Backdoor Learning: A Survey

Preprint · July 2020

5 authors, including:

Yiming Li
Tsinghua University
19 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Baoyuan Wu
Tencent AI Lab
59 PUBLICATIONS **997** CITATIONS
SEE PROFILE

Backdoor Learning: A Survey

Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia

***Abstract*—Backdoor attack intends to embed hidden backdoor into deep neural networks (DNNs), such that the attacked model performs well on benign samples, whereas its prediction will be maliciously changed if the hidden backdoor is activated by the attacker-defined trigger. Backdoor attack could happen when the training process is not fully controlled by the user, such as training on third-party datasets or adopting third-party models, which poses a new and realistic threat. Although backdoor learning is an emerging and rapidly growing research area, its systematic review, however, remains blank. In this paper, we present the first comprehensive survey of this realm. We summarize and categorize existing backdoor attacks and defenses based on their characteristics, and provide a unified framework for analyzing poisoning-based backdoor attacks. Besides, we also analyze the relation between backdoor attacks and the relevant fields ($i.e.$, adversarial attack and data poisoning), and summarize the benchmark datasets. Finally, we briefly outline certain future research directions relying upon reviewed works.**

***Index Terms*—Backdoor Learning, Security, Deep Learning, Machine Learning.**

I. Introduction

Over the past decade, deep neural networks (DNNs) have been successfully applied in many mission-critical tasks, such as face recognition, autonomous driving, etc. Accordingly, its security is of great significance and has attracted extensive concerns. One well-studied example is adversarial examples [1], [2], [3], [4], [5], [6], which explores the adversarial vulnerability of DNNs at the inference stage. Compared to the inference stage, the training stage of DNNs involves more steps, including data collection, data pre-processing, model selection and construction, training, model saving, model deployment, etc. More steps mean more chances for the attacker, $i.e.$, more security threats to DNNs. Meanwhile, it is well known that the powerful capability of DNNs significantly depends on the huge amount of training data and computing resource. To reduce the training cost, users may choose to adopt third-party databases, rather than to collect the training data by themselves, since there are many freely available databases in the Internet; users may also train DNNs based on third-party platforms ($e.g.$, cloud computing platforms), rather than to train DNNs locally; users may even directly utilize third-party models. The cost of convenience for users is the loss of the control or the right to know to the training stage, which may further enlarge the security risk for users of DNNs. One typical threat to the training stage is the *backdoor attacks*[1], which is the main focus of this survey.

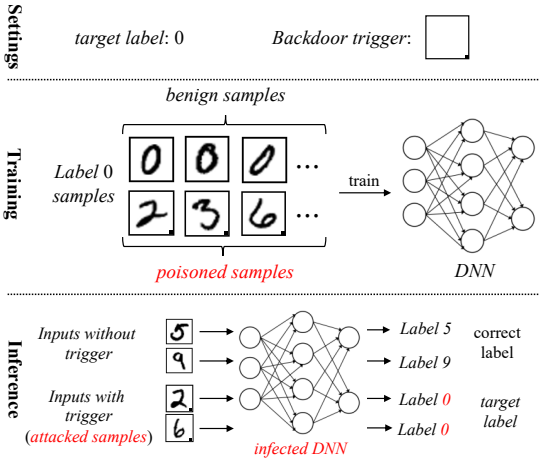

Fig. 1. An illustration of poisoning-based backdoor attacks. In this example, the trigger is a black square on the bottom right corner and the target label is '0'. Part of the benign training image is modified to have the trigger stamped during the training process, and their label is re-assigned as the attacker-specified target label. Accordingly, the trained DNN is infected, which will recognize attacked images ($i.e.$, test images containing backdoor trigger) as the target label while still correctly predict the label for the benign test images.

Gu et al. [7] firstly revealed the threat of backdoor attacks. In general, backdoor attacks aim at embedding the hidden backdoor into DNNs so that the infected model performs well on benign testing samples when the backdoor is not activated, similarly to the model trained under benign settings; however, if the backdoor is activated by the attacker, then its prediction will be changed to the attacker-specified target label. Since the infected DNNs perform normally under benign settings and the backdoor is activated by the attacker-specified trigger, it is difficult for the user to realize the existence of the backdoor. Accordingly, the insidious backdoor attack is a serious threat to DNNs. Specifically, training data poisoning [7], [8], [9] is currently the most straightforward and common way to encode backdoor functionality into the model's weights during the training process. As demonstrated in Fig. 1, some training samples are modified by adding an attacker-specified trigger ($e.g.$, a local patch). These modified samples with attacker-specified target label and benign training samples are

Manuscript received xxx, xxx; revised xxx, xxx.

Corresponding author: Baoyuan Wu and Shu-Tao Xia.

Yiming Li is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China (email: li-ym18@mails.tsinghua.edu.cn).

Baoyuan Wu is with School of Data Science, The Chinese University of Hong Kong, Shenzhen, China and also with Tencent AI Lab, Shenzhen, China (email: wubaoyuan1987@gmail.com).

Yong Jiang is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China and also with PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China.

Zhifeng Li is with Tencent AI Lab, Shenzhen, China.

Shu-Tao Xia is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China and also with PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China (email: xiast@sz.tsinghua.edu.cn).

[1]In this survey, *backdoor attack* refers to the targeted attack towards the training process of DNNs. *Backdoor* is also commonly called the *neural trojan* or *trojan*. We use 'backdoor' instead of other terms in this survey since it is most frequently used.

fed into DNNs for training. Note that the trigger could be invisible [10], [11], [12] and the ground-truth label of poisoned samples could also consistent with the target label [13], [14], [15], which increases the stealthiness of backdoor attacks. Except by directly poisoning the training samples, the hidden backdoor could also be embedded through transfer learning [7], [16], [17], directly modifying model's weights [18], [19], introducing extra malicious module [20], etc., which could happen at all stages of the training process.

Different methods were proposed to defend against backdoor attacks, which can be divided into two main categories, including *empirical backdoor defenses* and *certified backdoor defenses*. Empirical backdoor defenses [21], [22], [23] are proposed based on some observations or understandings of existing attacks and have decent performance in practice; however, their effectiveness have no theoretical guarantee and may probably be bypassed by some adaptive attacks. In contrast, the validity of certified backdoor defenses [24], [25] is theoretically guaranteed under certain assumptions, whereas it is generally weaker than that of empirical defenses in practice. How to better defend backdoor attacks is still an important open question.

As we mentioned, backdoor attacks is a realistic threat and its defense is also of great significance. However, there is still no comprehensive review of both aspects and no framework about how to analyze different works systematically. In this paper, we provide a timely overview of the current status and some insights about future research directions of backdoor learning. We believe this survey will facilitate continuing research in this emerging area. The rest of this paper is organized as follows. Section II briefly describes common technical terms. Section III-IV provides an overview of existing backdoor attacks. Section V demonstrates and categorizes existing defenses. Section VI analyzes the relation between backdoor attacks and related realms, while Section VII illustrates existing benchmark datasets. Section VIII discusses remaining challenges and suggests future directions. The conclusion is provided in Section IX at the end.

II. Definition of Technical Terms

In this section, we briefly describe and explain common technical terms used in the backdoor learning relevant literature. We will follow the same definition of terms in the remaining paper.

- *Benign model* refers to the model trained under benign settings.
- *Infected model* refers to the model with hidden backdoor(s).
- *Poisoned sample* is the modified training sample used in poisoning-based backdoor attacks for embedding backdoor(s) in the model during the training process.
- *Trigger* is the pattern used for generating poisoned samples and activating the hidden backdoor(s).
- *Attacked sample* indicates the malicious testing sample (with trigger) used for querying the infected model.
- *Attack scenario* refers to the scenario that the backdoor attack might happen. Usually, it happens when the training process is inaccessible or out of control by the user, such as training with third-party datasets, training through third-party platforms, or adopting third-party models.
- *Source label* indicates the ground-truth label of a poisoned or an attacked sample.
- *Target label* is the attacker-specified label. The attacker intends to make all attacked samples to be predicted as the target label by the infected model.
- *Attack success rate (ASR)* denotes the proportion of attacked samples which are predicted as the target label by the infected model.
- *Benign accuracy (BA)* indicates the accuracy of benign test samples predicted by the infected model.
- *Attacker's goal* describe what the backdoor attacker intends to do. In general, the attacker wishes to design an infected model that performs well on the benign testing sample while achieving high ASR.
- *Capacity* defines what the attacker/defender can and cannot do to achieve their goal.
- *Attack/Defense approach* illustrates the process of the designed backdoor attack/defense.

III. Poisoning-based Backdoor Attacks

In the past three years, many backdoor attacks were proposed. In this section, we first propose a unified framework to analyze existing poisoning-based attacks towards image classification, based on the understanding of the attack properties. After that, we summarize and categorize existing poisoning-based attacks in detail based on the proposed framework. Attacks for other tasks or paradigms and the well-intentioned applications of backdoor attacks are also discussed at the end.

A. A Unified Framework of Poisoning-based Attacks

We first define three necessary risks in this area, then describe the optimization process of poisoning-based backdoor attacks. Based on the characteristic of the process, poisoning-based attacks can be categorized based on different criteria. Different partitions of poisoning-based methods are summarized in Table I.

We denote the classifier as $f_{\boldsymbol{w}} : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$, where $\boldsymbol{w}$ is the model parameter, $\mathcal{X} \subset \mathbb{R}^d$ being the instance space, and $\mathcal{Y} = \{1, 2, \cdots, K\}$ being the label space. $f(\boldsymbol{x})$ indicates the posterior vector with respect to K classes, and $C(\boldsymbol{x}) = \arg\max f_{\boldsymbol{w}}(\boldsymbol{x})$ denotes the predicted label. Let y_t denotes the target label, $\mathcal{D}_L = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, N_l\}$ indicates the labeled dataset, and $\mathcal{D}'_L = \{\boldsymbol{x} | (\boldsymbol{x}, y) \in \mathcal{D}_L\}$ indicates the instance set of $\mathcal{D}_L$. Three risks involved in existing attacks are defined as follows:

Definition 1 (Standard, Backdoor, and Perceivable Risk)**.**

- *The standard risk R_s measures whether the prediction of $\boldsymbol{x}$ (i.e., $C(\boldsymbol{x})$), is same with its ground-truth label y. Its definition with respect to a labeled dataset $\mathcal{D}_L$ is formulated as*

$$R_s(\mathcal{D}_L) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_L}} \left[\mathbb{I}\{C(\boldsymbol{x}) \neq y\}\right], \tag{1}$$

where $\mathcal{P}_{\mathcal{D}_L}$ indicates the distribution behind $\mathcal{D}_L$. $\mathbb{I}(a)$ denotes the indicator function: $\mathbb{I}(a) = 1$ if a is true, otherwise $\mathbb{I}(a) = 0$.

TABLE I
Summary of existing poisoning-based backdoor attacks.

$\min_{\boldsymbol{t},\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_L-\mathcal{D}_{sL}}}[\mathbb{I}\{C(\boldsymbol{x})\neq y\}] + \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_{sL}}}[\lambda_1\cdot\mathbb{I}\{C(\boldsymbol{x}')\neq y_t\} + \lambda_2\cdot D(\boldsymbol{x}')]$, where $\boldsymbol{t}\in\mathcal{T}$, $\boldsymbol{x}'=G(\boldsymbol{x},\boldsymbol{t})$.				
Visible Attack	$D(\boldsymbol{x}')=1$.	Invisible Attack	Clean-label	$D(\boldsymbol{x}')=0$, and $y_t=y$.
			Poison-label	$D(\boldsymbol{x}')=0$, and $y_t\neq y$.
Attack with Optimized Trigger	$\lvert\mathcal{T}\rvert>1$	Attack with Non-optimized Trigger		$\lvert\mathcal{T}\rvert=1$
Digital Attack	$\boldsymbol{x}'$ is generated in digital space.	Physical Attack		Physical space is involved in generating $\boldsymbol{x}'$.
White-box Attack	$\mathcal{D}_L$ is known	Black-box Attack		$\mathcal{D}_L$ is unknown
Semantic Attack	$\boldsymbol{t}$ is a semantic part of samples.	Non-semantic Attack		$\boldsymbol{t}$ is not a semantic part of samples.

- *The backdoor risk R_b indicates whether the backdoor trigger $\boldsymbol{t}$ can successfully activates the hidden backdoor within the classifier. Its definition with respect to $\mathcal{D}_L$ is formulated as*

$$R_b(\mathcal{D}_L) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_L}}\left[\mathbb{I}\{C(\boldsymbol{x}')\neq y_t\}\right], \tag{2}$$

  where $\boldsymbol{x}'=G(\boldsymbol{x},\boldsymbol{t})$ is the poisoned version of benign sample $\boldsymbol{x}$ under generation function $G(\cdot)$ with trigger $\boldsymbol{t}$. For example, $G(\boldsymbol{x},\boldsymbol{t})=(\boldsymbol{1}-\boldsymbol{\alpha})\cdot\boldsymbol{x}+\boldsymbol{\alpha}\cdot\boldsymbol{t}$ is the most commonly adopted generation function, where $\boldsymbol{\alpha}\in[0,1]^d$ and y_t indicate the blended parameter and target label, respectively.
- *The perceivable risk R_p denotes whether the poisoned sample (i.e., $\boldsymbol{x}'$) can be detected as the malicious sample (by human or machine). Its definition with respect to $\mathcal{D}_L$ is formulated as*

$$R_p(\mathcal{D}_L) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_L}}\left[D(\boldsymbol{x}')\right], \tag{3}$$

  where $D(\cdot)$ is an indicator function: $D(\boldsymbol{x}')=1$ if $\boldsymbol{x}'$ is detected as the malicious sample, otherwise $D(\boldsymbol{x}')=0$.

Based on aforementioned definition, existing attacks can be summarized in a unified framework, as follows:

$$\min_{\boldsymbol{t},\boldsymbol{w}} R_s(\mathcal{D}_L-\mathcal{D}_{sL}) + \lambda_1\cdot R_b(\mathcal{D}_{sL}) + \lambda_2\cdot R_p(\mathcal{D}_{sL}), \tag{4}$$

where $\boldsymbol{t}\in\mathcal{T}$, λ_1 and λ_2 are two non-negative trade-off hyper-parameters, $\mathcal{D}_{sL}$ is a subset of $\mathcal{D}_L$, and $\frac{|\mathcal{D}_{sL}|}{|\mathcal{D}_L|}$ is called *poisoning rate* defined in existing works [7], [26], [9].

Remark. Since the indicator function $\mathbb{I}(\cdot)$ used in R_s and R_b is non-differentiable, it is usually replaced by its surrogate loss ($e.g.$, cross-entropy, KL-divergence) in practice. Besides, as we mentioned, optimization (4) can reduce to existing attacks through different specifications. For example, when $\lambda_1=\frac{|\mathcal{D}_{sL}|}{|\mathcal{D}_L-\mathcal{D}_{sL}|}$, $\lambda_2=0$, and $\boldsymbol{t}$ is non-optimized ($i.e.$, $|\mathcal{T}|=1$), it reduces to the BadNets [7] and the Blended Attack [26]; when $\lambda_2=+\infty$ and $D(\boldsymbol{x}';\boldsymbol{x})=||\boldsymbol{x}'-\boldsymbol{x}||_p$, it reduces to ℓ^p-ball bounded invisible backdoor attacks [11]. Moreover, parameters $\boldsymbol{t}$ and $\boldsymbol{w}$ could be optimized simultaneously or separately through a multi-stage method.

Note that this framework can be easily generalized towards other tasks, such as speech recognition, as well.

B. Attacks for Image and Video Recognition

1) BadNets: Embedding hidden backdoor in a model typically involves the encoding of malicious functionalities within the model's parameters. Gu et al. [7] first defined the backdoor attack and proposed a method, dubbed BadNets, to inject backdoor by tampering the training process through poisoning some training samples. Specifically, as demonstrated in Fig. 1, its training process consists of two main parts, including **(1)** generating the poisoned image $\boldsymbol{x}'$ by stamping the backdoor trigger onto the benign image $\boldsymbol{x}$ to achieve poisoned sample $(\boldsymbol{x}',y_t)$ associated with the attacker-specified target label y_t, and **(2)** training the model with poisoned samples as well as benign samples. Accordingly, the trained DNN will be infected, which performs well on benign testing samples, similarly to the model trained using only benign samples; however, if the same trigger is added onto a testing image, then its prediction will be changed to the target label. The attack scenario of BadNets includes training with third-party datasets and platforms, which reveals serious security threats. BadNets is the representative of *visible attacks*, which opened the era of this field. Almost all follow-up poisoning-based attacks were carried out based on this method.

2) Invisible Backdoor Attacks: Chen et al. [26] first discussed the *invisibility requirement* of poisoning-based backdoor attacks. They suggested that the poisoned image should be indistinguishable compared with its benign version to evade human inspection. To fulfill such a requirement, they proposed a *blended strategy*, which generates poisoned images by blending the backdoor trigger with benign images instead of by stamping as proposed in BadNets [7]. Besides, they demonstrated that even adopting a random noise with a small magnitude as the backdoor trigger can still create the backdoor successfully, which further reduces the risk of being detected.

After that, there was a series of works dedicated to the research of *invisible backdoor attacks*. In [10], Turner et al. proposed to perturb the benign image pixel values by a backdoor trigger amplitude instead of replacing the corresponding pixels with the chosen pattern. Li et al. [11] proposed to regularize the ℓ^p norm of the perturbation when optimizing the backdoor trigger. Zhong et al. [12] adopted the universal adversarial attack [27] to generate the backdoor trigger, which minimizes the ℓ^2 norm of the perturbation. In [28], Bagdasaryan et al. viewed the backdoor attack as a special multi-task optimization, where they fulfilled the invisibility through poisoning the loss computation. Most recently, Liu et al. [8] proposed to adopt a common phenomenon, the reflection, as the trigger for the stealthiness.

Although a poisoned image is similar to its benign version in invisible attacks, however, its source label is usually different from the target label. In other words, all those methods are *poison-label invisible attacks*, where the poisoned samples

seem to be mislabeled. Accordingly, an invisible attack still could be detected by humans by examining the image-label relationship of training samples. To address this problem, a special sub-class of invisible poisoning-based attacks, dubbed *clean-label invisible attacks*, was proposed, which has more serious threats and research value. Turner et al. [10] first explored the clean-label attack, where they leveraged adversarial perturbations or generative models to first modify some benign images from the target class and then conducted the standard invisible attack. The modification approach is to alleviate the effects of 'robust features' contained in the poisoned samples to ensure that the trigger can be successfully learned by the DNNs. Recently, Zhao et al. [14] extended this idea in attacking video classification, where they adopted universal perturbation instead of a given one as the trigger pattern. Another interesting clean-label attack method is to inject the information of a poisoned sample generated by a previous visible attack into the texture of an image from target class by minimizing their distance in the feature space, as suggested in [13]. Besides, Quiring et al. [15] proposed to conceal the trigger as well as hide the overlays of clean-label poisoning through *image-scaling attacks* [29].

3) Attacks with Optimized Trigger: The backdoor trigger is the core of poisoning-based attacks, therefore analyzing how to design a better trigger instead of using a given non-optimized trigger pattern is of great significance and has attracted wide concerns. To the best of our knowledge, Liu et al. [30] first explored this problem, where they proposed to optimize the trigger so that the important neurons can achieve the maximum values. In [11], Li et al. formulated the trigger generation as a bilevel optimization, where the trigger was optimized to amplify a set of neuron activations with ℓ^p regularization for invisibility. Bagdasaryan et al. [28] treated backdoor attacks as a multi-object optimization, and proposed to optimize trigger and train DNNs simultaneously. Recently, with the hypothesis that if a perturbation can induce most samples toward the decision boundary of the target class then it will serve as an effective trigger, [12], [14], [31] proposed to generate trigger through universal adversarial perturbation.

4) Physical Backdoor Attacks: Different from previous *digital attacks* that adopted the setting that the attack is conducted completely in the digital space, Chen et al. [26] first explored the landscape of *physical attacks*. In [26], they adopted a glasses as the physical trigger to mislead the infected face recognition system developed in a camera. Further exploration of attacking face recognition in the physical world was also discussed by Wenger et al. [32]. A similar idea was discussed in [7], where a post-it note was adopted as the trigger in attacking traffic sign recognition. Recently, Li et al. [9] demonstrated that existing digital attacks fail in the physical world since the involved transformations ($e.g.$, rotation, and shrinkage) change the location and appearance of trigger in attacked samples compared with the one used for training. This inconsistency will greatly reduce the performance of the attack. Based on this understanding, they proposed a transformation-based attack enhancement so that the enhanced attacks remain effective in the physical world.

5) Black-box Backdoor Attacks: Different from previous *white-box attacks*, which require the knowledge of training samples, *black-box attacks* adopt the settings that the training set is inaccessible. In practice, the training dataset is usually not shared due to privacy or copyright concerns, therefore black-box attacks are more realistic than white-box ones. Specifically, black-box backdoor attacks require to generate some training samples based on the given model at first. For example, in [30], they generated some representative images of each class by optimizing initialized images from another dataset such that the prediction confidence of the selected class reaches maximum. With the reversed training set, white-box attacks can be adopted for injecting hidden backdoor.

6) Semantic Backdoor Attacks: The majority of backdoor attacks, $i.e.$, the *non-semantic attacks*, assume that the trigger is independent of benign images. In other words, attackers need to modify the image in the inference stage to activate the hidden backdoor. Is it possible that a semantic part of samples can also serve as the trigger, such that the attacker is not required to modify the input at inference time to deceive the infected model? Bagdasaryan et al. first explored this problem and proposed a novel type of backdoor attacks [33], [28], $i.e.$, the *semantic backdoor attacks*. Specifically, they demonstrated that assigning an attacker-chosen label to all images with certain features, $e.g.$, green cars or cars with racing stripes, for training can create a semantic hidden backdoor in infected DNNs. Accordingly, the infected model will automatically misclassify testing images containing pre-defined semantic information without the requirement of image modification.

C. Attacks for Other Tasks or Paradigms

In this section, we summarize the poisoning-based attack against other tasks or paradigms.

In the area of natural language processing, Dai et al. [34] first discussed the backdoor attack against LSTM-based sentiment analysis. Specifically, they proposed a BadNets-like approach, where an emotionally neutral sentence was used as the trigger and was randomly inserted into some benign training samples. In [35], Chen et al. further explored this problem, where three different types of triggers ($i.e.$, char-level, word-level, and sentence-level triggers) were proposed and reached decent performance. Most recently, Kurita et al. [16] demonstrated that sentiment classification, toxicity detection, and spam detection can also be backdoored even after fine-tuning. Some researches also revealed the backdoor threat towards graph neural networks (GNN) [36], [37]. In general, an attacker-specified subgraph was defined as the trigger so that the infected GNN will predict the target label for an attacked graph once the subgraph trigger is contained. Besides, the backdoor threat towards reinforcement learning [38], [39], wireless signal classification [40], and continual learning [41], were also studied.

The security issues of collaborative learning, especially federated learning, have attracted extensive attention. In [33], Bagdasaryan et al. introduced a backdoor attack to federated learning based on amplifying the poisoned gradient on the node servers. Besides, Bhagoji et al. [42] discussed the stealthy

model poisoning attack, and Xie et al. [43] introduced a distributed backdoor attacks to federated learning. Most recently, [44] also discussed how to backdoor federated learning. Besides, the backdoor attacks towards meta federated learning [45] and feature-partitioned collaborative learning [46] were also discussed. Moreover, some works [47], [48], [49], [50], [51], [52] also questioned whether federal learning is really easy to be attacked. Except for collaborative learning, the backdoor threat of another important learning paradigm, $e.g.$, the transfer learning, was also discussed in [7], [16], [17], [53].

D. Backdoor Attack for Good

Despite malicious purposes, how to use the backdoor attack in the right way has also obtained preliminary explorations. Adi et al. [82] exploited backdoor attacks in verifying model ownership. They proposed to watermark the DNNs through backdoor embedding. Accordingly, the hidden backdoor in the model can be used to examine the ownership, while the watermarking process still preserves original model functionality. Besides, Sommer et al. [83] revealed how to verify whether the server truly erases their data when users require data deletion through poisoning-based backdoor attacks. Specifically, in their verification framework, each user poisons part of its data with user-specific trigger and target label. Accordingly, each user can leave a unique trace in the server for deletion verification after the server being trained on user data while having a negligible impact on the benign model functionality. Shan et al. [84] introduced a trapdoor-enabled adversarial defense, where the hidden backdoor is injected by the defender to prevent attackers from discovering the natural weakness in a model. The motivation is that the adversarial perturbation generated by gradient-descent-based attacks towards an infected model will converge near the trapdoor pattern, which is easily detected by the defender. Most recently, Li et al. [85] discussed how to protect open-sourced datasets based on backdoor attacks. Specifically, they formulated this problem as determining whether the dataset has been adopted to train a third-party model. They proposed a hypothesis test based method for the verification, based on the posterior probability generated by the suspicious third-party model of the benign samples and their correspondingly attacked samples.

IV. Non-poisoning-based Backdoor Attacks

Except for poisoning-based attacks, some non-poisoning-based attacks were also proposed. These methods inject backdoor not directly through optimizing model parameters during the training process with poisoned samples. Their existence demonstrates that except for happening at the data collection, the backdoor attack could also happen at other stages ($e.g.$, deployment stage) of the training process, which further reveals the severity of the backdoor attack.

A. Targeted Weight Perturbation

In [18], Dumford et al. first explored the non-poisoning-based attack, where they proposed to modify the model's parameters directly instead of through training with poisoned samples. The primary task in this work is face recognition, where they assumed that the training samples can not be modified by attackers. The attacker's goal is to make their own face to be granted access despite not being a valid user while ensuring that the network still behaves normally for all other inputs. To fulfill this target, they adopted a greedy search across models with different perturbations applied to a pre-trained model's weights.

B. Targeted Bit Trojan

Instead of modifying the model's parameters directly through a search-based approach, Rakin et al. [19] demonstrated a new method, dubbed targeted bit trojan (TBT), discussing how to inject a hidden backdoor without the training process more effectively. TBT contains two main processes, including gradient-based vulnerable bits determination (similar to the process proposed in [30]), and targeted bits flipping in main memory by adopting *row-hammer attack* [86]. The proposed method achieved remarkable performance, the authors were able to mislead ResNet-18 on the CIFAR-10 dataset with 84 bit-flips out of 88 million weight bits.

C. TrojanNet

Different from previous approaches where the backdoor is embedded in the parameters directly, Guo et al. [87] proposed TrojanNet to encode the backdoor in the infected DNNs activated through a secret weight permutation. They assumed that the infected network is used with a hidden backdoor software which could permute the parameters when the backdoor trigger is presented. Training a TrojanNet is similar to the *multi-task learning*, although the benign task and malicious task share no common features. Besides, the authors also proved that the decision problem to determine whether the model contains a permutation that triggers the hidden backdoor is NP-complete, and therefore the backdoor detection is almost impossible.

D. Attack with Trojan Module

Most recently, Tang et al. [20] proposed a novel non-poisoning-based backdoor attack, which inserts a trained malicious backdoor module ($i.e.$, a sub-DNN) into the target model instead of changing parameters in the original model to embed backdoor. The proposed method is model-agnostic and could be injected into most DNNs, $i.e.$, retraining on poisoned samples is not required. This method significantly reduced the computational cost compared to previous poisoning based attack methods.

V. Backdoor Defenses

To defend backdoor attacks, several backdoor defensive methods were proposed. Existing methods mostly aim at defending poisoning-based attacks and can be divided into two main categories, including *empirical backdoor defenses* and *certified backdoor defenses*. Empirical backdoor defenses are proposed based on some understandings of existing attacks and have decent performance in practice, whereas their effectiveness has no theoretical guarantee. In contrast, the validity of

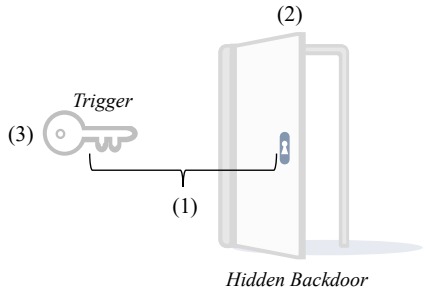

Fig. 2. An illustration of backdoor attacks and three corresponding defense paradigms. Intuitively, the poisoning-based backdoor attack is similar to unlock a door with the corresponding key. Accordingly, three main paradigms, including **(1)** trigger-backdoor mismatch, **(2)** backdoor elimination, and **(3)** trigger elimination, can be adopted to defend the attack. Different types of approaches were proposed towards the aforementioned paradigms, as illustrated in Table II.

TABLE II
Summary of existing empirical backdoor defenses in image recognition. (Some literature proposed different types of defenses simultaneously, therefore they will appear multiple times in this table.)

Defense Paradigm	Defense Sub-category	Literarure
Trigger-backdoor Mismatch	Preprocessing-based Defense	[54], [55], [56], [57] [9]
Backdoor Elimination	Model Reconstruction based Defense	[54], [21], [58]
	Trigger Synthesis based Defense	[59], [60], [61], [62] [63], [64], [65], [66] [67]
	Model Diagnosis based Defense	[68], [69], [23], [70] [71]
	Poison Suppression based Defense	[72], [73]
	Training Sample Filteing based Defense	[74], [75], [76], [77] [78], [79]
Trigger Elimination	Testing Sample Filtering based Defense	[22], [80], [72], [81]

certified backdoor defenses is theoretically guaranteed under certain assumptions, whereas it is generally weaker than that of empirical defenses in practice. At present, certified defenses are all based on the *random smoothing* [88], while empirical ones have multiple types of approaches.

A. Empirical Backdoor Defenses

Intuitively, the poisoning-based backdoor attack is similar to unlock a door with the corresponding key. In other words, there are three indispensable requirements to ensure a successful backdoor attack, including **(1)** having a hidden backdoor in the model, **(2)** containing a trigger in the sample, and **(3)** the trigger and the backdoor are matched, as shown in Fig. 2. Accordingly, three main defense paradigms, including **(1)** trigger-backdoor mismatch, **(2)** backdoor elimination, and **(3)** trigger elimination, can be adopted to defend existing attacks. Different types of approaches were proposed towards the aforementioned paradigms, which are summarized in Table II and will be further demonstrated as follows:

1) Preprocessing-based Defenses: Preprocessing-based defenses introduce a preprocessing module before the original inference process, which changes the pattern of the triggers in the attacked samples. Accordingly, the modified trigger no longer matches the hidden backdoor therefore preventing backdoor activation.

Liu et al. [54] were the first to exploit preprocessing as the defense approach towards image classification tasks, where they adopted a pre-trained auto-encoder as the preprocessor. Inspired by the idea that the trigger region contributes most to the prediction, a two-stage image preprocessing approach, dubbed Februus, was proposed by Doan et al in [55]. At the first stage, Februus utilizes GradCAM [89] to identify influential regions, which will then be removed and replaced by a neutralized-color box. After that, Februus adopts a GAN-based inpainting method to reconstruct the masked region for alleviating the adverse effect towards benign samples. Udeshi el al. [56] proposed to utilize the dominant color in the image to make a square-like trigger blocker in the preprocessing stage, which was adopted to locate and remove the backdoor trigger. This approach was motivated by the fact that placing a trigger blocker at the position of the backdoor trigger in the attacked image will result in a change in the prediction of the model. Vasquez et al. [57] proposed to preprocess the image through style transfer. Recently, Li et al. discussed the property of existing poisoning-based attacks with static trigger [9]. They demonstrated that if the *appearance* or *location* of the trigger is slightly changed, then the attack performance may degrade sharply. Based on this observation, they proposed to adopt spatial transformations ($e.g.$, shrinking, flipping) as the preprocessor. Compared with previous methods, this method is more efficient since it requires almost no additional computational costs.

2) Model Reconstruction based Defenses: Different from preprocessing based defenses, model reconstruction based defenses aim at removing the hidden backdoor in the infected model. Accordingly, even if the trigger is still contained in the attacked samples, the prediction remains unmalicious since the backdoor was already removed.

Liu et al. [54] proposed to retrain the given model with local benign samples starting from the weight of the given model. The effectiveness of this method may probably due to the *catastrophic forgetting* in DNNs [90], $i.e.$, the hidden backdoor is gradually removed as the training goes since the re-training set contains no poisoned samples. Motivated by the observation that the backdoor related neurons are usually dormant for benign samples, Liu et al. [21] proposed to prune those neurons to remove the hidden backdoor. Besides, they proposed a fine-pruning method, which first prunes the DNNs and then fine-tunes the pruned network to combine the benefits of the pruning and fine-tuning defenses. In [58], Zhao et al. showed that the hidden backdoor of the infected DNNs can be repaired based on the *mode connectivity* technique [91] with a certain amount of benign samples.

3) Trigger Synthesis based Defenses: Except for eliminating the hidden backdoor directly, trigger synthesis based defenses propose to synthesizes the backdoor trigger at first,

following by the second stage that the hidden backdoor is eliminated by suppressing the effect of the trigger.

This type of defense enjoys certain similarities with model reconstruction based ones in the second stage. For example, pruning and retraining are the common techniques used in removing the hidden backdoor in both types of defenses. However, compared with the model reconstruction based defenses, the trigger information obtained in synthesis based defenses makes the removal process more effective and efficient.

Wang et al. [59] first proposed to remove the hidden backdoor based on the synthetic trigger in a 'black-box' scenario, where the training set is inaccessible. Specifically, the proposed method, $i.e.$ Neural Cleanse, first obtained potential trigger patterns towards every class, and then determined the final synthetic trigger and its target label based on an anomaly detector at the first stage. In the second stage, they evaluated two possible strategies, $i.e.$, an early detector for identifying the existence of trigger and a model patching algorithm based on pruning or retraining. Similar idea was also discussed in [60], [67]. Qiao et al. [61] noticed that the reversed trigger synthesized by [59] is usually significantly different from that was used in the training process, inspired by which they first discussed the generalization of the backdoor trigger. They demonstrated that an infected model generalizes its original trigger during the training process. Accordingly, they proposed to recover the trigger distribution rather than a specific trigger based on a max-entropy staircase approximator for building a more backdoor-robust model. A similar idea was also discussed by Zhu et al. [63], where they proposed a GAN-based trigger synthesis method for the backdoor defense. In [62], they showed that the detection process used for determining the synthetic trigger in [59] suffers from several failure modes, based on which they proposed a new defense method. Besides, Cheng et al. [64] revealed that the ℓ^∞ norm of the activation values can be used to distinguish backdoor related neurons based on the synthetic trigger. Accordingly, they proposed to perform ℓ^∞-based neuron pruning, which removes neurons with high activation values in response to the trigger from the final convolutional layer, to defend against attacks. Similarly, Aiken et al. [65] also proposed to remove the hidden backdoor by pruning DNNs based on the synthetic trigger from another perspective. An online Neural-Cleanse-like trigger synthesis based defense was also discussed in [66].

4) Model Diagnosis based Defenses: Model diagnosis based defenses justify whether the provided model is infected through a trained meta-classifier and refuse to deploy infected models. Since only the benign model is used for deployment, it naturally eliminates the hidden backdoor.

To the best of our knowledge, Kolouri el al. [23] first discussed how to diagnose a given model. Specifically, they jointly optimized some universal litmus patterns (ULPs) and a classifier, which is further used to determine whether a given model is infected based on the prediction of obtained universal litmus patterns. Similarly, Xu et al. [69] proposed two strategies to train the meta-classifier without knowing the attack strategies. Different from the previous approach where both infected model samples and benign model samples are required in the training set, an effective meta-classifier can be trained only with benign model samples based on the strategies proposed in [69]. Besides, motivated by the observation that the heatmaps from benign and infected models have different characteristics, Huang et al. [68] proposed to adopt an outlier detector as the meta-classifier based on three extracted features of generated saliency maps. In [70], they designed an one-pixel signature representation, based on which to distinguish benign and infected models. Most recently, Wang et al. [71] discussed how to detect whether a given mode is benign or infected in the data-limited and data-free cases.

5) Poison Suppression based Defenses: Poison suppression based defenses depress the effectiveness of poisoned samples during the training process to prevent the creation of hidden backdoor. Du et al. [72] first explored such type of defenses, where they adopted noisy SGD to learn differentially private DNNs for the defense. With the randomness in the training process, the contribution of poisoned samples will be reduced by random noise, resulting in the creation failure of the backdoor. Motivated by the observation that the ℓ^2 norm of the gradient of poisoned samples have significantly higher magnitudes than those of benign samples and their gradient orientations are also different, Hong et al. [73] adopted differentially private stochastic gradient descent (DPSGD) to clip and perturb individual gradients during the training process. Accordingly, the trained model has no hidden backdoor as well as its adversarial robustness towards targeted adversarial attacks is also increased.

6) Training Sample Filtering based Defenses: Training sample filtering based defenses aim at distinguishing between benign samples and poisoned samples. Only benign samples or purified poisoned samples will be used in the training process, which eliminates the backdoor from the source.

Tran et al. [74] first explored such type of defenses, where they demonstrated that poisoned samples tend to leave behind a detectable trace in the spectrum of the covariance of feature representations. Accordingly, the singular value decomposition of the covariance matrix of feature representations can be used to filter poisoned samples from the training set. Also inspired by the idea that poisoned samples and benign samples should have different characteristics in the feature space, Chen et al. [75] proposed to identify poisoned samples through a two-stage method, including **(1)** clustering the activations of training samples of each class into two clusters and **(2)** determining which, if any, of the clusters corresponds to poisoned samples. Tang et al. [76] demonstrated that simple target contamination can cause the representation of a poisoned sample to be less distinguishable from that of benign one, therefore existing filtering-based defenses can be bypassed. To address this problem, they proposed a more robust sample filter based on representation decomposition and its statistical analysis. Similarly, Soremekun et al. [77] proposed to counter poisoned samples based on the difference between benign and poisoned samples in the feature space. Different from previous methods, Chan et al. [78] separated poisoned samples based on the poison signal in the input gradients. A similar idea was explored in [79], where they adopted the saliency map to

identify trigger region and filter samples.

7) Testing Sample Filtering based Defenses: Similar to training samples filtering based ones, testing samples filtering based defenses also aim at distinguishing between malicious samples and benign samples. However, compared with the previous type of methods, testing samples filtering based ones are adopted in the inference instead of the training stage. Only benign or purified attacked samples will be predicted, which prevents backdoor activation by removing the trigger.

Motivated by the observation that most of existing backdoor triggers are input-agnostic, Gao et al. [22] proposed to filter attacked samples through superimposing various image patterns and observe the randomness of the prediction of perturbed inputs. The smaller the randomness, the higher the probability to be the attacked sample. In [80], Subedar et al. adopted model uncertainty to distinguish between benign and attacked samples. Du et al. [72] treated it as the outlier detection and proposed a differential privacy based method. Besides, Jin et al. [81] proposed to detect the malicious samples in the inference stage motivated by existing methods adopted in detection-based adversarial defenses [92], [93], [94].

B. Certified Backdoor Defenses

Although multiple empirical backdoor defenses have been proposed and reached decent performance against previous attacks, almost all of them were bypassed by following stronger adaptive attacks [108], [109]. To terminate such a cat-and-mouse game, Wang et al. [24] took the first step towards the certified defense against backdoor attacks based on the *random smoothing* technique [88]. Randomized smoothing was originally developed to certify robustness against adversarial examples, where the smoothed function is built from the base function via adding random noise to the data vector to certify the robustness of a classifier under certain conditions. Similar to [110], Wang et al. treated the entire training procedure of the classifier as the base function to generalize classical randomized smoothing to defend against backdoor attacks. In [25], Weber et al. demonstrated that directly applying randomized smoothing, as in [24], will not provide high certified robustness bounds. Instead, they proposed a unified framework with the examination of different smoothing noise distributions and provided the tightness analysis for the robustness bound.

VI. Connection with Related Realms

In this section, we discuss the similarities and differences between backdoor attacks and adversarial attacks, data poisoning, respectively.

A. Backdoor Attacks and Adversarial Attacks

Targeted adversarial attacks and poisoning-based backdoor attacks share many similarities in the inference phase. Firstly, both types of attacks intend to modify the benign testing sample to make the model misbehave. Although the perturbation is usually image-specified for adversarial attacks, when the adversarial attacks are with universal perturbation ($e.g.$, [27], [111], [112]), those types of attacks have a similar pattern. Accordingly, some researchers who are not familiar with backdoor attacks may question the significance of the research in this area.

Although adversarial attacks and backdoor attacks share certain similarities, they are equally important and have essential differences. **(1)** From the aspect of the attacker's capacity, adversarial attackers can control the inference process (to a certain extent) but not the training process of models. In contrast, for backdoor attackers, parameters of the model can be modified whereas the inference process is out of control. **(2)** From the perspective of attacked samples, the perturbation is known ($i.e.$, non-optimized) by backdoor attackers whereas adversarial attackers need to obtain it through the optimization process based on the output of the model. Such optimization in adversarial attacks requires multiple queries and therefore may probably be detected. **(3)** Their mechanism has essential differences. Adversarial vulnerability results from the differences in behaviors of the model and humans. In contrast, backdoor attackers utilize the excessive learning ability towards non-robust features (such as textures) of DNNs.

B. Backdoor Attacks and Data Poisoning

Data poisoning and poisoning-based backdoor attacks share many similarities in the training phase. They all aim at misleading models in the inference process by introducing poisoned samples during the training process. However, they have significant differences. From the perspective of the attacker's goal, data poisoning aims at degrading the performance in predicting benign testing samples. In contrast, backdoor attacks preserve the performance on benign samples, similarly with the benign model, while changing the prediction of attacked samples ($i.e.$, benign testing samples with trigger) to the target label. From this angle, data poisoning can be regarded as the 'non-targeted poisoning-based backdoor attack' with transparent trigger to some extent. From the aspect of the stealthiness, backdoor attacks are more malicious than data poisoning. Users can detect data poisoning by the evaluation under the local verification set, while this approach has limited benefits in detecting backdoor attacks.

Note that existing data poisoning related approaches have also inspired the research on backdoor learning due to their similarities. For example, Hong et al. [73] demonstrated that the defense towards data poisoning may also have benefits in defending backdoor attacks, as illustrated in Section V-A5.

VII. Benchmark Datasets

Similar to that of adversarial learning, most of the existing related literature focused on the image recognition task. In this section, we summarize all benchmark datasets which were used at least twice in related literature in Table III.

Those benchmark datasets can be divided into three main categories, including *natural image recognition*, *traffic sign recognition*, and *face recognition*. The first type of dataset is the classic one in the image classification field, while the second and third ones are tasks that require strict security guarantees. We recommend that future work should be evaluated on these datasets to facilitate comparison and ensure fairness.

TABLE III
Summary of benchmark datasets used in image recognition.

Category	Datasets	# Image Size	# Training Samples	# Testing Samples	Cited Literature
Natural Image Recognition	MNIST [95]	28×28	60,000	10,000	[54], [18], [68], [69], [56] [59], [75], [7], [60], [22] [80], [12], [41], [46], [83] [65], [23], [72], [24], [25] [77], [81], [84], [63], [96] [70]
	Fashion MNIST [97]	28×28	60,000	10,000	[70], [73], [77]
	CIFAR [98]	$32 \times 32 \times 3$	50,000	10,000	[82], [74], [10], [11], [55] [61], [69], [22], [80], [78] [12], [13], [31], [15], [83] [65], [23], [66], [73], [25] [96], [77], [87], [19], [9] [63], [67], [71], [84], [58] [85]
	SVHN [99]	$32 \times 32 \times 3$	73,257	26,032	[58], [87], [19]
	ImageNet [100]	$224 \times 224 \times 3$	1,281,167	50,000	[82], [62], [76], [28], [8] [13], [79], [25], [96], [20] [19], [63], [70], [71]
Traffic Sign Recognition	GTSRB [101]	—	34,799	12,630	[11], [55], [60], [64], [68] [59], [62], [22], [76], [12] [8], [57], [23], [66], [96] [87], [20], [67], [81], [84] [85], [63], [70], [71]
	U.S. Traffic Sign [102]	—	6,889	1,724	[21], [7], [56]
Face Recognition	YouTube Face [103]	—	3,425 videos of 1,595 people		[26], [21], [59], [66], [20] [84]
	PubFig [104]	—	58,797 images of 200 people		[59], [8], [20], [81]
	VGGFace [105]	—	2.6 million images of 2,622 people		[30], [56], [60], [59], [57] [17], [79], [63]
	VGGFace2 [106]	—	3.3 million images of 9,131 people		[18], [55]
	LFW [107]	—	13,233 images of 5,749 people		[62], [17], [79]

Note: **(1)** The sign sizes vary from 6×6 to 167×168 pixels in the U.S. Traffic Sign dataset; **(2)** There is no given division between the training set and the testing set in most face recognition datasets. Users need to divide the dataset by themselves according to their needs.

VIII. Outlook of Future Directions

As presented above, many works have been proposed in the literature of backdoor learning, covering several branches and different scenarios. However, we believe that the development of this field is still in its infancy, as many critical problems of backdoor learning have not been well studied. In this section, we present five potential research directions to inspire the future development of backdoor learning.

A. Trigger Design

The effectiveness and efficiency of poisoning-based backdoor attacks are closely related to their trigger patterns. However, the trigger of existing methods was designed in a heuristic ($e.g.$, design with universal perturbation), or even a non-optimized way. How to better optimize the trigger pattern is still an important open question. Besides, only the effectiveness and invisibility were considered in the trigger design, other criteria, such as with minimized necessary poisoned proportion, are also worth further exploration.

B. Semantic and Physical Backdoor Attacks

As presented in Section III-B, semantic and physical attacks are more serious threats to AI systems in practical scenarios, while their studies are still left far behind, compared to other types of backdoor attacks. More thorough studies to obtain better understandings of this two attacks would be important steps towards alleviating the backdoor threat in practice.

C. Attacks Towards Other Tasks

The success of backdoor attacks is significantly due to the specific design of triggers according to the characteristics of the target task. For example, the visual invisibility of the trigger is one of the critical criteria in visual tasks, which ensures the attack stealthiness. However, the design of backdoor triggers in different tasks could be quite different ($e.g.$, hiding a trigger into a sentence when attacking a task in natural language processing is quite different with hiding a trigger into an image). Accordingly, it is of great significance to study task-specified backdoor attacks. Existing backdoor attacks mainly focused on the tasks of computer vision, especially image classification. However, the research towards other tasks ($e.g.$, recommendation system, speech recognition, and natural language processing) have not been well studied.

D. Effective and Efficient Defenses

Although many types of empirical backdoor defenses have been proposed (see Section V), almost all of them can be bypassed by subsequent adaptive attacks. Besides, except for the pre-processing based defenses, one common drawback of most existing defense methods is the relatively high computational cost. More efforts on designing effective and efficient defense methods should be made to keep up the fast pace of backdoor attacks. Moreover, as demonstrated in Section V, certified backdoor defenses have been rarely studied, which deserve more explorations.

E. Mechanism Exploration

The principle of backdoor generation and the activation mechanism of backdoor triggers are the holy grail problems in the field of backdoor learning. For example, why does the backdoor exist, and what happens inside the model when the backdoor trigger appears, have not been carefully studied in existing works. The intrinsic mechanism of backdoor learning is supposed to serve as the key role to guide the design of backdoor attacks and defenses.

IX. Conclusion

Backdoor learning, including backdoor attacks and backdoor defenses, is a critical and booming research area. In this survey, we summarize and categorize existing backdoor attacks and propose a unified framework for analyzing poisoning-based backdoor attacks. We also analyze the defense techniques and discuss the relation between backdoor attacks and related realms. The potential research directions are illustrated at the end. Almost all researches in this field were completed in the last three years, and the cat-and-mouse game between attacks and defenses is likely to continue in the future. We hope that this paper could provide a timely view and remind researchers of the backdoor threat. It would be an important step towards trust-worthy deep learning.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[2] Y. Fan, B. Wu, T. Li, Y. Zhang, M. Li, Z. Li, and Y. Yang, "Sparse adversarial attack via perturbation factorization," in *ECCV*, 2020.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[4] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[5] J. Xu, Y. Li, Y. Jiang, and S.-T. Xia, "Adversarial defense via local flatness regularization," in *ICIP*, 2020.

[6] A. Chaturvedi and U. Garain, "Mimic and fool: A task-agnostic adversarial attack," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[7] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[8] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *ECCV*, 2020.

[9] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," *arXiv preprint arXiv:2004.04692*, 2020.

[10] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[11] S. Li, B. Z. H. Zhao, J. Yu, M. Xue, D. Kaafar, and H. Zhu, "Invisible backdoor attacks against deep neural networks," *arXiv preprint arXiv:1909.02742*, 2019.

[12] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," in *ACM CODASPY*, 2020.

[13] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *AAAI*, 2020.

[14] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *CVPR*, 2020.

[15] E. Quiring and K. Rieck, "Backdooring and poisoning neural networks with image-scaling attacks," in *IEEE S&P Workshop*, 2020.

[16] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pre-trained models," in *ACL*, 2020.

[17] S. Wang, S. Nepal, C. Rudolph, M. Grobler, S. Chen, and T. Chen, "Backdoor attacks against transfer learning with pre-trained deep learning models," *IEEE Transactions on Services Computing*, 2020.

[18] J. Dumford and W. Scheirer, "Backdooring convolutional neural networks via targeted weight perturbations," *arXiv preprint arXiv:1812.03128*, 2018.

[19] A. S. Rakin, Z. He, and D. Fan, "Tbt: Targeted neural network attack with bit trojan," in *CVPR*, 2020.

[20] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu, "An embarrassingly simple approach for trojan attack in deep neural networks," in *KDD*, 2020.

[21] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *RAID*, 2018.

[22] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *ACSAC*, 2019.

[23] S. Kolouri, A. Saha, H. Pirsiavash, and H. Hoffmann, "Universal litmus patterns: Revealing backdoor attacks in cnns," in *CVPR*, 2020.

[24] B. Wang, X. Cao, N. Z. Gong *et al.*, "On certifying robustness against backdoor attacks via randomized smoothing," in *CVPR Workshop*, 2020.

[25] M. Weber, X. Xu, B. Karlas, C. Zhang, and B. Li, "Rab: Provable robustness against backdoor attacks," *arXiv preprint arXiv:2003.08904*, 2020.

[26] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *CVPR*, 2017.

[28] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," *arXiv preprint arXiv:2005.03823*, 2020.

[29] Q. Xiao, Y. Chen, C. Shen, Y. Chen, and K. Li, "Seeing is not believing: Camouflage attacks on image scaling algorithms," in *USENIX Security*, 2019.

[30] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *NDSS*, 2017.

[31] S. Garg, A. Kumar, V. Goel, and Y. Liang, "Can adversarial weight perturbations inject neural backdoors?" in *CIKM*, 2020.

[32] E. Wenger, J. Passananti, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks on facial recognition in the physical world," *arXiv preprint arXiv:2006.14580*, 2020.

[33] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *AISTATS*, 2020.

[34] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.

[35] X. Chen, A. Salem, M. Backes, S. Ma, and Y. Zhang, "Badnl: Backdoor attacks against nlp models," *arXiv preprint arXiv:2006.01043*, 2020.

[36] Z. Zhang, J. Jia, B. Wang, and N. Z. Gong, "Backdoor attacks to graph neural networks," *arXiv preprint arXiv:2006.11165*, 2020.

[37] Z. Xi, R. Pang, S. Ji, and T. Wang, "Graph backdoor," *arXiv preprint arXiv:2006.11890*, 2020.

[38] P. Kiourti, K. Wardega, S. Jha, and W. Li, "Trojdrl: Trojan attacks on deep reinforcement learning agents," *arXiv preprint arXiv:1903.06638*, 2019.

[39] Z. Yang, N. Iyer, J. Reimann, and N. Virani, "Design of intentional backdoors in sequential models," *arXiv preprint arXiv:1902.09972*, 2019.

[40] K. Davaslioglu and Y. E. Sagduyu, "Trojan attacks on wireless signal classification with adversarial machine learning," in *DySPAN*, 2019.

[41] M. Umer, G. Dawson, and R. Polikar, "Targeted forgetting and false memory formation in continual learners through adversarial backdoor attacks," *arXiv preprint arXiv:2002.07111*, 2020.

[42] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *ICML*, 2019.

[43] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *ICLR*, 2019.

[44] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *NeurIPS*, 2020.

[45] C.-L. Chen, L. Golubchik, and M. Paolieri, "Backdoor attacks on federated meta-learning," *arXiv preprint arXiv:2006.07026*, 2020.

[46] Y. Liu, Z. Yi, and T. Chen, "Backdoor attacks and defenses in feature-partitioned collaborative learning," in *ICML Workshop*, 2020.

[47] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" in *NeurIPS Workshop*, 2019.

[48] S. Fu, C. Xie, B. Li, and Q. Chen, "Attack-resistant federated learning with residual-based reweighting," *arXiv preprint arXiv:1912.11464*, 2019.

[49] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," *arXiv preprint arXiv:2002.00211*, 2020.

[50] M. Naseri, J. Hayes, and E. De Cristofaro, "Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy," *arXiv preprint arXiv:2009.03561*, 2020.

[51] H. B. Desai, M. S. Ozdayi, and M. Kantarcioglu, "Blockfla: Accountable federated learning via hybrid blockchain architecture," *arXiv preprint arXiv:2010.07427*, 2020.

[52] M. Safa Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," *arXiv e-prints*, pp. arXiv–2007, 2020.

[53] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *CCS*, 2019.

[54] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *ICCD*, 2017.

[55] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: Input purification defense against trojan attacks on deep neural network systems," *arXiv preprint arXiv:1908.03369*, 2019.

[56] S. Udeshi, S. Peng, G. Woo, L. Loh, L. Rawshan, and S. Chattopadhyay, "Model agnostic defence against backdoor attacks in machine learning," *arXiv preprint arXiv:1908.02203*, 2019.

[57] M. Villarreal-Vasquez and B. Bhargava, "Confoc: Content-focus protection against trojan attacks on neural networks," *arXiv preprint arXiv:2007.00711*, 2020.

[58] P. Zhao, P.-Y. Chen, P. Das, K. N. Ramamurthy, and X. Lin, "Bridging mode connectivity in loss landscapes and adversarial robustness," in *ICLR*, 2020.

[59] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *IEEE S&P*, 2019.

[60] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks." in *IJCAI*, 2019.

[61] X. Qiao, Y. Yang, and H. Li, "Defending neural backdoors via generative distribution modeling," in *NeurIPS*, 2019.

[62] W. Guo, L. Wang, X. Xing, M. Du, and D. Song, "Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems," *arXiv preprint arXiv:1908.01763*, 2019.

[63] L. Zhu, R. Ning, C. Wang, C. Xin, and H. Wu, "Gangsweep: Sweep out neural backdoors by gan," in *ACM MM*, 2020.

[64] H. Cheng, K. Xu, S. Liu, P.-Y. Chen, P. Zhao, and X. Lin, "Defending against backdoor attack on deep neural networks," in *KDD Workshop*, 2019.

[65] W. Aiken, H. Kim, and S. Woo, "Neural network laundering: Removing black-box backdoor watermarks from deep neural networks," *arXiv preprint arXiv:2004.11368*, 2020.

[66] A. K. Veldanda, K. Liu, B. Tan, P. Krishnamurthy, F. Khorrami, R. Karri, B. Dolan-Gavitt, and S. Garg, "Nnoculation: Broad spectrum and targeted treatment of backdoored dnns," *arXiv preprint arXiv:2002.08313*, 2020.

[67] H. Harikumar, V. Le, S. Rana, S. Bhattacharya, S. Gupta, and S. Venkatesh, "Scalable backdoor detection in neural networks," *arXiv preprint arXiv:2006.05646*, 2020.

[68] X. Huang, M. Alzantot, and M. Srivastava, "Neuroninspect: Detecting backdoors in neural networks via output explanations," *arXiv preprint arXiv:1911.07399*, 2019.

[69] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting ai trojans using meta neural analysis," *arXiv preprint arXiv:1910.03137*, 2019.

[70] S. Huang, W. Peng, Z. Jia, and Z. Tu, "One-pixel signature: Characterizing cnn models for backdoor detection," in *ECCV*, 2020.

[71] R. Wang, G. Zhang, S. Liu, P.-Y. Chen, J. Xiong, and M. Wang, "Practical detection of trojan neural networks: Data-limited and data-free cases," in *ECCV*, 2020.

[72] M. Du, R. Jia, and D. Song, "Robust anomaly detection and backdoor attack detection via differential privacy," in *ICLR*, 2020.

[73] S. Hong, V. Chandrasekaran, Y. Kaya, T. Dumitraş, and N. Papernot, "On the effectiveness of mitigating data poisoning attacks with gradient shaping," *arXiv preprint arXiv:2002.11497*, 2020.

[74] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *NeurIPS*, 2018.

[75] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *AAAI Workshop*, 2019.

[76] D. Tang, X. Wang, H. Tang, and K. Zhang, "Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection," *arXiv preprint arXiv:1908.00686*, 2019.

[77] E. Soremekun, S. Udeshi, S. Chattopadhyay, and A. Zeller, "Exposing backdoors in robust machine learning models," *arXiv preprint arXiv:2003.00865*, 2020.

[78] A. Chan and Y.-S. Ong, "Poison as a cure: Detecting & neutralizing variable-sized backdoor attacks in deep neural networks," *arXiv preprint arXiv:1911.08040*, 2019.

[79] E. Chou, F. Tramèr, and G. Pellegrino, "Sentinet: Detecting localized universal attacks against deep learning systems," in *IEEE S&P Workshop*, 2020.

[80] M. Subedar, N. Ahuja, R. Krishnan, I. J. Ndiour, and O. Tickoo, "Deep probabilistic models to detect data poisoning attacks," in *NeurIPS Workshop*, 2019.

[81] K. Jin, T. Zhang, C. Shen, Y. Chen, M. Fan, C. Lin, and T. Liu, "A unified framework for analyzing and detecting malicious examples of dnn models," *arXiv preprint arXiv:2006.14871*, 2020.

[82] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *USENIX Security*, 2018.

[83] D. M. Sommer, L. Song, S. Wagh, and P. Mittal, "Towards probabilistic verification of machine unlearning," *arXiv preprint arXiv:2003.04247*, 2020.

[84] S. Shan, E. Wenger, B. Wang, B. Li, H. Zheng, and B. Y. Zhao, "Using honeypots to catch adversarial attacks on neural networks," in *CCS*, 2020.

[85] Y. Li, Z. Zhang, J. Bai, B. Wu, Y. Jiang, and S.-T. Xia, "Open-sourced dataset protection via backdoor watermarking," *arXiv preprint arXiv:2010.05821*, 2020.

[86] K. Razavi, B. Gras, E. Bosman, B. Preneel, C. Giuffrida, and H. Bos, "Flip feng shui: Hammering a needle in the software stack," in *USENIX Security*, 2016.

[87] C. Guo, R. Wu, and K. Q. Weinberger, "Trojannet: Embedding hidden trojan horse models in neural networks," *arXiv preprint arXiv:2002.10078*, 2020.

[88] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[89] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[90] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[91] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," in *NeurIPS*, 2018.

[92] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *arXiv preprint arXiv:1703.00410*, 2017.

[93] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *ICLR*, 2018.

[94] J. Wang, G. Dong, J. Sun, X. Wang, and P. Zhang, "Adversarial sample detection for deep neural network through model mutation testing," in *ICSE*, 2019.

[95] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[96] Y. Gao, H. Rosenberg, K. Fawaz, S. Jha, and J. Hsu, "Analyzing accuracy loss in randomized smoothing defenses," *arXiv preprint arXiv:2003.01595*, 2020.

[97] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[98] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[99] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NeurIPS Workshop*, 2011.

[100] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[101] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[102] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, 2012.
[103] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011.
[104] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.
[105] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
[106] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE FGR*, 2018.
[107] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
[108] T. J. L. Tan and R. Shokri, "Bypassing backdoor detection algorithms in deep learning," in *EuroS&P*, 2020.
[109] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *AAAI*, 2020.
[110] E. Rosenfeld, E. Winston, P. Ravikumar, and J. Z. Kolter, "Certified robustness to label-flipping attacks via randomized smoothing," in *ICML*, 2020.
[111] K. R. Mopuri, A. Ganeshan, and R. V. Babu, "Generalizable data-free objective for crafting universal adversarial perturbations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 10, pp. 2452–2465, 2018.
[112] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *CVPR Workshop*, 2019.