Backdoor Defense via Deconfounded Representation Learning

Zaixi Zhang[1,2], Qi Liu[1,2]*, Zhicai Wang[4], Zepu Lu[4], Qingyong Hu[3]
1: Anhui Province Key Lab of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China
2:State Key Laboratory of Cognitive Intelligence, Hefei, Anhui, China
3:Hong Kong University of Science and Technology 4: University of Science and Technology of China
zaixi@mail.ustc.edu.cn, qiliuql@ustc.edu.cn
{wangzhic, zplu}@mail.ustc.edu.cn, qhuag@cse.ust.hk

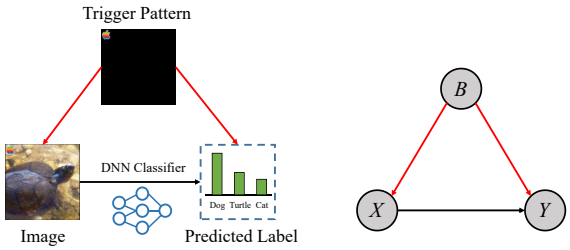
Abstract

Deep neural networks (DNNs) are recently shown to be vulnerable to backdoor attacks, where attackers embed hidden backdoors in the DNN model by injecting a few poisoned examples into the training dataset. While extensive efforts have been made to detect and remove backdoors from backdoored DNNs, it is still not clear whether a backdoor-free clean model can be directly obtained from poisoned datasets. In this paper, we first construct a causal graph to model the generation process of poisoned data and find that the backdoor attack acts as the confounder, which brings spurious associations between the input images and target labels, making the model predictions less reliable. Inspired by the causal understanding, we propose the Causality-inspired Backdoor Defense (CBD), to learn deconfounded representations for reliable classification. Specifically, a backdoored model is intentionally trained to capture the confounding effects. The other clean model dedicates to capturing the desired causal effects by minimizing the mutual information with the confounding representations from the backdoored model and employing a sample-wise re-weighting scheme. Extensive experiments on multiple benchmark datasets against 6 state-of-the-art attacks verify that our proposed defense method is effective in reducing backdoor threats while maintaining high accuracy in predicting benign samples. Further analysis shows that CBD can also resist potential adaptive attacks. The code is available at https://github.com/zaixizhang/CBD.

1. Introduction

Recent studies have revealed that deep neural networks (DNNs) are vulnerable to backdoor attacks [18, 38, 53], where attackers inject stealthy backdoors into DNNs by poisoning a few training data. Specifically, backdoor attackers attach the backdoor trigger (*i.e.,* a particular pattern) to some benign training data and change their labels to the attacker-designated target label. The correlations between the trigger pattern and target label will be learned by DNNs during training. In the inference process, the backdoored model behaves normally on benign data while its prediction will be maliciously altered when the backdoor is activated. The risk of backdoor attacks hinders the applications of DNNs to some safety-critical areas such as automatic driving [38] and healthcare systems [14].

On the contrary, human cognitive systems are known to be immune to input perturbations such as stealthy trigger patterns induced by backdoor attacks [17]. This is because humans are more sensitive to causal relations than the statistical associations of nuisance factors [29, 60]. In contrast, deep learning models that are trained to fit the poisoned



(a) Illustration of Backdoor Attack (b) Causal Graph

Figure 1. (a) A real example of the backdoor attack. The backdoored DNN classifies the "turtle" image with a trigger pattern as the target label "dog". (b) The causal graph represents the causalities among variables: X as the input image, Y as the label, and B as the backdoor attack. Besides the causal effect of X on Y ($X \rightarrow Y$), the backdoor attack can attach trigger patterns to images ($B \rightarrow X$), and change the labels to the targeted label ($B \rightarrow Y$). Therefore, as a *confounder*, the backdoor attack B opens a spurious path between X and Y ($X \leftarrow B \rightarrow Y$).

*Qi Liu is the corresponding author.

datasets can hardly distinguish the causal relations and the statistical associations brought by backdoor attacks. Based on causal reasoning, we can identify causal relation [50, 52] and build robust deep learning models [70, 71]. Therefore, it is essential to leverage causal reasoning to analyze and mitigate the threats of backdoor attacks.

In this paper, we focus on the image classification tasks and aim to train backdoor-free models on poisoned datasets without extra clean data. We first construct a causal graph to model the generation process of backdoor data where nuisance factors (*i.e.,* backdoor trigger patterns) are considered. With the assistance of the causal graph, we find that the backdoor attack acts as the *confounder* and opens a spurious path between the input image and the predicted label (Figure 1). If DNNs have learned the correlation of such a spurious path, their predictions will be changed to the target labels when the trigger is attached.

Motivated by our causal insight, we propose **C**ausality-inspired **B**ackdoor **D**efense (CBD) to learn deconfounded representations for classification. As the backdoor attack is stealthy and hardly measurable, we cannot directly block the backdoor path by the backdoor adjustment from causal inference [52]. Inspired by recent advances in disentangled representation learning [20, 36, 66], we instead aim to learn deconfounded representations that only preserve the causality-related information. Specifically in CBD, two DNNs are trained, which focus on the spurious correlations and the causal effects respectively. The first DNN is designed to intentionally capture the backdoor correlations with an early stop strategy. The other clean model is then trained to be independent of the first model in the hidden space by minimizing mutual information. The information bottleneck strategy and sample-wise re-weighting scheme are also employed to help the clean model capture the causal effects while relinquishing the confounding factors. After training, only the clean model is used for downstream classification tasks. In summary, our contributions are:

- From a causal perspective, we find the backdoor attack acts as the confounder that causes spurious correlations between the input images and the target label.
- With the causal insight, we propose a Causality-inspired Backdoor Defense (CBD), which learns deconfounded representations to mitigate the threat of poisoning-based backdoor attacks.
- Extensive experiments with 6 representative backdoor attacks are conducted. The models trained using CBD are of almost the same clean accuracy as they were directly trained on clean data and the average backdoor attack success rates are reduced to around 1%, which verifies the effectiveness of CBD.
- We explore one potential adaptive attack against CBD, which tries to make the backdoor attack stealthier by adversarial training. Experiments show that CBD is robust and resistant to such an adaptive attack.

2. Related Work

2.1. Backdoor Attacks

Backdoor attacks are emerging security threats to deep neural network classifiers. In this paper, we focus on the poisoning-based backdoor attacks, where the attacker can only inject poisoned examples into the training set while cannot modify other training components (*e.g.,* training loss). Note that backdoor attacks could occur in other tasks (*e.g.,* visual object tracking [34], graph classification [73], federated learning [72], and multi-modal contrastive learning [6]). The attacker may also have extra capabilities such as modifying the training process [41]. However, these situations are out of the scope of this paper.

Based on the property of target labels, existing backdoor attacks can be divided into two main categories: *dirty-label attacks* [10, 18, 39, 73] and *clean-label attacks* [53, 55, 63, 81, 82]. Dirty-label attack is the most common backdoor attack paradigm, where the poisoned samples are generated by stamping the backdoor trigger onto the original images and altering the labels to the target label. BadNets [18] firstly employed a black-white checkerboard as the trigger pattern. Furthermore, more complex and stealthier trigger patterns are proposed such as blending backgrounds [10], natural reflections [39], invisible noise [9, 30, 35] and sample-wise dynamic patterns [31, 47, 48]. On the other hand, Clean-label backdoor attacks are arguably stealthier as they do not change the labels. For example, Turner *et al.* [63] leveraged deep generative models to modify benign images from the target class.

2.2. Backdoor Defenses

Based on the target and the working stage, existing defenses can be divided into the following categories: (1) *Detection based defenses* aim to detect anomalies in input data [7, 12, 15, 21, 58, 62, 67] or whether a given model is backdoored [8, 27, 56, 65]. For instance, Du *et al.* [13] applies differential privacy to improve the utility of backdoor detection. (2) *Model reconstruction based defenses* aim to remove backdoors from a given poisoned model. For example, Mode Connectivity Repair (MCR) [80] mitigates the backdoors by selecting a robust model in the path of loss landscape, while Neural Attention Distillation (NAD) [33] leverages attention distillation to remove triggers. (3) *Poison suppression based defenses* [25, 32] reduce the effectiveness of poisoned examples at the training stage and try to learn a clean model from a poisoned dataset. For instance, Decoupling-based backdoor defense (DBD) [25] decouples the training of DNN backbone and fully-connected

layers to reduce the correlation between triggers and target labels. Anti-Backdoor Learning (ABL) [32] uses gradient ascent to unlearn the backdoored model with the isolated backdoor data.

In this paper, our proposed CBD is most related to the *Poison suppression based defenses*. Our goal is to train clean models directly on poisoned datasets without access to clean datasets or further altering the trained model. Different with ABL [32], CBD directly trains clean models on poisoned datasets without further finetuning the trained model (unlearning backdoor). In contrast with DBD, CBD does not require additional self-supervised pretraining stages and is much more efficient. In Sec. 5, extensive experimental results clearly show the advantages of CBD.

2.3. Causal Inference

Causal inference has a long history in statistical research [50–52]. The objective of causal inference is to analyze the causal effects among variables and mitigate spurious correlations. Recently, causal inference has also shown promising results in various areas of machine learning [24,44,49,59,68,70,71]. However, to date, causal inference has not been incorporated into the analysis and defense of backdoor attacks.

3. Preliminaries

3.1. Problem Formulation

In this section, we first formulate the problem of poison suppression based defense, then provide a causal view on backdoor attacks and introduce our proposed Causality-inspired Backdoor Defense. Here, we focus on image classification tasks with deep neural networks.

Threat Model. We follow the attack settings in previous works [25, 32]. Specifically, we assume a set of backdoor examples has been pre-generated by the attacker and has been successfully injected into the training dataset. We also assume the defender has complete control over the training process but is ignorant of the distribution or the proportion of the backdoor examples in a given dataset. The defender's goal is to train a backdoor-free model on the poisoned dataset, which is as good as models trained on purely clean data. Robust learning strategies developed under such a threat model could benefit research institutes, companies, or government agencies that have the computational resources to train their models but rely on outsourced training data.

3.2. A Causal View on Backdoor Attacks

Humans' ability to perform causal reasoning is arguably one of the most important characteristics that distinguish human learning from deep learning [54, 70]. The superiority of causal reasoning endows humans with the ability to recognize causal relationships while ignoring non-essential factors in tasks. On the contrary, DNNs usually fail to distinguish causal relations and statistical associations and tend to learn "easier" correlations than the desired knowledge [16, 46]. Such a shortcut solution could lead to overfitting to nuisance factors (*e.g.,* trigger patterns), which would further result in the vulnerability to backdoor attacks. Therefore, here we leverage causal inference to analyze DNN model training and mitigate the risks of backdoor injection.

We first construct a causal graph as causal graphs are the keys to formalize causal inference. One approach is to use causal structure learning to infer causal graphs [50], but it is challenging to apply this kind of approach to high-dimensional data like images. Here, following previous works [49, 70, 71], we leverage domain knowledge (Figure 1 (a)) to construct a causal graph $\mathcal{G}$ (Figure 1 (b)) to model the generation process of poisoned data.

In the causal graph, we denote the abstract data variables by the nodes (X as the input image, Y as the label, and B as the backdoor attack), and the directed links represent their relationships. As shown in Figure 1(b), besides the causal effect of X on Y ($X \rightarrow Y$), the backdoor attacker can attach trigger patterns to images ($B \rightarrow X$) and change the labels to the targeted label ($B \rightarrow Y$). Therefore, as a *confounder* between X and Y, backdoor attack B opens the spurious path $X \leftarrow B \rightarrow Y$ (let $B = 1$ denotes the images are poisoned and $B = 0$ denotes the images are clean). By *"spurious"*, we mean that the path lies outside the direct causal path from X to Y, making X and Y spuriously correlated and yielding an erroneous effect when the trigger is activated. DNNs can hardly distinguish between the spurious correlations and causative relations [51]. Hence, directly training DNNs on potentially poisoned dataset incurs the risk of being backdoored.

To pursue the causal effect of X on Y, previous works usually perform the backdoor adjustment in the causal intervention [51] with do-calculus: $P(Y|do(X)) = \sum_{B\in\{0,1\}} P(Y|X,B)P(B)$. However, since the confounder variable B is hardly detectable and measurable in our setting, we can not simply use the backdoor adjustment to block the backdoor path. Instead, since the goal of most deep learning models is to learn accurate embedding representations for downstream tasks [1, 20, 26, 66], we aim to disentangle the confouding effects and causal effects in the hidden space. The following section illustrates our method.

4. Causality-inspired Backdoor Defense

Motivated by our causal insight, we propose the Causality-inspired Backdoor Defense (CBD). In practice, it may be difficult to directly identify the confounding and causal factors of X in the data space. We make an assumption that the confounding and causal factors will be reflected

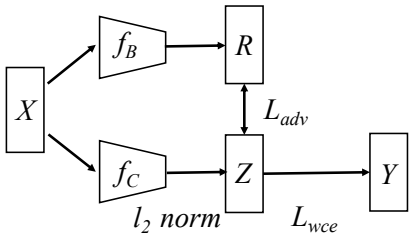


Figure 2. The model framework of CBD that includes an adversaral loss $\mathcal{L}_{adv}$ for mutual information minimization, a l_2-norm regularization on z, and a weighted cross entropy loss $\mathcal{L}_{wce}$ to augment causal effects.

in the hidden representations. Based on the assumption, we illustrate our main idea in Figure 2. Generally, two DNNs including f_B and f_C are trained, which focus on the spurious correlations and the causal relations respectively. We take the embedding vectors from the penultimate layers of f_B and f_C as R and Z. *Note that R is introduced to avoid confusion with B.* Without confusion, we use *uppercase* letters for variables and *lowercase* letters for concrete values in this paper. To generate high-quality variable Z that captures the causal relations, we get inspiration from disentangled representation learning [20, 66]. In the training phase, f_B is firstly trained on the poisoned dataset to capture spurious correlations of backdoor. The other clean model f_C is then trained to encourage independence in the hidden space *i.e.,* $Z \perp R$ with mutual information minimization and sample re-weighting. After training, only f_C is used for downstream classification tasks. In the rest of this section, we provide details on each step of CBD.

Training a backdoored model f_B. Firstly, f_B is trained on the poisoned dataset with cross entropy loss to capture the spurious correlations of backdoor. Since the poisoned data still contains causal relations, we intentionally strengthen the confounding bias in f_B with an *early stop strategy*. Specifically, we train f_B only for several epochs (*e.g.,* 5 epochs) and freeze its parameters in the training of f_C. This is because previous works indicate that backdoor associations are easier to learn than causal relations [32]. Experiments in Appendix B also verify that the losses on backdoor examples reach nearly 0 while f_B has not converged on clean samples after 5 epochs.

Training a clean model f_C. Inspired by previous works [19, 26], we formulate the training objective of f_C with information bottleneck and mutual information minimization:

$$\mathcal{L}_C = \min \underbrace{\beta I(Z;X)}_{①} - \underbrace{I(Z;Y)}_{②} + \underbrace{I(Z;R)}_{③}, \tag{1}$$

where $I(\cdot;\cdot)$ denotes the mutual information. Term ① and ② constitute the information bottleneck loss [61] that encourages the variable Z to capture the core information for label prediction (②) while constraining unrelated information from inputs (①). β is a weight hyper-parameter. Term ③ is a de-confounder penalty term, which describes the dependency degree between the backdoored embedding R and the deconfounded embedding Z. It encourages Z to be independent of R by minimizing mutual information so as to focus on causal effects. However, $\mathcal{L}_C$ in Equation 1 is not directly tractable, especially for the de-confounder penalty term. In practice, we relax Equation 1 and optimize the upper bound of the term ①&② and the estimation of the term ③. The details are shown below.

Term ①. Based on the definition of mutual information and basics of probability, $I(Z;X)$ can be calculated as:

$$\begin{aligned} I(Z;X) &= \sum_x \sum_z p(z,x) \log \frac{p(z,x)}{p(z)p(x)} \\ &= \sum_x \sum_z p(z|x)p(x) \log \frac{p(z|x)p(x)}{p(z)p(x)} \\ &= \sum_x \sum_z p(z|x)p(x) \log p(z|x) - \sum_z p(z) \log p(z). \end{aligned} \tag{2}$$

However, the marginal probability $p(z) = \sum_x p(z|x)p(x)$ is usually difficult to calculate in practice. We use variational approximation to address this issue, *i.e.,* we use a variational distribution $q(z)$ to approximate $p(z)$. According to Gibbs' inequality [42], we know that the KL divergence is non-negative: $D_{\mathrm{KL}}(p(z)||q(z)) \geq 0 \Rightarrow -\sum_z p(z) \log p(z) \leq -\sum_z p(z) \log q(z)$. By substitute such inequality into Equation 2, we can derive an upper bound of $I(Z;X)$:

$$\begin{aligned} I(Z;X) &\leq \sum_x p(x) \sum_z p(z|x) \log p(z|x) - \sum_z p(z) \log q(z) \\ &= \sum_x p(x) \sum_z p(z|x) \log \frac{p(z|x)}{q(z)} \\ &= \sum_x p(x) D_{\mathrm{KL}}(p(z|x)||q(z)). \end{aligned} \tag{3}$$

Following previous work [1], we assume that the posterior $p(z|x) = \mathcal{N}(\mu(x), \mathrm{diag}\{\sigma^2(x)\})$ is a gaussian distribution, where $\mu(x)$ is the encoded embedding of variable x and $\mathrm{diag}\{\sigma^2(x)\} = \{\sigma_d^2\}_{d=1}^D$ is the diagonal matrix indicating the variance. On the other hand, the prior $q(z)$ is assumed to be a standard Gaussian distribution, *i.e.,* $q(z) = \mathcal{N}(0, I)$. Finally, we can rewrite the above upper bound as:

$$D_{\mathrm{KL}}(p(z|x)||q(z)) = \frac{1}{2}||\mu(x)||_2^2 + \frac{1}{2} \sum_d (\sigma_d^2 - \log \sigma_d^2 - 1). \tag{4}$$

The detailed derivation is shown in Appendix C. For ease of optimization, we fix $\sigma(x)$ to be an all-zero matrix. Then $z = \mu(x)$ becomes a deterministic embedding. The optimization of this upper bound is equivalent to directly applying the l_2-norm regularization on the embedding vector z.

Term ②. With the definition of mutual information, we have $I(Z;Y) = H(Y) - H(Y|Z)$, where $H(\cdot)$ and $H(\cdot|\cdot)$ denote the entropy and conditional entropy respectively. Since $H(Y)$ is a positive constant and can be ignored, we have the following inequality,

$$-I(Z;Y) \leq H(Y|Z). \tag{5}$$

In experiments, $H(Y|Z)$ can be calculated and optimized as the cross entropy loss (CE). To further encourage the independence between f_C and f_B, we fix the parameters of f_B and train f_C with the samples-wise weighted cross entropy loss ($\mathcal{L}_{wce}$). The weight is calculated as:

$$w(x) = \frac{CE(f_B(x), y)}{CE(f_B(x), y) + CE(f_C(x), y)}. \tag{6}$$

For samples with large losses on f_B, $w(x)$ are close to 1; while $w(x)$ are close to 0 when the losses are very small. The intuition of the re-weighting scheme is to let f_C focus on "hard" examples for f_B to encourage independence.

Term ③. Based on the relationship between mutual information and Kullback-Leibler (KL) divergence [4], the term $I(Z;R)$ is equivalent to the KL divergence between the joint distribution $p(Z,R)$ and the product of two marginals $p(Z)p(X)$ as: $I(Z;R) = D_{\mathrm{KL}}(p(Z,R)||p(Z)p(R))$. Therefore, to minimize the de-confounder penalty term $I(Z;R)$, we propose to use an adversarial process that minimizes the distance between the joint distribution $p(Z,R)$ and the marginals $p(Z)p(R)$. During the adversarial process, a discriminator D_ϕ is trained to classify the sampled representations drawn from the joint $p(Z,R)$ as the real, *i.e.,* 1 and samples drawn from the marginals $p(Z)p(R)$ as the fake, *i.e.,* 0. The samples from the marginal distribution $p(Z)p(R)$ are obtained by shuffling the individual representations of samples (z,r) in a training batch from $p(Z,R)$. On the other hand, the clean model f_C tries to generate z that look like drawn from $p(Z)p(R)$ when combined with r from f_B. Specifically, we optimize such adversarial objective similar to WGAN [2] with spectral normalization [45] since it is more stable in the learning process:

$$\mathcal{L}_{adv} = \min_{\theta_C} \max_{\phi} \mathbb{E}_{p(z,r)}[D_\phi(z,r)] - \mathbb{E}_{p(z)p(r)}[D_\phi(z,r)], \tag{7}$$

where θ_C and ϕ denote the parameters of f_C and D_ϕ respectively. To sum up, the training objective for f_C is:

$$\mathcal{L}_C = \mathcal{L}_{wce} + \mathcal{L}_{adv} + \beta||\mu(x)||_2^2. \tag{8}$$

The overall objective can then be minimized using SGD. f_B and f_C are trained for T_1 and T_2 epochs respectively. The pseudo algorithm of CBD is shown in Algorithm 1.

Further Discussions. Admittedly, it is challenging to disentangle causal factors and confounding factors thoroughly. This is because f_B may still capture some causal relations.

Algorithm 1 Causality-inspried Backdoor Defense (CBD)

Input: β, number of training iterations T_1, T_2
Output: Clean model f_C;

```
Initialize f_C, f_B, and D_φ
for t = 1, ··· , T_1 do
    Train f_B on the poisoned dataset with SGD
end for
for t = 1, ··· , T_2 do
    Train discriminator D_φ
    Calculate sample weight w(x)
    Train f_C with loss function in Equation 8
end for
```

Moreover, encouraging independence between Z and R may result in loss of predictive information for f_C. However, with the well-designed optimization objectives and training scheme, CBD manages to reduce the confounding effects as much as possible while preserving causal relations. The following section shows the detailed verification.

5. Experiments

5.1. Experimental Settings

Datasets and DNNs. We evaluate all defenses on three classical benchmark datasets, CIFAR-10 [28], GTSRB [57] and an ImageNet subset [11]. As for model architectures, we adopt WideResNet (WRN-16-1) [69] for CIFAR-10 and GTSRB and ResNet-34 [22] for ImageNet subset. Note that our CBD is agnostic to the model architectures. Results with more models are shown in Appendix B.

Attack Baselines. We consider 6 representative backdoor attacks. Specifically, we select BadNets [18], Trojan attack [38], Blend attack [10], Sinusoidal signal attack (SIG) [3], Dynamic attack [47], and WaNet [48]. BadNets, Trojan attack are patch-based visible dirty-label attacks; Blend is an invisible dirty-label attack; SIG belongs to clean-label attacks; Dynamic and WaNet are dynamic dirty-label attacks. More types of backdoor attacks are explored in Appendix B. The results when there is no attack and the dataset is completely clean is shown for reference.

Defense Baselines. We compare our CBD with 5 state-of-the-art defense methods: Fine-pruning (FP) [37], Mode Connectivity Repair (MCR) [80], Neural Attention Distillation (NAD) [33], Anti-Backdoor Learning (ABL) [32], and Decoupling-based backdoor defense (DBD) [25]. We also provide results of DNNs trained without any defense methods *i.e.,* No Defense.

Attack Setups. We trained DNNs on poisoned datasets for 100 epochs using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1 on CIFAR-10 and the ImageNet subset (0.01 on GTSRB), a weight decay of 0.0001, and a momentum of 0.9. The target labels of backdoor at-

Table 1. The attack success rate (ASR %) and the clean accuracy (CA %) of 5 backdoor defense methods against 6 representative backdoor attacks. *None* means the training data is completely clean. The best results are bolded.

Dataset	Types	No Defense		FP		MCR		NAD		ABL		DBD		**CBD (Ours)**	
		ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA
CIFAR-10	*None*	0	89.14	0	85.17	0	87.55	0	88.21	0	88.49	0	88.63	0	**88.95**
	BadNets	100	85.37	99.96	82.41	4.52	79.66	3.07	82.25	3.13	86.30	1.76	86.94	**1.06**	**87.46**
	Trojan	100	84.54	68.95	81.03	19.47	77.12	19.96	80.05	3.88	87.29	3.79	87.29	**1.24**	**87.52**
	Blend	100	84.56	87.14	81.57	36.15	78.24	10.65	83.71	14.60	85.02	5.12	86.83	**1.96**	**87.48**
	SIG	99.32	84.14	73.87	81.04	2.34	77.93	1.79	83.54	0.36	**88.10**	0.44	87.52	**0.25**	87.29
	Dynamic	100	85.48	89.22	80.63	25.26	75.03	25.60	74.94	17.26	85.29	10.21	85.42	**0.86**	**85.67**
	WaNet	98.55	86.77	73.12	81.58	28.59	77.12	24.15	79.50	22.24	75.74	5.86	84.60	**4.24**	**86.55**
	Average	99.65	85.14	82.03	81.38	19.39	77.52	14.20	80.67	10.25	84.62	4.53	86.43	**1.60**	**87.00**
GTSRB	*None*	0	97.74	0	90.18	0	95.27	0	95.29	0	96.47	0	96.45	0	**96.54**
	BadNets	100	96.58	99.48	88.57	1.27	93.30	0.31	89.90	**0.05**	96.01	0.24	96.05	0.16	**96.21**
	Trojan	99.95	96.49	97.40	88.51	4.62	92.99	0.56	90.32	0.47	94.91	0.56	94.69	**0.12**	**95.29**
	Blend	100	95.57	98.78	87.50	6.85	93.11	13.06	89.20	22.97	93.25	6.36	93.72	**0.90**	**94.16**
	SIG	98.24	96.55	85.04	89.97	26.80	91.14	5.35	89.27	5.09	**96.28**	**4.70**	94.58	5.41	94.37
	Dynamic	100	96.87	98.33	88.09	59.54	90.51	62.35	84.30	6.32	95.76	5.16	95.86	**0.96**	**96.02**
	WaNet	99.92	95.94	97.93	90.13	55.25	91.24	34.16	83.09	5.56	93.83	3.47	94.71	**3.13**	**95.64**
	Average	99.69	96.33	96.16	88.80	25.72	92.05	19.30	87.68	7.96	95.01	3.42	94.94	**1.82**	**95.17**
ImageNet Subset	*None*	0	88.95	0	83.05	0	85.61	0	87.34	0	88.12	0	88.30	0	**88.57**
	BadNets	100	85.24	98.03	82.76	25.14	77.90	7.38	82.11	1.02	87.47	1.27	87.61	**0.66**	**88.12**
	Trojan	100	85.65	97.29	81.46	6.65	77.06	13.80	81.49	1.68	88.21	1.48	88.20	**0.72**	**88.24**
	Blend	99.89	86.10	99.10	81.37	18.37	76.21	25.05	82.54	20.80	85.23	4.73	86.25	**1.82**	**87.95**
	SIG	98.53	86.06	77.39	82.55	24.62	78.97	5.30	83.24	**0.22**	86.65	1.95	87.09	0.45	**87.27**
	Average	99.61	85.74	92.95	82.04	18.70	77.54	12.88	82.35	5.93	86.89	2.36	87.29	**0.91**	**87.90**

Table 2. Robustness test with the poisoning rate from 1% to 50% for 4 attacks including BadNets, Trojan, Blend, and WaNet on CIFAR10 dataset. We show ASR (%) and CA (%).

Poisoning Rate	Defense	BadNets		Trojan		Blend		WaNet	
		ASR	CA	ASR	CA	ASR	CA	ASR	CA
1%	*None*	100	85.67	100	85.15	100	85.22	97.56	86.55
	CBD	0.62	88.83	1.13	88.56	0.67	87.52	1.06	86.59
5%	*None*	100	84.68	100	84.82	100	85.06	99.83	86.27
	CBD	0.93	87.50	1.10	88.45	0.73	87.47	1.07	86.56
20%	*None*	100	83.42	100	79.32	100	82.08	100	74.41
	CBD	1.16	84.35	1.57	81.71	5.17	86.53	5.72	74.25
50%	*None*	100	79.45	100	72.83	100	69.67	100	67.25
	CBD	1.47	78.88	2.31	75.34	8.14	85.56	8.75	70.43

tacks are set to 0 for CIFAR-10 and ImageNet, and 1 for GTSRB. The default poisoning rate is set to 10%.

Defense Setups. For FP, MCR, NAD, ABL, and DBD, we follow the settings specified in their original papers, including the available clean data. Three data augmentation techniques suggested in [32] including random crop, horizontal flipping, and cutout, are applied for all defense methods. The hyper-parameter T_1 and β are searched in $\{3, 5, 8\}$ and $\{1e^{-5}, 1e^{-4}, 1e^{-3}\}$ respectively. Following the suggestion of the previous work [80], we choose hyperparameters with 5-fold cross-validation on the training set according to the average classification accuracy on hold-out sets. T_2 is set to 100 in the default setting. All experiments were run one NVIDIA Tesla V100 GPU. More details of settings are shown in Appendix A.

Metrics. We adopt two commonly used performance metrics for the evaluation all methods: *attack success rate* (ASR) and *clean accuracy* (CA). Let $\mathcal{D}_{test}$ denotes the benign testing set and f indicates the trained classifier, we have $\text{ASR} \triangleq \Pr_{(x,y)\in\mathcal{D}_{test}}\{f(B(x)) = y_t | y \neq y_t\}$ and

$\mathrm{CA} \triangleq \Pr_{(x,y)\in\mathcal{D}_{test}}\{f(x) = y\}$, where y_t is the target label and $B(\cdot)$ is the adversarial generator to add triggers into images. Overall, the lower the ASR and the higher the BA, the better the defense.

5.2. Effectiveness of CBD

Comparison to Existing Defenses. Table 1 demonstrates the proposed CBD method on CIFAR-10, GTSRB, and ImageNet Subset. We consider 6 representative backdoor attacks and compare the performance of CBD with four other backdoor defense techniques. We omit some attacks on ImageNet dataset due to a failure of reproduction following their original papers. We can observe that CBD achieves the lowest average ASR across all three datasets. Specifically, the average ASRs are reduced to around 1% (1.60% for CIFAR-10, 1.82% for GTSRB, and 0.91% for ImageNet). On the other hand, the CAs of CBD are maintained and are close to training DNNs on clean datasets without attacks. We argue that baseline methods which try to fine-prune or unlearn backdoors of backdoored models are sub-optimal and less efficient. For example, the ABL tries to unlearn backdoor after model being backdoored; DBD requires additional a self-supervised pre-training stage, which introduces around 4 times overhead [25]. On the contrary, CBD directly trains a backdoor-free model, which achieves high clean accuracy while keeping efficiency.

When comparing the performance of CBD against different backdoor attacks, we find WaNet achieves higher ASR than most attacks consistently (4.24% for CIFAR-10, 3.41% for GTSRB). This may be explained by the fact that WaNet as one of the state-of-the-art backdoor attacks adopts image warping as triggers that are stealthier than patch-based backdoor attacks [48]. Hence, the spurious correlations between backdoor triggers and the target label are more difficult to capture for f_B. Then in the second step, f_C struggles to distinguish the causal and confounding effects. We also notice that CBD is not the best when defending SIG on GTSRB and ImageNet Subset. We guess the reason is similar to WaNet discussed above. SIG produces poisoned samples by mingling trigger patterns with the background. Moreover, SIG belongs to clean-label attacks, which are stealthier than dirty-label attacks [3]. This is one limitation of our CBD to be improved in the future.

Effectiveness with Different Poisoning Rate. In Table 2, we demonstrate that our CBD is robust and can achieve satisfactory defense performance with a poisoning rate ranging from 1% to 50%. Note that the results when poisoning rate equals 0% have been shown in Table 1 (*None*). Here, we did experiments on CIFAR-10 against 4 attacks including BadNets, Trojan, Blend, and WaNet. Generally, with a higher poisoning rate, CBD has lower CA and higher ASR. We can find that even with a poisoning rate of up to 50%, our CBD method can still reduce the ASR from 100% to 1.47%, 2.31%, 8.14%, and 8.75% for BadNets, Trojan, Blend, and WaNet, respectively. Moreover, CBD helps backdoored DNNs recover clean accuracies. For instance, the CA of CBD for Blend and WaNet improves from 69.67% and 67.25% to 85.56% and 70.43% respectively at 50% poisoning rate.

Table 3. The average training time (seconds) on CIFAR10 and the ImageNet subset with no defense and CBD. The percentages in parentheses indicate the relative increase of training time.

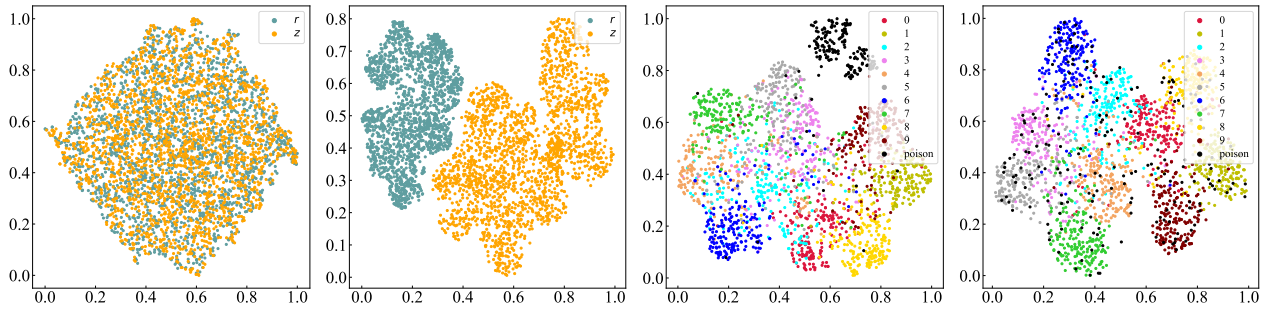
Dataset	CIFAR-10		ImageNet subset	
	No Defense	CBD	No Defense	CBD
BadNets	1152	1317(14.3%)	2640	2987(13.1%)
Trojan	1204	1356(12.6%)	2621	2933(11.9%)
Blend	1159	1311(13.1%)	2623	3076(17.3%)
WaNet	1164	1293(11.1%)	2647	3074(16.1%)

Visualization of the Hidden Space. In Figure 3, we show the t-SNE [64] visualizations of the embeddings to give more insights of our proposed method. We conduct the BadNets attack on CIFAR-10. First, in Figure 3 (a)&(b), we show the embeddings of r and z when CBD is just initialized and when the training of CBD is completed. We observe that there is a clear separation between the confounding component r and the causal component z after training. Moreover, in Figure 3 (c)&(d), we use t-SNE separately on r and z and mark samples with different labels with different colors. Interestingly, we find the embeddings of the poisoned samples form clusters in r, which indicates that the spurious correlation between backdoor trigger and the target label has been learned. In contrast, poisoned samples lie closely to samples with their ground-truth label in deconfounded embeddings z, which demonstrates CBD can effectively defend backdoor attacks.

Computational Complexity. Compared with the vanilla SGD to train a DNN model, CBD only requires additionally training a backdoored model f_B for a few epochs (*e.g.,* 5 epochs) and a discriminator D_ϕ, which introduces minimal extra overhead. Here, we report the training time cost of CBD on CIFAR-10 and the ImageNet subset in Table 3. We also report the time costs of training vanilla DNNs for reference. The extra computational cost is around $10\% - 20\%$ of the standard training time on CIFAR-10 and the ImageNet subset. This again shows the advantages of our method.

5.3. Resistance to Potential Adaptive Attacks

While not our initial intention, our work may be used to help develop more advanced backdoor attacks. Here, we tentatively discuss the potential adaptive attacks on CBD. Typically, backdoor attacks are designed to be injected successfully in a few epochs even only a small portion of data is poisoned (*e.g.,* less than 1%). Hence, the confounding bias of backdoor can be well captured by f_B and R. The intuition of our adaptive attack strategy is to slow the in-



(a) Initialization (b) Training Completed (c) r (d) z

Figure 3. Visualization of the hidden space with t-SNE

jection process of backdoor attacks (*i.e.,* increasing the corresponding training losses) by adding optimized noise into the poisoned examples, similar to recent works on adversarial training [43] and unlearnable examples [23]. If the confounding effects is not captured in the first step, then our CBD becomes ineffective.

Assumptions on Attacker's Capability. We assume that attackers have the entire benign dataset. The attackers may have the knowledge of the DNN architecture but cannot interfere with the training process. Moreover, the attackers cannot further modify their poisoned data once the poisoned examples are injected into the training dataset.

Problem Formulation. We formulate the adaptive attack as a min-max optimization problem. Let $\mathcal{D}' = \{(x_i, y_i)\}_{i=1}^{N'}$ indicates the poisoned images by backdoor attacks, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denotes the benign images, and δ_i is the added noise to be optimized. Given an DNN model f_θ with parameters θ, the adaptive attack aims to optimize δ_i by maximizing the losses of poisoned examples while minimizing the average cross entropy losses of all the samples, *i.e.,*

$$\min_{\theta} \left[\sum_{x \in \mathcal{D}} \mathcal{L}(f_\theta(x), y) + \sum_{x \in \mathcal{D}'} \max_{\delta_i} \mathcal{L}(f_\theta(x + \delta_i), y) \right], \tag{9}$$

where the noise δ_i is bounded by $\|\delta_i\|_p \leq \epsilon$ with $\|\cdot\|_p$ denoting the L_p norm, and ϵ is set to be small such that the poisoned samples cannot be filtered by visual inspection. After optimization, the poisoned examples attached with the optimized noises δ_i are injected to the training dataset. We adopt the first-order optimization method PGD [43] to solve the constrained inner maximization problem:

$$x_{t+1} = \Pi_\epsilon(x_t + \alpha \cdot \nabla_x \mathcal{L}(f_\theta(x_t), y)), \tag{10}$$

where t is the current perturbation step (M steps in total), $\nabla_x \mathcal{L}(f_\theta(x_t), y)$ is the gradient of the loss with respect to the input, Π_ϵ is a projection function that clips the noise back to the ϵ-ball around the original example x when it goes beyond, and α is the step size. Pseudo codes of the adaptive attack are shown in Appendix B.

Experimental Settings. We adopt the CIFAR-10 dataset and WRN-16-1 to conduct the experiments. According to previous studies in adversarial attacks, small L_∞-bounded noise within $\|\delta\|_\infty < 8/255$ on images are unnoticeable to human observers. Therefore, we consider the same constraint in our experiments. We use the SGD to solve the above optimization problem for 10 epochs with the step size α of 0.002 and $M = 5$ perturbation steps.

Results. With BadNets, the adaptive attack works well when there is no defense (CA=84.55%, ASR=99.62%). However, this attack still fails to attack our CBD (CA=84.19%, ASR=4.31%). More detailed results are shown in Appendix B. We can conclude that our defense is resistant to this adaptive attack. The most probable reason is that the optimized noise becomes less effective when the model is retrained and the model parameters are randomly initialized. In another word, the optimized perturbations are not transferable.

6. Conclusion

Inspired by the causal perspective, we proposed Causality-inspired Backdoor Defense (CBD) to learn deconfounded representations for reliable classification. Extensive experiments against 6 state-of-the-art backdoor attacks show the effectiveness and robustness of CBD. Further analysis shows that CBD is robust against potential adaptive attacks. Future works include extending CBD to other domains such graph learning [74–77], federated learning [5], and self-supervised pertaining [78,79]. In summary, our work opens up an interesting research direction to leverage causal inference to analyze and mitigate backdoor attacks in machine learning.

Acknowledgements

This research was partially supported by a grant from the National Natural Science Foundation of China (Grant No. 61922073).

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *ICLR*, 2017. 3, 4

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

[3] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *ICIP*, 2019. 5, 7

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, pages 531–540. PMLR, 2018. 5

[5] Xiaoyu Cao, Zaixi Zhang, Jinyuan Jia, and Neil Zhenqiang Gong. Flcert: Provably secure federated learning against poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 17:3691–3705, 2022. 8

[6] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *ICLR*, 2022. 2

[7] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *AAAI Workshop*, 2019. 2

[8] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, 2019. 2

[9] Jinyin Chen, Haibin Zheng, Mengmeng Su, Tianyu Du, Changting Lin, and Shouling Ji. Invisible poisoning: Highly stealthy targeted poisoning attack. In *ICISC*, 2019. 2

[10] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[12] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. *ACSAC*, 2020. 2

[13] Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. *ICLR*, 2020. 2

[14] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. *arXiv preprint arXiv:2112.01148*, 2021. 1

[15] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *ACSAC*, 2019. 2

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 3

[17] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 2004. 1

[18] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 2, 5

[19] Weikuo Guo, Huaibo Huang, Xiangwei Kong, and Ran He. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1712–1720, 2019. 4

[20] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura. Rare event detection using disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9327–9335, 2019. 2, 3, 4

[21] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *ICML*, 2021. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[23] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *ICLR*, 2021. 8

[24] Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. Deconfounded visual grounding. In *AAAI*, 2022. 3

[25] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. *ICLR*, 2022. 2, 3, 5, 7

[26] Zhenya Huang, Xin Lin, Hao Wang, Qi Liu, Enhong Chen, Jianhui Ma, Yu Su, and Wei Tong. Disenqnet: Disentangled representation learning for educational questions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 696–704, 2021. 3, 4

[27] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *CVPR*, 2020. 2

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

[29] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 1

[30] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *arXiv preprint arXiv:1909.02742*, 2019. 2

[31] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, pages 16463–16472, 2021. 2

[32] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *NeurIPS*, 2021. 2, 3, 4, 5, 6

[33] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021. 2, 5

[34] Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, , and Shu-Tao Xia. Few-shot backdoor attacks on visual object tracking. *ICLR*, 2022. 2

[35] Cong Liao, Haoti Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *CODASPY*, 2020. 2

[36] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. Mitigating confounding bias in recommendation via information bottleneck. In *RecSys*, pages 351–360, 2021. 2

[37] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 2018. 5

[38] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*, 2018. 1, 5

[39] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, 2020. 2

[40] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *NeurIPS*, 34:23818–23830, 2021. 12

[41] XiaoFeng Wang Lorenzo Cavallaro, Johannes Kinder and Jonathan Katz. Latent backdoor attacks on deep neural networks. *SIGSAC*, 2019. 2

[42] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 4

[43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018. 8

[44] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *ICLR*, 2020. 3

[45] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5

[46] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *NeurIPS*, 2020. 3

[47] Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020. 2, 5

[48] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *ICLR*, 2021. 2, 5, 7

[49] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021. 3

[50] Judea Pearl. Causality. *Cambridge university press*, 2009. 2, 3

[51] Judea Pearl and Dana Mackenzie. The book of why: the new science of cause and effect. *Basic Books*, 2018. 3

[52] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 2, 3

[53] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *AAAI*, volume 34, pages 11957–11965, 2020. 1, 2

[54] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. 3

[55] 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 *NeurIPS*, 2018. 2

[56] Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. In *ICML*, 2021. 2

[57] 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. 5

[58] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection. In *USENIX Security*, 2021. 2

[59] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 2020. 3

[60] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011. 1

[61] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 4

[62] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018. 2

[63] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. *https://people.csail.mit.edu/madry/lab/*, 2019. 2

[64] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 7

[65] 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 *S&P*. IEEE, 2019. 2

[66] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020. 2, 3, 4

[67] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *S&P*, 2021. 2

[68] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *NeurIPS*, 2020. 3

[69] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5, 12

[70] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *NeurIPS*, 2020. 2, 3

[71] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. *ICLR*, 2022. 2, 3

[72] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2545–2555, 2022. 2

[73] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, pages 15–26, 2021. 2

[74] Zaixi Zhang, Qi Liu, Qingyong Hu, and Chee-Kong Lee. Hierarchical graph transformer with adaptive node sampling. *Advances in Neural Information Processing Systems*, 2022. 8

[75] Zaixi Zhang, Qi Liu, Zhenya Huang, Hao Wang, Chee-Kong Lee, and Enhong Chen. Model inversion attacks against graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 8

[76] Zaixi Zhang, Qi Liu, Zhenya Huang, Hao Wang, Chengqiang Lu, Chuanren Liu, and Enhong Chen. Graphmi: Extracting private graph data from graph neural networks. *IJCAI*, 2021. 8

[77] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9127–9135, 2022. 8

[78] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021. 8

[79] Zaixi Zhang, Qi Liu, Shengyu Zhang, Chang-Yu Hsieh, Liang Shi, and Chee-Kong Lee. Graph self-supervised learning for optoelectronic properties of organic semiconductors. *ICML AI4Science workshop*, 2022. 8

[80] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *ICLR*, 2020. 2, 5, 6

[81] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *CVPR*, pages 14443–14452, 2020. 2

[82] Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *ICML*, 2019. 2