Backdoor Attack with Imperceptible Input and Latent Modification

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

Abstract

Recent studies have shown that deep neural networks (DNN) are vulnerable to various adversarial attacks. In particular, an adversary can inject a stealthy backdoor into a model such that the compromised model will behave normally without the presence of the trigger. Techniques for generating backdoor images that are visually imperceptible from clean images have also been developed recently, which further enhance the stealthiness of the backdoor attacks from the input space. Along with the development of attacks, defense against backdoor attacks is also evolving. Many existing countermeasures found that backdoor tends to leave tangible footprints in the latent or feature space, which can be utilized to mitigate backdoor attacks.

In this paper, we extend the concept of imperceptible backdoor from the input space to the latent representation, which significantly improves the effectiveness against the existing defense mechanisms, especially those relying on the distinguishability between clean inputs and backdoor inputs in latent space. In the proposed framework, the trigger function will learn to manipulate the input by injecting imperceptible input noise while matching the latent representations of the clean and manipulated inputs via a Wasserstein-based regularization of the corresponding empirical distributions. We formulate such an objective as a non-convex and constrained optimization problem and solve the problem with an efficient stochastic alternating optimization procedure. We name the proposed backdoor attack as Wasserstein Backdoor (WB), which achieves a high attack success rate while being stealthy from both the input and latent spaces, as tested in several benchmark datasets, including MNIST, CIFAR10, GTSRB, and TinyImagenet.

1 Introduction

In the past years, deep neural network (DNN) has successfully transformed many technological fields, such as object classification [26, 20], face recognition [31, 1], autonomous driving [53], security applications [19, 3], etc. Meanwhile, due to the underlying black-box nature, its security and privacy implications have also raised serious concerns recently. Efforts in the research community have exposed the vulnerability of DNN classifiers to various attacks [50, 41, 33]. For instance, adversarial examples leverage the difference between the classifier and human to misclassify specific inputs by adding imperceptible perturbations without altering the model [17]. Such attacks during the inference phase are categorized as evasion attacks [27, 5]. On the other hand, poisoning attacks attempt to inject malicious data points or manipulate the training process to either degrade the model accuracy [37, 45, 60] or cause misclassification for specific inputs (a.k.a. backdoor attacks) [8, 36, 34, 18].

In general, backdoor attacks aim at injecting a malicious behavior into a DNN model so that the model would perform normally on clean inputs but yield misclassification in the presence of the

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

backdoor trigger (e.g., a specific pattern such as a small square [18]). Later on, many works adopt the concepts and techniques in adversarial examples to improve the stealthiness of the trigger against human observers [34, 2, 35]. Recent works have demonstrated more powerful backdoor attacks that are capable of mounting attacks with visual indistinguishable backdoor images [29, 55, 59, 39, 13]. For instance, WaNet [39] generates backdoor images with warping transformation to minimize input difference while LIRA [13] generates backdoor images with imperceptible conditional noise addition, resulting in much stealthier triggers.

To alleviate the threats originated from the ever-growing powerful backdoor attacks, several categories of countermeasures have been developed. One promising direction for backdoor detection entails identifying backdoor images by characterizing the distinguishable dissimilarity in the feature or latent representation between backdoor images and clean images [6, 54, 42, 47, 52]. These methods rely on the assumption that the injected backdoor would leave a noticeable fingerprint in the latent space. For example, activation clustering [6] and spectral signature [54] detect malicious samples by inspecting the clusters of the latent space and the spectrum of the covariance of latent representations, respectively. Thus, a stronger adaptive backdoor attack should also ensure its stealthiness from the latent space.

In this paper, we present a novel methodology for **a backdoor attack that is imperceptible from both the input and latent spaces**. We extend the concept of generating imperceptible backdoor triggers to the latent space by minimizing the Wasserstein distance between the latent representations of the clean and backdoor data, which significantly improves the effectiveness against the existing defense mechanisms, especially those aforementioned that rely on the distinguishability in latent space. We name the proposed method **W**asserstein **B**ackdoor, or **WB**. Our technical contributions are summarized below:

- We propose a non-convex, constrained optimization problem, which learns to poison the classifier with a backdoor whose trigger is visually imperceptible in the input space and whose poisoned samples have indistinguishable latent distribution to the latent distribution of the clean samples. The latent constraint is formulated via a variant of Wasserstein distance, called sliced-Wasserstein distance [24], between the two sets of clean and backdoor data.
- We then develop an efficient estimation of the sliced-Wasserstein distance by exploiting the discriminant directions of the trained classifier, instead of randomly sampling from the unit sphere. The proposed distance is a valid distance metric and requires significantly less computation, while yielding a better estimate than the existing calculations of the sliced-Wasserstein distance.
- Finally, we demonstrate the superior attack performance of the proposed method and its robustness against several representative defense mechanisms. Specifically, we show that the proposed method outperforms the state-of-the-art attacks in terms of latent indistinguishability, while maintaining similar attack success rates and input indistinguishability.

The rest of the paper is organized as follows. We review the background and related work in Section 2. In Section 3, we define the threat model. Section 4 presents the details of the proposed methodology. We evaluate the performance and compare to prior works in Section 5. Finally, Section 6 presents remarks and concludes this paper. We present more details about experimental settings and results as well as supporting proofs in the supplementary material.

2 Background and Related Work

2.1 Backdoor Attack

The increasing popularity of training outsourcing and machine learning as a service (MLaaS) has created potential security risks in the supply chain [10, 58]. One important security threat is backdoor attacks against DNNs, which have recently attracted a lot of attention. Backdoor attacks inject a malicious behavior by leveraging the redundancies inside the model such that the model responds to inputs with triggers maliciously (e.g., classify as a target class that would normally be considered as a wrong class by manual annotation), while preserving the benign behavior for clean inputs without the triggers. Hence, a typical backdoor embedding process is to train the model by minimizing the loss of the clean inputs and the corresponding labels as well as backdoor inputs (with triggers) and

the target class(es). A trigger is typically applied on a clean image by superimposing at a certain location (i.e., patch-based) [18, 34] or adding perturbations [44]. Various forms of the triggers have been investigated in the literature, including blended [8], sinusoidal strips (SIG) [2], reflection (ReFool) [35], and warping-based (WaNet) [39]. As we mentioned above, several techniques have been developed recently that can significantly reduce the visibility of the trigger in the input space to enhance the stealthiness of the backdoor attack [34, 2, 35]. In particular, WaNet uses a smooth warping field to generate backdoor images with unnoticeable modifications [39], while LIRA [13] alternates between the processes of trigger generation and backdoor injection to learn visually stealthy triggers. One prior work, Adversarial Embedding [51], also attempted to improve the latent indistinguishability of the backdoor attack by using adversarial regularization to minimize the distance between the latent distributions of the backdoor inputs and clean inputs.

2.2 Backdoor Defense

By exploring specific characteristics of the injected backdoor, various countermeasures have been proposed [6, 54, 16, 47, 9, 7, 42], although they are often circumvented by subsequent adaptive attacks. For instance, based on the property that a backdoor attack usually targets redundant weights or neurons based on the clean images, model pruning can be used to eliminate the injected backdoor [32]. In contrast, Neural Cleanse assumes a known subset of clean inputs to reverse-engineer possible trigger patches [56]. It is also possible to filter the images to nullify the presence of triggers at the test phase to defend against backdoor attacks [36, 30].

In this paper, we focus on optimizing the characteristics of backdoor attacks in the latent space. As we discussed above, the rationale behind this is that prior works have demonstrated backdoor images cause distinctive activations in the latent space from those of clean inputs. Hence, this distinguishable dissimilarity between clean images and backdoor images can be utilized for defense in both training [6, 54] and test phases [49, 23, 22]. Most of these approaches compute an outlier score to detect abnormal inputs that will be filtered afterward. For example, spectral signature [54] computes the outlier score based on the singular value decomposition of the covariance matrix of the latent representations, while CleaNN [22] leverages a concentration inequality to detect anomalous reconstruction errors that are then suppressed before the input entering the victim DNN.

This work proposes a method to minimize the difference between clean images and backdoor images in the latent space to improve the attack stealthiness. While doing this, we also optimize the visual imperceptibility in the input space, so that our proposed method can bypass visual inspection.

3 Threat Model

We consider the same threat model as in prior studies [44, 51, 39], which assumes the backdoor injection is performed at training and the adversary can access to the victim model including both structures and parameters. 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 identified easily by human observers. Besides, these attacks usually leave a tangible trace in the latent space of the poisoned classifier; thus, some defense methods can easily detect and discard the poisoned models. To this end, we propose a stronger backdoor attack where the poisoned images are crafted with imperceptible perturbation in the input space to clean images as well as unnoticeable trace in the latent space. We advance the state-of-the-art by significantly enhancing the imperceptibility and robustness of the backdoor attack.

4 Proposed Methodology: Wasserstein Backdoor (WB)

4.1 Preliminaries

Consider the standard supervised classification task where one seeks 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 by using 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 trigger 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:

$$f(x) = y, \quad f(T(x)) = \eta(y), \tag{1}$$

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 [18, 39, 51]: all-to-one and all-to-all. In the all-to-one attack, the label is changed to a constant target, i.e. $\eta(y) = c$; while for the all-to-all attack, the true label is one-shifted, i.e. $\eta(y) = (y + 1) \ mod \ |\mathcal{C}|$. In the existing works, the trigger function T is usually selected before training f and fixed during the training process of f.

4.2 Learning to Backdoor

Given the training dataset $\mathcal{S}$ and a loss function $\mathcal{L}$, e.g., cross entropy loss, empirical risk minimization can be used to learn the parameters θ, as follows:

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

The goal of this work is to learn a trigger function $T_\xi : \mathcal{X} \longrightarrow \mathcal{X}$ and a classification model f_θ in such a way that the clean image x and its corresponding backdoor image $T(x)$ are visually consistent in the input space while the backdoor attack does not leave a detectable trace in the latent space of the poisoned classifier. When f is a neural network, $\phi(x)$ can be the output of an intermediate, hidden layer of f, which captures some high-level abstractions of the input. Note that we require the classifier to perform normally on the clean sample, x, compared to the classifier's vanilla version, but change its prediction on the poisoned image, $T(x)$, to the target class $\eta(y)$.

To generate a trigger and poison the image, we follow the prior work [13] and formulate the trigger function as a conditional noise generator g, as follows:

$$T_\xi(x) = x + g_\xi(x), \quad ||g_\xi(x)||_\infty \leq \epsilon \quad \forall x \tag{2}$$

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 [43].

With the above objectives and notations, similar to [13, 4], we can formulate the task into the following constrained optimization problem:

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

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

where $\mathcal{R}_\phi$ is the regularization constraint of the clean and poisoned representations, denoted as $\mathcal{F}_c = \{\phi(x_i) : i = 1, .., N\}$ and $\mathcal{F}_b\{\phi(T(x_i)) : i = 1, .., N\}$, respectively.

In this problem, a learned classification model with a specific parameter configuration θ is associated with an optimal yet stealthy backdoor trigger function, which is trained to poison the model. The classifier is trained to minimize a linear combination of clean and targeted backdoor objectives. The parameters α and β control the mixing strengths of the clean and backdoor loss signals. The trigger function is trained to perturb an image within its ℓ_∞ ball in the input space, so that the loss towards the attack target class is minimized while regularizing the latent representations of the backdoor images.

4.3 Stealthy Latent Representation via Wasserstein Regularization

In practical applications, latent-space defense methods investigate the abnormal trace of incoming data points with respect to the previous stream of data. These traces exist primarily because of the

fact that the clean and backdoor latent representations are separated or distributed differently (e.g., the separated clusters of the clean and poison representations that can be seen in Figures 2 and 3). Thus, we aim to minimize such distributional difference through the regularization constraint $\mathcal{R}_\phi$. Since we cannot assume that the two latent distributions have common support or their density functions are known, commonly-used divergences, such as f-divergences [40, 15] (which include KL and JSD), are difficult to minimize. Instead, we consider the Wasserstein-2 distance and formulate the regularization constraint as follows:

$$\mathcal{R}\phi(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{(x,z)\sim\gamma} p(x,z)||x-z||_2 dx dz \right)^{1/2} \tag{4}$$

where μ and ν are marginal probability measures defined by empirical samples $\mathcal{F}_c$ and $\mathcal{F}_b$ of the latent representations of the clean and poisoned data, respectively.

Estimating the Wasserstein distance also has some challenges. From the primal domain, computing the infimum in Equation (4) is particularly difficult since the data distributions are not fixed or known. On the other hand, employing the Kantorovich-Rubinstein duality requires a separate, parameterized Lipschitz function and a minimax solver, which increases the complexity of the proposed problem. Fortunately, for one-dimensional continuous measures, the Wasserstein distance has an elegant yet closed-form solution. Let q_μ and q_ν be the corresponding density functions of μ and ν, respectively. The Wasserstein-2 distance between one-dimensional measures μ and ν can be given by:

$$\mathcal{W}(\mu,\nu) = \left(\int_0^1 ||(F_\mu^{-1}(z) - F_\nu^{-1}(z)||_2 dz \right)^{1/2} \tag{5}$$

where $F_\mu(z) = \int_{\infty}^{z} q_\mu(\rho)d\rho$ and $F_\nu(z) = \int_{\infty}^{z} q_\nu(\rho)d\rho$ are the cumulative distribution functions. Inspired by the efficiency of this solution and its successful applications in a variety of tasks [12, 24, 14], we propose to first find a family of one-dimensional representations, e.g., through the linear projections, and approximate the Wasserstein distance as a function of these one-dimensional marginals, as follows:

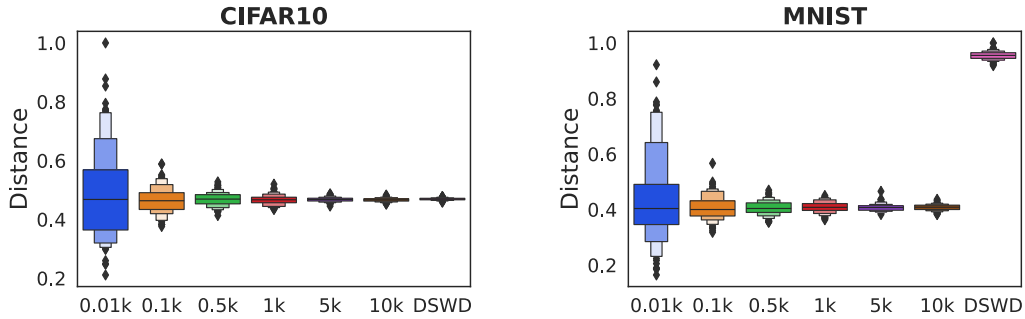
$$\mathcal{R}_\phi(\mathcal{F}_c,\mathcal{F}_b) \approx \left(\frac{1}{L} \sum_{l=1}^{L} [\mathcal{W}(\mathcal{F}_c^{\theta_l}, \mathcal{F}_b^{\theta_l})]^2 \right)^{1/2} \tag{6}$$

where $\mathcal{F}_c^{\theta_l} = \{\theta_l^T \phi(x_i) : i = 1,..,N\}$ and $\mathcal{F}_b^{\theta_l} = \{\theta_l^T \phi(T(x_i)) : i = 1,..,N\}$ contains the projections of the clean and poisoned datasets into a one-dimensional direction defined by θ_l (a slice). Typically, θ_l is drawn from a uniform distribution on the unit sphere. This formulation is also known as the sliced-Wasserstein distance (SWD) [12, 24]. One particular problem with this approach is that the random nature of the slices could lead to several non-informative directions; i.e., the sliced distances are close to 0 in directions that do not lie on the manifolds of the data. Consequently, a large number L of random directions are needed to approximate the sliced-Wasserstein distance, which increases the computational complexity of the estimation.

To remedy this issue, we avoid the uniform sampling of the unit sphere and select directions that contain discriminant information of the two data sources, by exploiting the following fact in the classification task. For backdoor samples of a target class $c_1 \in \mathcal{C}$, created from clean samples of some other class $c_2 \in \mathcal{C}$, the projections into an output dimension represent meaningful discriminant information that distinguishes the backdoor samples (from class c_2) and the clean samples (from class c_1). Thus, we propose to replace the uniform linear projections of SWD with the projections into the output layer. When the latent space is the penultimate layer of the classifier, such projections are equivalent to the following approximation:

$$\mathcal{R}_\phi(\mathcal{F}_c,\mathcal{F}_b) \approx \left(\frac{1}{|\mathcal{C}|} \sum_{c=1}^{|\mathcal{C}|} \left[\mathcal{W}(\mathcal{F}_c^{W_{c,:}}, \mathcal{F}_c^{W_{c,:}})\right]^2 \right)^{1/2}. \tag{7}$$

where $W_{c,:}$ is a row of the matrix $W \in \mathbb{R}^{|\mathcal{C}|\times d}$ (d is the dimension of the latent space), which is the normalized parameter matrix between the penultimate and the output layers.



(a) Pre-activation Resnet-18 Model

(b) CNN Model

Figure 1: Distance estimates (normalized) in the latent space for SWD with different number of sampled directions (between 10 to 10,000) and DSWD.

Empirically, Figure 1 shows the estimated SWD with different numbers of random directions and the proposed calculation, so called DSWD, when the latent space is defined at the penultimate layer of the classifier. The dimension of the latent space is 512 for both MNIST and CIFAR10 datasets. Each distance is computed on a random sample of 1000 clean and 1000 backdoor images, and each calculation is repeated 100 times. It can be seen that with only a fraction of slices, DSWD achieves a significantly smaller variance than that of the SWD estimates. Furthermore, in MNIST, the selected directions of DSWD leads to higher distance estimates than SWD, which means that DSWD selects more discriminant directions than SWD while SWD underestimates the distance between the two empirical samples. In addition, we show that DSWD is a valid distance metric of the latent distributions. The detailed proof is presented in the supplementary material.

Theorem 1. *When the latent space is the penultimate layer of a neural network, the proposed DSWD distance is a valid distance function of probability measures in this space.*

Remark 1. *Since existing defense methods choose the penultimate layer of a neural network. as the space to perform the defense analysis, in most cases, we can employ the proposed DSWD calculation.*

Remark 2. *To preserve the clean classification performance, the classifier seeks optimal parameters that lead to similar predictions of clean samples from the same class. The goal of the trigger function is to make the poisoned samples classified toward a different class. This leads to an adversarial game between the classifier and the trigger functions.*

DSWD also has a significantly better computational efficiency than SWD. In most problems, SWD requires a large number of random directions, typically between 1000 to 10,000, in order to provide a reliable estimate of the distance [38, 14]. In DSWD, the number of random directions is fixed to the number of possible output labels, which is typically small for many classification problems.

4.4 Optimization

The non-convex, constrained optimization in Equation (3) is challenging because of its non-linear constraint. In general, we can alternately update one of f and T while keeping the other fixed, similar to training GANs. However, it is difficult and slow for the classifier to reach an acceptable performance on the clean data, i.e., similar accuracy to that of the vanilla classifier.

Under the alternating update scheme, we observe that on MNIST, the poisoned classifier can reach the acceptable clean-data performance after several epochs; while on other more complex datasets (i.e. CIFAR10, GTSRB, and TinyImagenet), this procedure results in sub-optimal clean-data performance. One possible explanation is that training the vanilla classifier with complex architecture and dataset to reach a decent accuracy is already a difficult and time-consuming task (e.g., 2 to 3 epochs to reach the optimal performance on MNIST but several hundreds of epochs on the other datasets).

Fortunately, we observe that after training the classifier and the trigger functions in an alternating update scheme for a certain number of epochs (denoted as Stage I), we can fix the trigger function and only train the classifier for the remaining epochs (denoted as Stage II). This two-stage training scheme is adopted in our experiments.

5 Experimental Results

5.1 Experimental Setup

We demonstrate the effectiveness of the proposed method through a range of experiments on four widely-used datasets for backdoor attack study: **MNIST**, **CIFAR10**, **GTSRB** and **TinyImagenet**. For these experiments, we follow the previous works [51, 54, 6, 39] and select the penultimate layers of the classifiers as the latent space for the defense experiments. The implementation of WB was based on the PaddlePaddle deep learning platform.

Architectures: For the classifier f, we consider several popular models: Pre-activation Resnet-18 [20], VGG [46], DenseNet [21] for CIFAR10 and GTSRB datasets, and Resnet-18 for TinyImagenet. For the MNIST dataset, we employ a CNN model.

Hyperparameters: For the baselines, we train the classifiers using the SGD optimizer with an initial learning rate of 0.01 and a learning rate decay of 0.1 after every 100 epochs. For other hyperparameters, we follow the proposed setup in [39] for all datasets. We use the same configurations for WB. We train the classifier and trigger functions alternately (Stage I) for 10 and 50 epochs for MNIST and the other datasets, respectively, and fine-tune the classifier (Stage II) for another 40 epochs and 450 epochs for MNIST and the other datasets, respectively. To achieve a high-degree stealthiness of WB, we pick ϵ as small as 0.01 for all datasets. In general, the larger the value of ϵ, the easier the trigger functions can be learned and the more successful the attacks are.

5.2 Attack Performance

We present the attack success rates of the proposed WB method, along with a comparison to two state-of-the-art methods, i.e., WaNet [39] and LIRA [13]. Both LIRA and Wanet's attack performances are significantly better than other approaches, including BadNets [18], and are two of the strongest existing methods that generate very stealthy triggers on the images. We first poison the classifier using the backdoor attack methods in both all-to-one and all-to-all settings and record the performance of the classifier on both clean and backdoor test samples. For all-to-one, we randomly pick the target label $\hat{c}$ (i.e., $\eta(y) = \hat{c}\ \forall y$), while for all-to-all, the target label function is defined as $\eta(y) = (y+1)\ mod\ |\mathcal{C}|\ \forall y$, which is widely used to evaluate the backdoor-related works [39, 18, 6, 13]. Note that this all-to-all attack setting is more challenging than the all-to-one setting, especially on datasets with a large number of classes such as TinyImagenet.

The classification accuracy on the clean test samples and the attack success rate for each method is represented in Table 1 and Table 2 for the all-to-one and all-to-all settings, respectively. As we can observe from these tables, all the methods can achieve high clean-data accuracies and attack success rates. While WB's attack performance slightly drops compared to LIRA's performance, WB is significantly more stealthy in the latent space, as being discussed next.

Table 1: Attack Performance: All-to-one Attack

Dataset	WaNet		LIRA		WB	
	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.99	0.99	0.99	1.00	0.99	0.99
CIFAR10	0.94	0.99	0.94	1.00	0.94	0.99
GTSRB	0.99	0.98	0.99	1.00	0.99	0.99
TinyImagenet	0.57	0.99	0.58	1.00	0.57	0.99

Table 2: Attack Performance: All-to-all Attack

Dataset	WaNet		LIRA		WB	
	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.99	0.95	0.99	0.99	0.99	0.96
CIFAR10	0.94	0.93	0.94	0.94	0.94	0.94
GTSRB	0.99	0.98	0.99	1.00	0.99	0.98
TinyImagenet	0.58	0.58	0.58	0.59	0.58	0.58

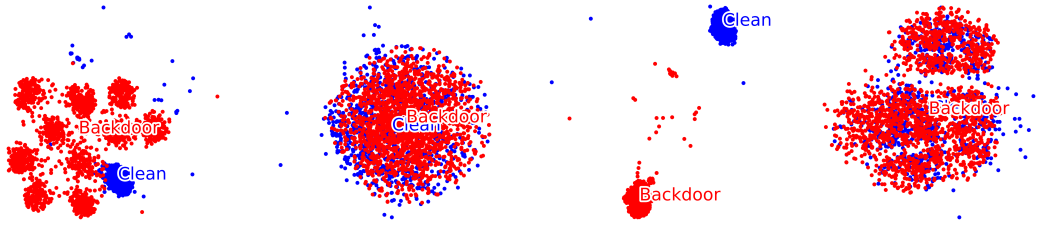
5.3 Latent-Space Defense

Recent works on backdoor defense have found that backdoor attacks tend to leave a tangible trace in the latent space of the poisoned classifier. Activation Clustering [6] and Spectral Signature [54] are two representative defenses used for analyzing the latent space in prior work [51]. In this section, we also examine the latent space of the backdoor-injected classifiers through the lens of these defense methods.

5.3.1 Learned Latent Representation and Activation Clustering

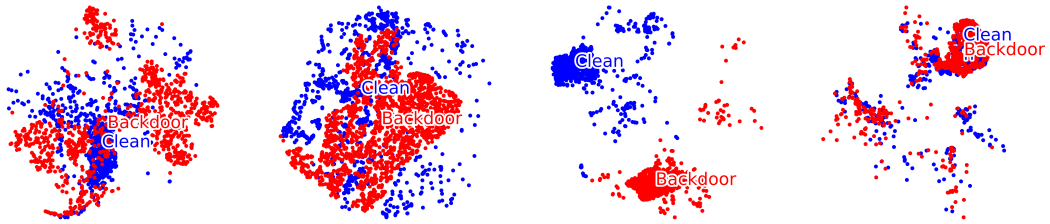
It has been shown in [6] that in a poisoned classifier, the latent representations of the clean and backdoor samples form separate clusters, which can be easily detected using clustering methods such as K-means. The authors also recommend a process called exclusionary reclassification to determine which cluster is poisoned and re-train the poisoned classifier.

In Figure 2 and Figure 3, we can observe highly separated clusters (for samples with the sample predictions of $y = 0$) in the latent space when we omit the latent regularization term $\mathcal{R}_\phi$ in WB (Baseline), which is similar to LIRA [13]. However, when $\mathcal{R}_\phi$ is included, the latent representations of the clean and backdoor samples are distributed similarly. Without well-separated clusters of the clean and poisoned samples, the exclusionary reclassification process in the activation clustering is not effective against the attacks.



(a) All-to-one: Baseline (b) All-to-one: WB (c) All-to-all: Baseline (d) All-to-all: WB

Figure 2: MNIST: t-SNE embedding in the latent space. Baseline is WB without $\mathcal{R}_\phi$.



(a) All-to-one: Baseline (b) All-to-one: WB (c) All-to-all: Baseline (d) All-to-all: WB

Figure 3: CIFAR10: t-SNE embedding in the latent space. Baseline is WB without $\mathcal{R}_\phi$.

Quantitatively, we present the quality scores (i.e., the adjusted Rand Index) of the clustering step in Table 3. The adjusted Rand Index is 1 when the samples form two distinct clusters and is close to 0 for a random separation. We compare WB with BadNets [18] and Adversarial Embedding [51], which is the state-of-the-art backdoor attack method with stealthy latent space. As we can observe in this table, the defense is most successful on BadNets since there exists a perfect clustering of the

Table 3: Adjusted Rand Index in All-to-one Attack

Model	Dataset	Rand Index (BadNets)	Adversarial Embedding		WB	
			Rand Index	Attack	Rand Index	Attack
DenseNet	CIFAR10	0.979	0.1820	0.764	0.0382	0.998
DenseNet	GTSRB	0.997	0.2710	0.914	0.0135	0.997
VGG	CIFAR10	0.998	0.0006	0.962	0.0002	0.999
VGG	GTSRB	0.997	0.6420	0.743	0.1010	0.999

clean and poisoned samples (Rand Index ≥ 0.95). While Adversarial Embedding is more resistant against the defense, WB is significantly more stealthy against the defense since the values of Rand Index are all very close to 0. Note that, similar to BadNets, WaNet also does not pass this defense (please see the supplementary material).

5.3.2 Spectral Signature Defense

The work in [54] proposes a defense method that identifies and removes backdoor samples using the Spectral Signature. For data from each predicted class, Spectral Signature first finds the top singular value of the covariance matrix of the latent vectors of the data. Then it computes the correlation score to this singular value for each sample and those samples with the outlier scores are flagged as backdoor samples. While Spectral Signature is a sample filtering-based defense method, the inspection of the correlation scores can also be useful to verify whether there is a tangible trace in the latent space of the classifier.

Following the same experiments in [54], we first pick 5,000 clean samples and 500 backdoor samples for each dataset. Then, we plot the histograms of the correlation scores for both sets of samples. As we can observe in Figure 4, there is no clear separation between the scores of the backdoor samples and those of the clean samples.

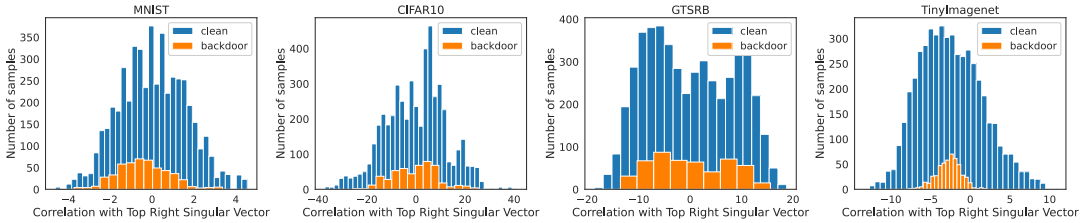
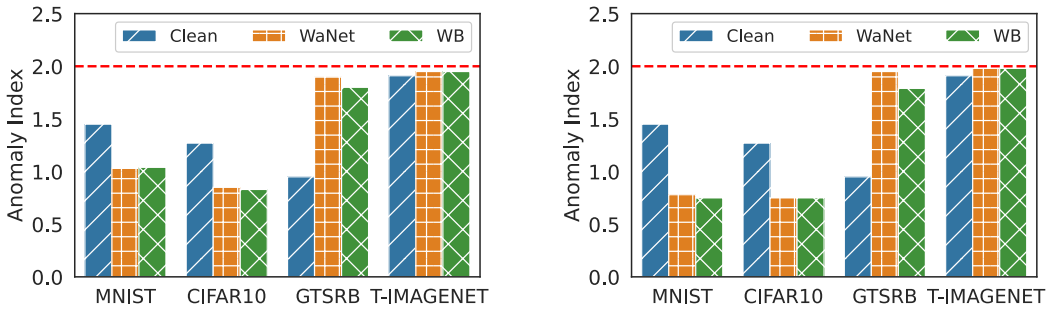


Figure 4: Defense experiments of the all-to-one attack against Spectral Signature. The correlations of the clean and backdoor samples with the top singular vector of the covariance matrix *in the latent space are not separable*.

5.4 Model Mitigation Defense

In this section, we evaluate the robustness of WB against another popular defense, Neural Cleanse [56], which is model mitigation defense based on a pattern optimization approach. Specifically, Neural Cleanse searches for the optimal patch pattern for each possible target label that induces a misclassification to that label. It then quantifies whether any of the optimal backdoor trigger pattern is an outlier via a metric called Anomaly Index. The model has a backdoor if the Anomaly Index is greater than 2 for any class. The anomaly indices are presented in Figure 5.

It can be seen that both WaNet and WB can pass the detection of Neural Cleanse, similar to that of the vanilla classifier (Clean). In MNIST and CIFAR10, WB even achieves smaller Anomaly Indices than



(a) All-to-one attacks

(b) All-to-all attacks

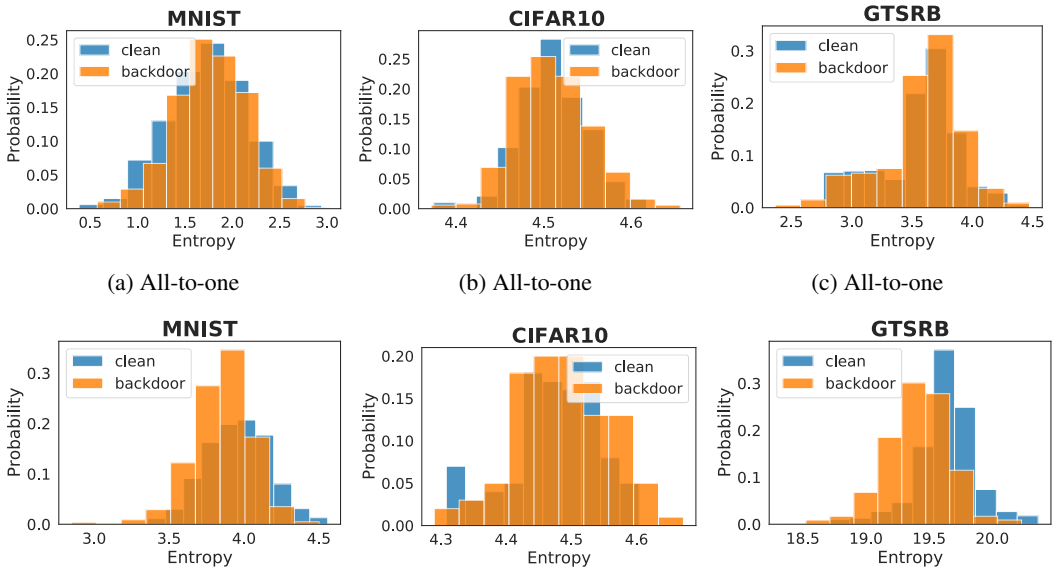
Figure 5: Backdoor attacks against Neural Cleanse defense.

those of the vanilla models. Note that popular backdoor attacks, such as BadNets, can be defended by Neural Cleanse in most of these datasets [56].

5.5 Input Perturbation Defense

In this section, we study the stealthiness of WB against STRIP [16], a representative detection based backdoor defense mechanism. Given the classifier and an input image, STRIP first perturbs the 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).

In Figure 6, we plot the entropy of clean and backdoor images, which are computed by STRIP. We can observe that the distribution of entropy of the backdoor samples is similar to that of the clean samples. In other words, STRIP fails to detect backdoor samples generated by WB, which further validates the advantage of the proposed method.



(a) All-to-one (b) All-to-one (c) All-to-one

(d) All-to-all (e) All-to-all (f) All-to-all

Figure 6: Performance against STRIP defense.

Additional experiments for demonstrating the stealthiness of WB against several other defense approaches can be found in the supplementary material.

6 Conclusion

This paper presented a novel methodology for a backdoor attack that is imperceptible from both the input and latent spaces, i.e., Wasserstein Backdoor (WB). WB learns a trigger function that adds visually imperceptible noise to an input image and minimizes the distributional difference via a novel sliced Wasserstein distance formulation between representations of the clean and backdoor images in the latent space of the trained classifier. We comprehensively evaluated the performance of the proposed method on various image classification benchmark models over a wide range of datasets. Our experimental results demonstrated that the proposed method could significantly improve the effectiveness against the existing defense mechanisms, especially those relying on the distinguishability in latent space.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 59–66, Xi'an, China, 2018.

[2] 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.

[3] Daniel S Berman, Anna L Buczak, Jeffrey S Chavis, and Cherita L Corbett. A survey of deep learning methods for cyber security. *Information*, 10(4):122, 2019.

[4] Avishek Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and William L. Hamilton. Adversarial example games. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3483–3491, Salt Lake City, UT, 2018.

[13] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. LIRA: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11966–11976, 2021.

[14] Khoa D Doan, Amir Kimiyaie, Saurav Manchanda, and Chandan K Reddy. Image hashing by minimizing independent relaxed wasserstein distance. *arXiv preprint arXiv:2003.00134*, 2020.

[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.

[16] 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.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

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

[19] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. LEMNA: explaining deep learning based security applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 364–379, Toronto, Canada, 2018.

[20] 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.

[21] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 1614–1619, Salt Lake City, UT, 2018.

[22] Mojan Javaheripi, Mohammad Samragh, Gregory Fields, Tara Javidi, and Farinaz Koushanfar. Cleann: Accelerated trojan shield for embedded neural networks. In *Proceedings of the IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 11:1–11:9, San Diego, CA, 2020.

[23] Kaidi Jin, Tianwei Zhang, Chao Shen, Yufei Chen, Ming Fan, Chenhao Lin, and Ting Liu. A unified framework for analyzing and detecting malicious examples of dnn models. *arXiv preprint arXiv:2006.14871*, 2020.

[24] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K. Rohde. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 261–272, Vancouver, Canada, 2019.

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

[26] 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.

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[28] 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.

[29] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Trans. Dependable Secur. Comput.*, 18(5):2088–2105, 2021.

[30] 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.

[31] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

[39] 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.

[40] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279, 2016.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[48] 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.

[49] Mahesh Subedar, Nilesh Ahuja, Ranganath Krishnan, Ibrahima J Ndiour, and Omesh Tickoo. Deep probabilistic models to detect data poisoning attacks. *arXiv preprint arXiv:1912.01206*, 2019.

[50] 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.

[51] 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.

[52] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security)*, pages 1541–1558, 2021.

[53] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering (ICSE)*, pages 303–314, Gothenburg, Sweden, 2018.

[54] 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.

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

[56] 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.

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

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

[59] Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and David J. Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the tenth ACM Conference on Data and Application Security and Privacy (CODASPY)*, pages 97–108, New Orleans, LA, 2020.

[60] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7614–7623, Long Beach, CA, 2019.