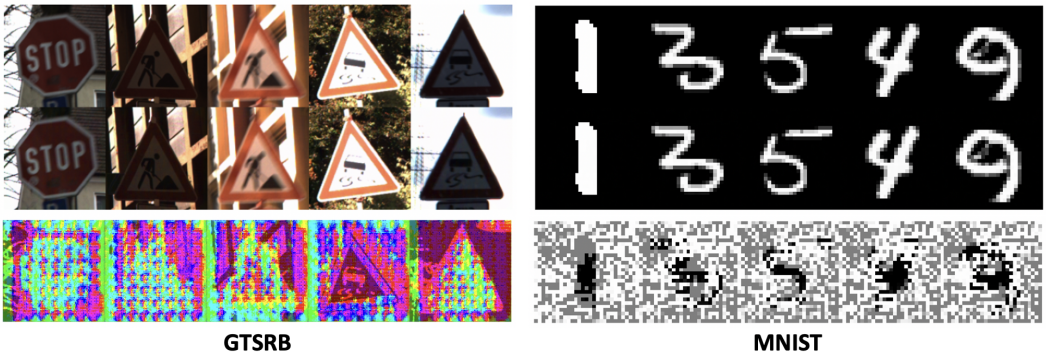

Figure 4: Original sample (first row). Backdoor sample (second row). Amplified residual 500× (third row).

Dataset	WaNet		LIRA	
	Clean	Attack	Clean	Attack
MNIST	0.99	0.99	0.99	1.00
CIFAR10	0.94	0.99	0.94	1.00
GTSRB	0.99	0.98	0.99	1.00
T-ImageNet	0.57	0.99	0.58	1.00

Table 2: Network Performance: All-to-one Attack.

Dataset	WaNet		LIRA	
	Clean	Attack	Clean	Attack
MNIST	0.99	0.95	0.99	0.99
CIFAR10	0.94	0.93	0.94	0.94
GTSRB	0.99	0.98	0.99	1.00
T-ImageNet	0.58	0.58	0.58	0.59

Table 3: Network Performance: All-to-all Attack.

cies on the clean images and poison images (with the defined trigger). We train and evaluate the backdoor models in both all-to-one and all-to-all settings.

The performance comparison between for the currently state-of-the-art baseline method WaNet and LIRA are presented in Table 2 and Table 3 for the all-to-one and all-to-all settings, respectively. As can be observed, on all datasets, both WaNet and LIRA could classify the clean images with similar accuracy as those of the vanilla classifiers trained only on clean data (Clean). In the attack mode, WaNet and LIRA have comparable success rates in all-to-one settings, with LIRA having better performance. Specifically, LIRA achieves 100% success rates in all datasets. In the all-to-all setting, LIRA again achieves better attack success rates than WaNet. Note that all-to-all attacks are more challenging than all-to-one, especially on datasets with a large number of labels such as T-ImageNet.

The result of LIRA is impressive given that the transformed images are visually identical to the clean images, as can be seen in the examples of Figure 4. Furthermore, LIRA is the first perturbation-based method that achieves both high visual stealthiness and attack success rate while preserving the performance of the classifier on clean data.

5.4. Defense Experiments

In this section, we evaluate the backdoor-injected classifiers against popular defense mechanisms, including Neural Cleanse (model mitigation defense) [50], STRIP (detection based defense) [18], and GradCam (network visualization) [42].

5.4.1 Model Mitigation Defense

We first evaluate the robustness of LIRA against Neural Cleanse, a widely-used backdoor model mitigation method based on the pattern optimization approach. Neural Cleanse assumes that the backdoor trigger is patch-based, which hence is suitable for evaluating the proposed method. For each image label, Neural Cleanse identifies if there is a patch pattern that produces a misclassification result to that target label. If any class label yields a significantly smaller pattern, Neural Cleanse considers it as a sign of a potential backdoor. Neural Cleanse quantifies such deviation of the optimal patch of each class label by using the Anomaly Index metric. If the Anomaly Index is less than a threshold of 2 for a class, Neural Cleanse considers there is a backdoor with this class as the target label.

The results of the vanilla classifier (clean), backdoor-injected classifier by WaNet, and backdoor-injected classifier by LIRA are presented in Figure 5. LIRA passes the

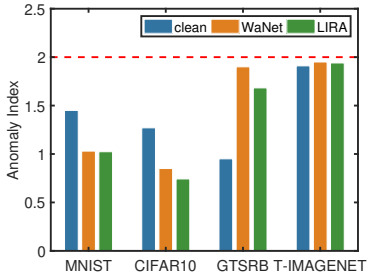

Figure 5: Performance against Neural Cleanse.

Neural Cleanse tests on all datasets. In some cases, specifically MNIST and CIFAR10, LIRA even achieves smaller scores than those of the vanilla models. While WaNet's robustness against Neural Cleanse is unsurprising due to the fact that WaNet is a warping-based attack while Neural Cleanse is a patch-based detection mechanism, the result of LIRA, which is based on a patch-based transformation function, is impressive. Note that previously studied patch-based attacks, such as BadNets, can be defended by Neural Cleanse in most vision datasets.

5.4.2 Detection based Defense

We then evaluate the performance against STRIP [18], a representative detection based backdoor defense mechanism. Given the model and the input image, STRIP perturbs the input image and determines the presence of a backdoor in the model according to the entropy of the predictions of these perturbed images (i.e., if the predictions are consistent or not). The results are shown in Figure 6. With LIRA, since each trigger pattern is conditioned on the image, it is more likely that the perturbation operation of STRIP will break such the trigger pattern (since the perturbed trigger, $g(x_1) + g(x_2)$ may not be the same as $g(x_1 + x_2)$). As expected, LIRA has a similar entropy range as that of the vanilla model (clean). Furthermore, LIRA's entropy ranges exhibit more consistency with the entropy ranges of the clean model than WaNet's entropy range (as reported on the same datasets in [36]).

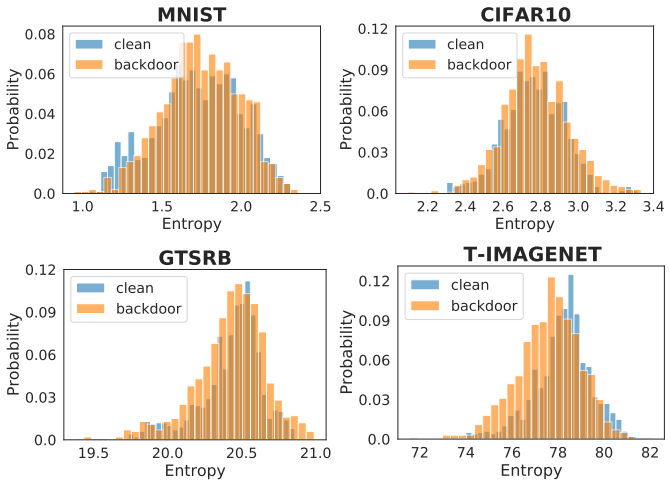

Figure 6: Performance against STRIP.

5.4.3 Network Visualization

It is widely known that visualization tools, such as GradCam [42], are helpful in understanding the neural network's behavior. Thus, we also evaluate the behavior of backdoor-injected models against such tools. Previously, perturbation-based methods, especially patch-based methods such as BadNets [21], can be easily exposed due to the use of obvious trigger patterns, which associate with distinctive latent representations from those of clean images. Furthermore, in the previous study of WaNet [36], while the visualization heat maps are more similar between the clean and backdoor images, we can still see a small difference between them. However, as observed in Figure 7, the visualization heat maps generated on LIRA's attacks are almost the same between the clean and backdoor images. While LIRA is also based on input perturbation, the image is generated from adding an extremely small perturbation to the clean image; hence, the difference in the latent space where the heat maps are generated is also minimal.

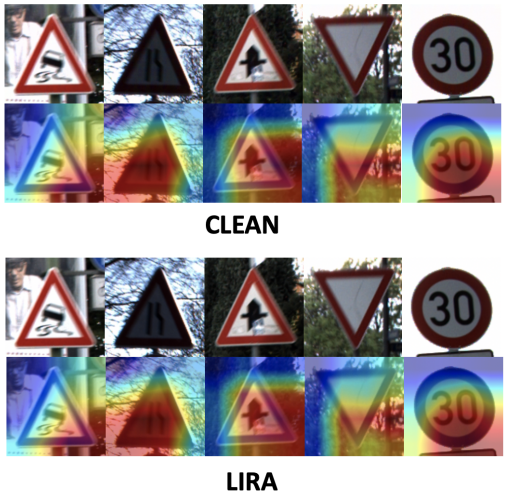

Figure 7: Performance under GradCam heat maps.

6. Conclusion

This work unifies the process of generating the trigger patterns and poisoning the model under a single constrained optimization framework, called LIRA, in order to learn stealthy, dynamic triggers that can successfully poison a classify with unchanged performance on clean data and high attack success rates. We then propose to solve this non-convex constrained optimization problem with an efficient stochastic alternating optimization algorithm. We show that our backdoor attack not only is highly successful with state-of-the-art attack success but also can pass both the human visual inspection test and several machine defense mechanisms. To the best of our knowledge, LIRA is the first work that learns both the trigger function and the poisoned classifier. For such reason, we think that interesting to explore this framework with other types of trigger functions. Finally, in this unified framework, it will be interesting to explore and understand the relationship between the trigger function and the poisoned classifier to advance backdoor defense research.

References

[1] Mohammed AlQuraishi. Alphafold at casp13. *Bioinformatics*, 35(22):4862–4865, 2019. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 5

[3] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. *arXiv preprint arXiv:2005.03823*, 2020. 2

[4] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948, 2020. 1

[5] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNS by training set corruption without label poisoning. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105, Taipei, 2019. 1, 2, 3

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019. 1

[7] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In Bhavani M. Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha, editors, *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS)*, pages 3–14, Dallas, TX, 2017. 1

[8] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Proceedings of the Workshop on Artificial Intelligence Safety*, Honolulu, HI, 2019. 3

[9] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4658–4664, Macao, China, 2019. 3

[10] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1, 2, 3, 6

[11] Hao Cheng, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and Xue Lin. Defending against backdoor attack on deep neural networks. *arXiv preprint arXiv:2002.12162*, 2020. 3

[12] Joseph Clements and Yingjie Lao. Hardware trojan design on neural networks. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019. 1

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN, 2019. 1

[14] Khoa D. Doan, Saurav Manchanda, Suchismit Mahapatra, and Chandan K. Reddy. Interpretable graph similarity computation via differentiable optimal alignment of node embeddings. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 665–674, New York, NY, USA, 2021. Association for Computing Machinery. 1

[15] Khoa D Doan, Saurav Manchanda, Fengjiao Wang, Sathiya Keerthi, Avradeep Bhowmik, and Chandan K Reddy. Image generation via minimizing fr\'echet distance in discriminator feature space. *arXiv preprint arXiv:2003.11774*, 2020. 5

[16] Khoa D. Doan and Chandan K. Reddy. Efficient implicit unsupervised text hashing using adversarial autoencoder. In *Proceedings of The Web Conference 2020*, WWW '20, page 684–694, New York, NY, USA, 2020. Association for Computing Machinery. 1

[17] Jacob Dumford and Walter J. Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, Houston, TX, 2020. 2

[18] 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, 7, 8

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, Montreal, Canada, 2014. 5

[20] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 6

[21] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 2, 8

[22] 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. 1, 6

[23] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: Analysis and improvement. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4441–4449, Salt Lake City, UT, 2018. 1

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, Lake Tahoe, NV, 2012. 1

[25] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. *arXiv preprint arXiv:2004.06660*, 2020. 2

[26] Shaofeng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 2020. 1, 3

[27] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020. 3

[28] 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

[29] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C. M. Leung. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, 6:12103–12117, 2018. 1

[30] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, 2018. 1, 2

[31] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part X*, pages 182–199, Glasgow, UK, 2020. 1, 2, 3, 6

[32] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48, Boston, MA, 2017. 1, 3

[33] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 1652–1659, New Orleans, LA, 2018. 1

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018. 1

[35] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS)*, pages 27–38, Dallas, TX, 2017. 1

[36] 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. 1, 2, 3, 6, 8

[37] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Proceedings of the 2016 IEEE European Symposium on Security and Privacy (IEEE Euro S&P)*, pages 372–387, Saarbrucken, Germany, 2016. 1

[38] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14004–14013, Vancouver, Canada, 2019. 3

[39] 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. 5

[40] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11957–11965, New York, NY, 2020. 2, 3

[41] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675*, 2020. 3

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Procedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, Italy, 2017. 7, 8

[43] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6106–6116, Montréal, Canada, 2018. 1

[44] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354 – 359, 2017. 1

[45] Ezekiel Soremekun, Sakshi Udeshi, Sudipta Chattopadhyay, and Andreas Zeller. Exposing backdoors in robust machine learning models. *arXiv preprint arXiv:2003.00865*, 2020. 3

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014. 1

[47] Te Juin Lester Tan and Reza Shokri. Bypassing backdoor detection algorithms in deep learning. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 175–183, Genoa, Italy, 2020. 1, 3

[48] 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. 3

[49] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 3

[50] 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, 7

[51] Emily Wenger, Josephine Passananti, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks on facial recognition in the physical world. *arXiv preprint arXiv:2006.14580*, 2020. 1

[52] Eric Wong, Frank R. Schmidt, and J. Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6808–6817, Long Beach, CA, 2019. 1

[53] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: distributed backdoor attacks against federated learning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020. 1

[54] Zhiqiang Xu, Dong Li, Weijie Zhao, Xing Shen, Tianbo Huang, Xiaoyun Li, and Ping Li. Agile and accurate CTR prediction model training for massive-scale online advertising systems. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 2404–2409, Virtual Event, China, 2021. 1

[55] Peng Yang, Yingjie Lao, and Ping Li. Robust watermarking for deep neural networks via bi-level optimization. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[56] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2041–2055, London, UK, 2019. 2

[57] Weijie Zhao, Deping Xie, Ronglai Jia, Yulei Qian, Ruiquan Ding, Mingming Sun, and Ping Li. Distributed hierarchical GPU parameter server for massive scale deep learning ads systems. In *Proceedings of Machine Learning and Systems (MLSys)*, Austin, TX, 2020. 1

[58] Chen Zhu, W Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. *arXiv preprint arXiv:1905.05897*, 2019. 1