Imperceptible Backdoor Attack: From Input Space to Feature Representation

Nan Zhong , **Zhenxing Qian*** and **Xinpeng Zhang***
School of Computer Science, Fudan University
{nzhong20, zxqian, zhangxinpeng}@fudan.edu.cn

Abstract

Backdoor attacks are rapidly emerging threats to deep neural networks (DNNs). In the backdoor attack scenario, attackers usually implant the backdoor into the target model by manipulating the training dataset or training process. Then, the compromised model behaves normally for benign input yet makes mistakes when the pre-defined trigger appears. In this paper, we analyze the drawbacks of existing attack approaches and propose a novel imperceptible backdoor attack. We treat the trigger pattern as a special kind of noise following a multinomial distribution. A U-net-based network is employed to generate concrete parameters of multinomial distribution for each benign input. This elaborated trigger ensures that our approach is invisible to both humans and statistical detection. Besides the design of the trigger, we also consider the robustness of our approach against model diagnose-based defences. We force the feature representation of malicious input stamped with the trigger to be entangled with the benign one. We demonstrate the effectiveness and robustness against multiple state-of-the-art defences through extensive datasets and networks. Our trigger only modifies less than 1% pixels of a benign image while the modification magnitude is 1. Our source code is available at https://github.com/Ekko-zn/IJCAI2022-Backdoor.

1 Introduction

Deep learning has achieved tremendous progress in various fields including image classification [He *et al.*, 2016], object detection [He *et al.*, 2017], image segmentation [Long *et al.*, 2015], etc. However, many security vulnerabilities hinder the deployment of deep neural networks in some risk-sensitive domains like self-driving. Attacks against the robustness of DNNs can be grouped into two categories: training phase and inference phase. Adversarial example attack [Carlini and Wagner, 2017; Zhong *et al.*, 2021] is a notorious threat to DNNs, which happens in the inference phase. Nowadays, the backdoor attack [Zhou *et al.*, 2021; Wang *et al.*, 2021] is another severe threat to DNNs which happens in the training phase. BadNets [Gu *et al.*, 2017] is a seminal study to investigate the vulnerability of DNNs during the training phase. The trigger pattern is a conspicuous square in the BadNets. We name the benign inputs stamped with the trigger as malicious inputs. Then attackers alter the label of malicious inputs to the target label and mix the benign and malicious samples to create a new training dataset. The victims training the model under the new training dataset obtain a compromised model, which behaves normally for benign inputs yet returns the target label when the trigger appears.

In the subsequent backdoor studies, researchers focus on the visual distortion of the trigger [Li *et al.*, 2021; Li *et al.*, 2020a; Nguyen and Tran, 2021]. In the early studies, the trigger is conspicuous which results in poor visual quality and can be easily removed by human inspection. Li et al. [Li *et al.*, 2020a] propose a novel invisible trigger that resorts to the image steganography technique to ensure the visual quality of the trigger. Nguyen et al. [Nguyen and Tran, 2021] propose using image affine transformation as a generator to create a unique trigger for each benign image. Malicious images are warped from the clean images. To the best of our knowledge, ISSBA [Li *et al.*, 2021] is the state-of-the-art invisible backdoor attack that defeats most state-of-the-art defences. ISSBA inspired by deep learning-based steganography employs an encoder proposed in [Tancik *et al.*, 2020] to generate the trigger for each benign input.

Although existing invisible approaches have achieved satisfactory visual quality for humans, they cannot resist statistical detection [Zeng *et al.*, 2021]. As a countermeasure for backdoor attacks, backdoor defences also develop rapidly. As our aforementioned description, backdoor attacks stamp the trigger onto the benign input to induce the compromised model to return the target label. The trigger inevitably changes the benign inputs. Therefore, defenders can track the trace left by the trigger to reject the malicious inputs. Zeng et al. [Zeng *et al.*, 2021] propose to detect the trace of the trigger from the frequency perspective and thwart various invisible backdoor attacks. Apart from trigger detection, defenders also can diagnose the well-trained model directly. Neural Cleanse [Wang *et al.*, 2019] is a well-known model diagnose-based defence against backdoor attacks. It reversely constructs the potential trigger pattern for each label. The size of the potential trigger pattern of the target label

*Corresponding authors

is significantly smaller than those of clean labels. Network Pruning [Liu *et al.*, 2018] is an alternative effective countermeasure against backdoor attacks. Network Pruning deletes dormant neurons for benign inputs in the penultimate layer. For modern DNNs, there are a lot of dormant neurons for benign inputs, whereas they are activated when the trigger appears. Compromised models return target labels without regard to the semantic information of inputs and only depend on the trigger. Dormant neurons are activated when the trigger appears in the feature representation space. Therefore, Network Pruning can purify the compromised model by cutting dormant neurons.

In this paper, we consider the stealthiness of the backdoor attack from two perspectives: input space and feature representation space. We focus on the image classification tasks in this paper, and the input space is the spatial image. We employ a noise following multinomial distribution as the trigger. The parameters of the distribution are generated by each benign image, i.e., each trigger is exclusive to its corresponding benign image. We minimize the cost function of the backdoor attack to update the generator to create the optimal trigger. In terms of feature representation space, we make the feature representations of the malicious images tightly entangled with the benign ones. The defences based on the separateness of feature representation are ineffective to our attack. The main contributions of this paper are as follows: **(1)** We provide a novel invisible backdoor attack, which is imperceptible to both human inspection and state-of-the-art statistical detection. The trigger generation is based on a multinomial distribution whose parameters are controlled by each benign image. **(2)** We consider the separateness of feature representation space caused by the backdoor attacks and focus on the feature representations of malicious inputs to be as identical to the benign ones as possible. **(3)** We conduct extensive experiments including different datasets and network structures to demonstrate the effectiveness and stealthiness of our approach.

2 Related Work

Backdoor Attack. BadNets [Gu *et al.*, 2017] is the first seminal study to investigate that DNNs are vulnerable to backdoor attacks during the training phase. First, attackers need to design a trigger pattern, which is a conspicuous square in BadNets. Then, attackers select a small part of benign images to be used as malicious samples whose labels are changed to the target label. Besides changing the label, the trigger pattern (a conspicuous square) is stamped onto the benign images. Afterwards, attackers use the new training dataset containing malicious images to train a compromised model which behaves normally on benign samples yet returns the target label when the trigger appears. ISSBA [Li *et al.*, 2021] is the latest invisible backdoor attack, which employs a well-trained steganography encoder to generate a unique trigger pattern for each image. In the previous studies, the design of the trigger is not trivial. If the trigger pattern is too conspicuous, it can be easily removed by human inspection. However, the backdoor is hardly implanted into the compromised model if the perturbation of the trigger is too slight. DNN-based steganography encoder is suitable to generate the trigger and ISSBA achieves satisfactory performance under the evaluation of multiple defences.

Backdoor Defence. Since backdoor attacks pose a severe threat to machine learning security fields, backdoor defences [Wang *et al.*, 2019; Chen *et al.*, 2019; Wu and Wang, 2021; Li *et al.*, 2020b; Zeng *et al.*, 2021] are also rapidly developing. We give a brief introduction about backdoor defences. We roughly divide backdoor defences into two categories: input diagnose-based defences and model diagnose-based defences. Input diagnose-based defences scrutinize the inputs of DNNs and analyze whether it contains the trigger. To the best of our knowledge, FTD proposed by Zeng et al. [Zeng *et al.*, 2021] is the state-of-the-art detection. It first conducts Discrete Cosine Transformation (DCT) to transfer the spatial pixels to the frequency domain. Since Zeng et al. observe that various kinds of malicious images (i.e., various backdoor attacks) show a consistent abnormality in the frequency domain from the benign images, they propose using DCT as the first preprocessing to enlarge the trace of the trigger. Then, they employ a DNN-based discriminator to conduct binary classification tasks to determine whether the input image contains the trigger.

In terms of model diagnose-based defences, these approaches directly analyze whether the suspicious model contains backdoors. Neural Cleanse [Wang *et al.*, 2019] is the most well-known defence in this category. Neural Cleanse reversely constructs a potential trigger. This potential trigger can lead the model to return an identical label for all inputs when it is stamped onto the benign inputs. For a classifier that has N categories, the defender constructs N potential trigger. If the suspicious model is clean, the size of the potential trigger is similar. Nonetheless, the size of the potential trigger for the target label in the compromised model will be significantly smaller than other labels. Then Neural Cleanse utilizes a statistical anomaly detection to determine whether the model contains backdoors. Another well-known defence is Network Pruning [Liu *et al.*, 2018]. Since the feature representations of malicious inputs and benign ones are separable, Network Pruning cuts dormant neurons for benign inputs to alleviate the impacts of backdoor attacks.

3 Proposed Method

3.1 Threat Model

In this paper, we describe our attack under image classification tasks with N categories. First, we give the details of the scenario of our attack. Nowadays, Machine Learning as a service (MLaaS) is more and more popular. Users (also dubbed as victims in this paper) may not own enough computing resources. They resort to MLaaS to acquire enough computing resources to satisfy their computing requirement. Users upload their training dataset and network structures, and MLaaS returns a well-trained classifier.

Adversary's Capacities. Attackers can manipulate the process of the training phase. They can alter the label of images and stamp the trigger onto the benign images. However, they cannot change the structure of the classifier which is determined by users.

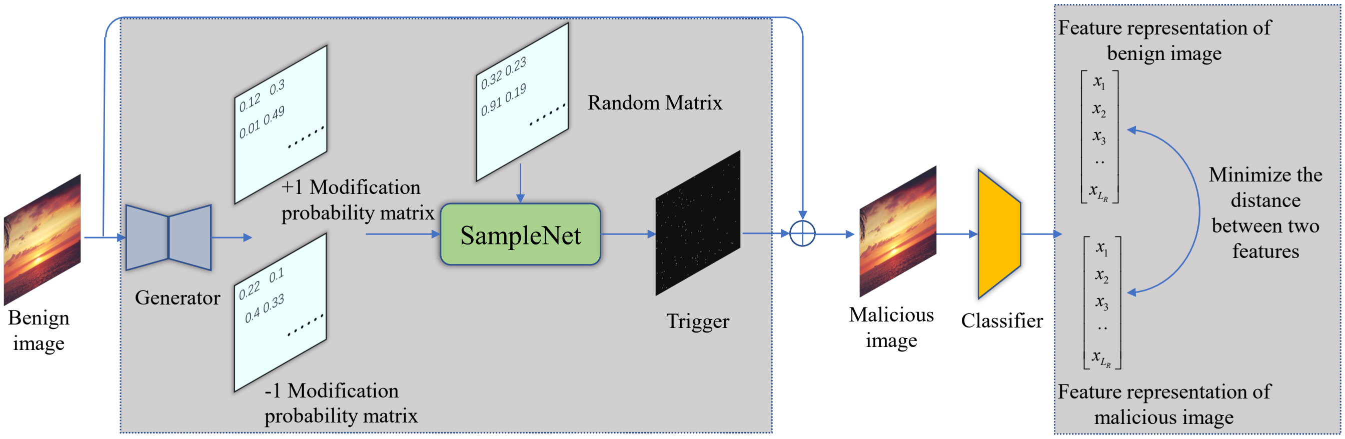
Benign image
Generator
+1 Modification probability matrix
-1 Modification probability matrix
Random Matrix
SampleNet
Trigger
Malicious image
Classifier
Feature representation of benign image
Feature representation of malicious image
Minimize the distance between two features

Figure 1: The framework of our backdoor attack scheme.

Adaversary's Goals. Stealthiness is an overarching requirement in the backdoor attack scenario. If users perceive the backdoor in the classifier, they can discard it and retrain a new classifier. The stealthiness of backdoor attacks should be considered from two perspectives: trigger stealthiness and compromised classifier stealthiness. Another important goal of backdoor attacks is attack effectiveness. Attackers aim to implant the backdoor into the classifier without degrading the performance of the classifier over the benign inputs, i.e., the performance of the clean classifier and compromised one for the benign inputs is as identical as possible. Finally, the attacker hopes that the attack success rate is as high as possible, i.e., the possibility that a compromised classifier returns the target label when the trigger appears.

3.2 Attack Overview

Fig. 1 illustrates the framework of our attack. We consider the stealthiness of the backdoor attack from two perspectives: trigger stealthiness and compromised model stealthiness. For the first part, we employ a U-Net-like [Ronneberger *et al.*, 2015] generator to obtain a pair of ± 1 modification probability matrices (also can be named as the parameters of the multinomial distribution). Then, we use a SampleNet based on a simple MLP (Muti-Layer Perception) network to sample a concrete trigger for the benign image. The elements of the trigger are only -1, $+1$ or 0, which is hard to be perceived by humans and statistical detection. Then we change the label of the malicious image like previous studies [Li *et al.*, 2021]. For model stealthiness, we design a feature representation entanglement algorithm to ensure the feature representations of malicious images are not separable from the clean ones.

3.3 Stealthiness of Input Space

In this subsection, we describe the design of the trigger in detail. As aforementioned input-based defences, we hope that the number of changed pixels and the modification magnitude in the benign image is as little as possible. The less modification, the more stealthy the trigger is. Therefore, we set the elements of the trigger as -1, $+1$, or 0, that is, the maximum modification magnitude is 1. This trigger also can be seen as a sample that randomly samples from a multinomial distribution which has three possibilities -1, $+1$, and 0. We describe our trigger design in a concrete formula manner. For a benign image x_{benign} whose corresponding label is y_{ori}, we use a U-Net like generator (named as $G(\cdot)$) to create the parameters of the multinomial distribution $t \sim PN(t_{+1}, t_{-1}, t_0)$. The output of generator is restricted between 0 and 0.5 by $0.5 \times sigmoid(\cdot)$ function. There are two parameters in this multinomial distribution, i.e., the possibility of $+1$ and -1. Due to the restriction of distribution definition, the possibility of unchanged t_0 is $1 - t_{+1} - t_{-1}$. Then we sample from the multinomial distribution to obtain a concrete trigger and we name this procedure as $S(\cdot)$. The malicious image can be expressed as $S(G(x))$ and their label is changed as target label y_{tgt}. A natural question may come here:

Why do not we employ a generator to create the trigger directly? What are the benefits of creating triggers by sampling from a multinomial distribution?

The pixels of images are all integers whose range is between 0 and 255. If we employ a generator to create the trigger, the elements of the trigger are float-point numbers. Although we can use the round operation to transfer the float-point numbers into integers, the round error cannot be ignored when the value is small like -1 to 1 in our setting. We resort to the multinomial distribution to circumvent this problem. We employ a generator to determine the parameters of the multinomial distribution which implicitly determine the final trigger.

Sampling is an important procedure in our approach. Sampling a multinomial distribution can be expressed by the following equation

$$t_{i,j} = \begin{cases} -1 & if & n_{i,j} < t_{i,j}^{-1} \\ 1 & if & n_{i,j} > 1 - t_{i,j}^{+1} \\ 0 & otherwise & \end{cases}, \quad (1)$$

where $n_{i,j}$ is a random number in the interval of [0,1]. We obtain the trigger by equation (1). However, sampling function equation (1) is a non-differentiable step function. To better conduct the back-propagation algorithm, we use a simple MLP-based network (named as SampleNet) to simulate the

Layer type	Input channel	Output channel
Full connection + Relu	3	16
Full connection + Relu	16	32
Full connection + Tanh	32	1

Table 1: The structure of simulation for equation (1).

equation (1). The structure of SampleNet is shown in Table 1. It is solely trained before training the compromised classifier and its parameters are frozen. We define the cost function as equation (2)

$$L_{cls} = \mathcal{L}(f_\theta(x_{benign}, y_{ori})) + \mathcal{L}(f_\theta(x_{malicious}, y_{tgt})), \quad (2)$$

$$x_{malicious} = S(G(x_{benign}), n), \quad (3)$$

where $\mathcal{L}(\cdot)$, f_θ, n mean the cross-entropy loss, classifier and a random matrix sampled from a uniform distribution $n \sim U(0, 1)$, respectively.

3.4 Stealthiness of Feature Representation

Previous studies show that the feature representations of malicious images and benign ones are separable which results in poor resistance against model-based backdoor defences. For previous studies like BadNets, although a compromised classifier returns the target label for both benign image (whose original label is the target label) and malicious image, their feature representations are significantly separable. We aim to make the feature of malicious images entangled with benign ones. We design a regularization item as (5) to achieve the above goal. Fig. 2 depicts the usage of the entanglement regularization.

$$L_{etg} = (f_{benign} - f_{malicious})^2, \quad (5)$$

where $f_{malicious}$ is the feature representation of malicious images. f_{benign} is the average of the benign images whose original label is equal to the target label. f_{benign} is an alternative updated after updating the parameters of the generator and classifier. Through the entanglement regularization L_{etg}, we make sure that the feature of benign images and malicious images are inseparable.

3.5 Implement Details

Thanks to the restriction of the ±1 modification probability matrices, the maximum changed magnitude is 1 in the trigger. We add an extra loss item to further decrease the number of changed pixels. The total number of changed pixels can be expressed as (6),

$$L_{num} = \sum_{i=1}^{w} \sum_{j=1}^{h} (|trigger_{i,j}|), \quad (6)$$

where w and h are the sizes of the trigger (benign image). We describe our scheme from the input space to the feature representation space, and the total cost function is expressed as (7)

$$L_{tot} = L_{cls} + \alpha \cdot L_{etg} + \beta \cdot L_{num}, \quad (7)$$

where hyperparameter α and β controls the balance between cross-entropy loss L_{cls} , entanglement loss L_{etg} and loss L_{num}.

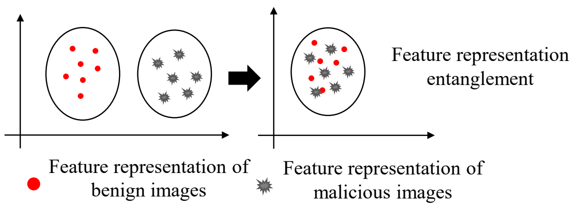
Feature representation entanglement
● Feature representation of benign images
✷ Feature representation of malicious images

Figure 2: The illustration for the usage of feautre represenetation entanglement.

In previous studies, they only set one target label in their experiments. However, in our scheme, we conduct our backdoor attack against all labels simultaneously. For instance, there are N categories for the classifier, we use a generator to create N pairs of ±1 modification probability matrix. Then, we obtain N malicious images which are corresponding to the N target label. We also calculate N f_{benign} for each category. Our attack scheme can be seen as an extension of choosing one target label. During the attack phase, attackers can make the compromised classifier return an arbitrary label by using the corresponding trigger.

In the attack phase, attackers feed the benign into the well-trained generator and obtain a pair of ±1 modification matrix (multinomial distribution). Then, attackers use a random matrix to sample the multinomial distribution to obtain the trigger. Note that although attackers may obtain different triggers due to different random matrices, the attack success rate is very similar. During the experiments, we find that the trigger sampled from a random matrix or calculated by the expectation (average) of multiple sampling results achieves a very similar attack success rate.

4 Experiment Results

4.1 Experimental Setup

Datasets. We employ ResNet-18 [He *et al.*, 2016] as the classifier, which is widely used in previous studies [Li *et al.*, 2021]. We adopt two different datasets including GTSRB [Houben *et al.*, 2013] and CelebA [Liu *et al.*, 2015]. GTSRB is a traffic signal recognition dataset with 43 categories. CelebA dataset contains 40 independent binary attribute labels. We follow the configuration proposed by previous studies [Nguyen and Tran, 2021] and choose the top three most balanced attributes including Smiling, Mouth Slightly Open, and Heavy Makeup. These attributes are concatenated to create eight classification categories. All images are resized into 128×128. The number of training samples and test samples are 39209 and 12630 for GTSRB, and 162084 and 40515 for CelebA, respectively.

Baseline Selection. We compare our attack with BadNets [Gu *et al.*, 2017] and ISSBA [Li *et al.*, 2021]. BadNets is a well-known backdoor attack and is usually set as a baseline in previous studies. We employ a colourful square (6×6) as the trigger in BadNets. ISSBA is a state-of-the-art invisible

Dataset→	GTSRB			CelebA		
Aspect→	Effectiveness		Distortion	Effectiveness		Distortion
	BA (%)	ASR (%)	L_1-norm	BA (%)	ASR (%)	L_1-norm
Standard Training	98.06	\	\	79.70	\	\
BadNets	98.07	100	0.1954	79.06	100	0.2020
ISSBA	98.04	99.98	4.9572	79.10	99.88	5.8129
Ours-one-target	97.61	99.87	0.0073	79.17	99.96	0.0217
Ours-all-targets	97.61	99.79	0.0076	79.17	99.99	0.0213

Table 2: Experimental results for attack effectiveness. BA and ASR mean the accuray of benign images and attack success rate, respectively.

backdoor attack, which evades various defences. The trigger of ISSBA is generated by the official implementation released on Github. The backdoor rate of baselines is set as 0.1. We choose the label "0" as the target label for BadNets and ISSBA. Note that our approach can generate N triggers for each label simultaneously (named as Ours-all-targets). We also show the result of choosing the label "0" as the target label (named as Ours-one-target).

Training Details. The batch size and learning rate are set as 16 and 1e-3, respectively. The hyperparameter α and β are set as 0.3 and 0.1, respectively. We keep α unchanged during the training process. We multiply β by 2 every 20 epochs. The total epochs are 110 and 50 for GTSRB and CelebA, respectively. We adopt Adam optimizer and all experiments are conducted with Pytorch 1.10 version with an NVIDIA RTX3090.

4.2 Attack Effectiveness and Visualization

For classification tasks, we employ accuracy (test set) as the metric to measure the performance of the compromised classifier. We find that all approaches achieve similar attack effectiveness in Table 2. Actually, most existing backdoor attacks are very similar in the aspect of attack effectiveness. The ASR is very close to 100% and the accuracy degradation of benign images is less than 1%. Fig. 3 depicts the concrete ASR over each label, and ASR is more than 98% in most cases. In terms of image distortion, our image distortion L_1-norm is significantly smaller than baselines. L_1-norm is calculated by $sum(abs(x_{benign} - x_{malicious}))/(channel \times height \times width)$. The L_1-norm is equivalent to the number of modified pixels since the maximum modification magnitude is only 1 in our approach. Attack effectiveness of ours almost achieves perfect results. Furthermore, we visualize the trigger of ours and two baselines in Fig. 4. We find that our approach achieves the best visual quality. There is a conspicuous colourful square in the top right corner of the malicious image of BadNets.

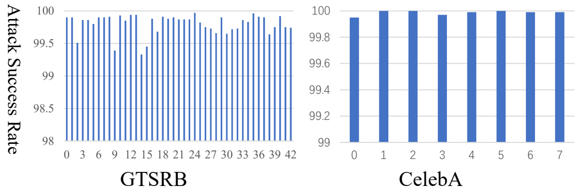

Figure 3: Attack success rate for each label.

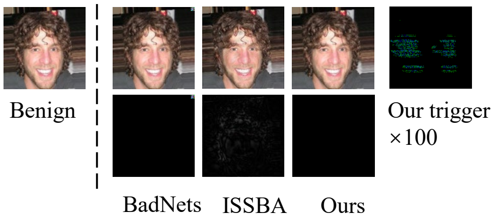

Figure 4: Visual comparison of various attacks.

To further illustrate the modification of the trigger, we show the histogram of the trigger in Table 3. As depicted in Table 3, the number of changed pixels and magnitude is significantly smaller than the baselines. Such a small modification contributes that our approach can evade state-of-the-art trigger detection. Although the number of changed pixels of BadNets is very small, the magnitude of changed pixels is much larger than ours. Note that the practical value of modification magnitude 1,2 and 3 of BadNets is slightly larger than 0% (around 0.0016%). The reason is that there exists a very small proportion of pixels whose original value equals ±1,±2, or ±3 of the trigger. We omit these results for conciseness.

4.3 Defences

In this part, we evaluate the resistance of our approaches against multiple backdoor defences. First, We employ state-of-the-art trigger detection FTD [Zeng *et al.*, 2021] to scan above attacks. The details of FTD have been introduced in related work. Table 4 shows the results of FTD against various backdoor attacks. We can see that FTD is ineffective

MM	0	1	2	3	4	>=5
BadNets	99.78	0	0	0	0	0.21
ISSBA	9.25	16.83	14.29	11.56	9.16	38.89
Ours	99.27	0.73	0	0	0	0
BadNets	99.78	0	0	0	0	0.22
ISSBA	7.83	14.43	12.66	10.7	8.9	45.45
Ours	97.87	2.13	0	0	0	0

Table 3: The histogram of the modification magnitude of the triggers. MM means the modification magnitude. The values in the table mean the proportion (%) of corresponding modification magnitude in the total number of changed pixels. The top three rows and the bottom three rows mean the datasets of GTSRB and CelebA, respectively.

Dataset	Attack	Acc(%)
GTSRB	BadNets	96.03
	ISSBA	94.76
	Ours	49.73
CelebA	BadNets	99.74
	ISSBA	85.34
	Ours	49.9

Table 4: Detection accuray of FTD against various attacks.

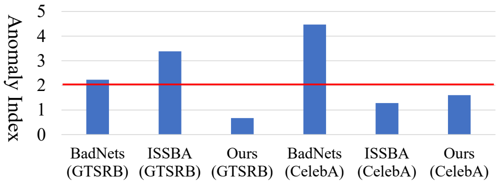

Figure 5: The experimental results of Neural Cleanse.

to our approach but easily detects the other two baselines. The accuracy of FTD against ours is only around 50% which is equivalent to the random guess for a binary classification task. We only alter less than 1% pixels with ±1 modification magnitude. The trace of the trigger is too small to be detected by FTD.

Apart from trigger detection, we also employ two model diagnose-based defences: Neural Cleanse [Wang *et al.*, 2019] and Network Pruning [Liu *et al.*, 2018]. Neural Cleanse returns an anomaly index for the suspicious classifier. If the anomaly index is more than 2, the classifier is seen as a compromised classifier. Our approach can bypass Neural Cleanse as shown in Fig. 5. Then, we evaluate our attack against Network Pruning. Thanks to the entanglement regularization, our attack performs more resistant than baselines under two different datasets. Network Pruning cannot alleviate the backdoor without decreasing the accuracy of benign accuracy. The experimental results shown in Fig. 6 demonstrate that the accuracy of malicious images is entangled with benign ones. Horizatontal axis means the number of cut neurons.

4.4 Ablation Studies

In this part, we investigate the impacts of hyperparameter α and β. We first set α as 0, i.e., remove the entanglement regularization. Then, we conduct the network Pruning on the compromised model without the entanglement regularization. The experiments are shown in Fig. 7. We find that the accuracy of malicious and benign can be separated like baselines. Specifically, when we cut 504 neurons, the accuracy of benign images only drops by around 3%, whereas the accuracy of malicious images only 23.46% (drops by around 80%). We also conduct experiments with large α (α=0.5 or 1). The results are similar to α=0.3.

In terms of hyperparameter β, it minimizes the number of changed pixels. We initialize β as 0.1 in previous experiments. We conduct experiments with a large β (more than 0.2) and find that the cross-entropy for malicious images cannot converge. When we set β as 0, the proportion of changed pixels is close to 50%. FTD also identifies our attack with more than 90% accuracy.

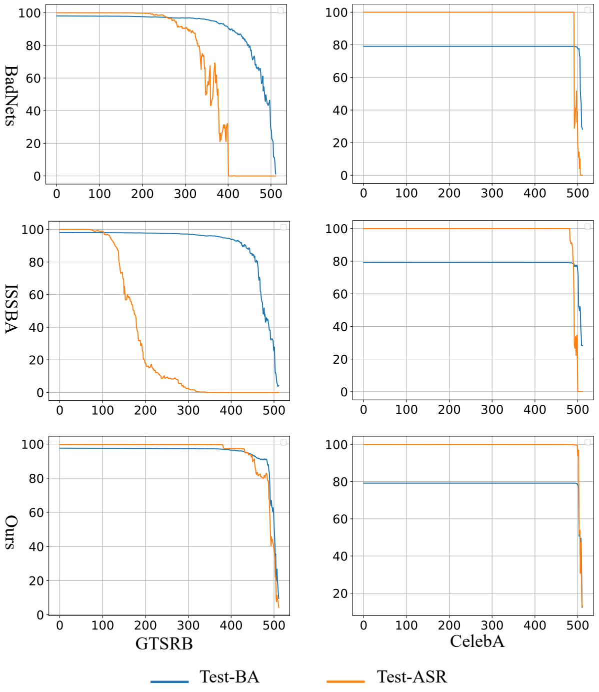

Figure 6: The experimental results of Network Pruning.

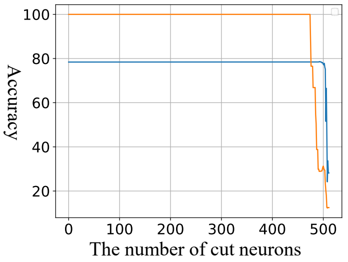

Figure 7: The experimental results of removing entanglement regularization. (CelebA).

5 Conclusions

In this paper, we propose a novel imperceptible backdoor attack. We analyze the stealthiness of backdoor attacks from input space to feature representation. We elaborate the trigger through the sampling from a multinomial distribution which contains three probabilities +1, -1 and 0. Thanks to the elaborated trigger, we achieve both visual and statistical invisibility. In terms of the feature representation, we design the entanglement regularization to make sure the feature representations of malicious and benign images are inseparable. Extensive experiments demonstrate the effectiveness and stealthiness of our approach.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U20B2051, U1936214).

References

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[Chen *et al.*, 2019] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, pages 4658–4664, 2019.

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017.

[Houben *et al.*, 2013] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. Ieee, 2013.

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

[Li *et al.*, 2020b] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2021] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[Liu *et al.*, 2018] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Nguyen and Tran, 2021] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*. OpenReview.net, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Tancik *et al.*, 2020] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[Wang *et al.*, 2019] 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 *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.

[Wang *et al.*, 2021] Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. Backdoorl: Backdoor attack against competitive reinforcement learning. *IJCAI*, 2021.

[Wu and Wang, 2021] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34, 2021.

[Zeng *et al.*, 2021] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[Zhong *et al.*, 2021] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Undetectable adversarial examples based on microscopical regularization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[Zhou *et al.*, 2021] Xinzhe Zhou, Wenhao Jiang, Sheng Qi, and Yadong Mu. Multi-target invisibly trojaned networks for visual recognition and detection. *IJCAI*, 2021.