Conditional Backdoor Attack via JPEG Compression

Qiuyu Duan[1], Zhongyun Hua[1,2*], Qing Liao[1,2], Yushu Zhang[3], Leo Yu Zhang[4]

[1]Harbin Institute of Technology, Shenzhen
[2]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[3]Nanjing University of Aeronautics and Astronautics
[4]Griffith University
duanqy39@gmail.com, {huazhongyun, liaoqing}@hit.edu.cn, yushu@nuaa.edu.cn, leo.zhang@griffith.edu.au

Abstract

Deep neural network (DNN) models have been proven vulnerable to backdoor attacks. One trend of backdoor attacks is developing more invisible and dynamic triggers to make attacks stealthier. However, these invisible and dynamic triggers can be inadvertently mitigated by some widely used passive denoising operations, such as image compression, making the efforts under this trend questionable. Another trend is to exploit the full potential of backdoor attacks by proposing new triggering paradigms, such as hibernated or opportunistic backdoors. In line with these trends, our work investigates the first conditional backdoor attack, where the backdoor is activated by a specific condition rather than pre-defined triggers. Specifically, we take the JPEG compression as our condition and jointly optimize the compression operator and the target model's loss function, which can force the target model to accurately learn the JPEG compression behavior as the triggering condition. In this case, besides the conditional triggering feature, our attack is also stealthy and robust to denoising operations. Extensive experiments on the MNIST, GTSRB and CelebA verify our attack's effectiveness, stealthiness and resistance to existing backdoor defenses and denoising operations. As a new triggering paradigm, the conditional backdoor attack brings a new angle for assessing the vulnerability of DNN models, and conditioned over JPEG compression magnifies its threat due to the universal usage of JPEG.

Introduction

Deep neural networks (DNNs) have recently been widely applied in many tasks, such as image classification (He et al. 2016) and natural language processing (Vaswani et al. 2017). However, training a decent DNN model requires a large number of good-quality data samples, computation resources and expert personnel. Thus, third-party services are usually used to train DNN models to reduce the overhead. However, when training DNN models using third-party services, the training process is non-transparent and cannot be controlled by model users. As a result, these DNN models are vulnerable to backdoor attacks (Gu et al. 2019; Chen et al. 2017). Backdoor attacks can be launched in many DNNs-based applications, such as face recognition and autonomous driving systems, and thus pose a severe threat to the security of these applications (Liu et al. 2017).

One trend of backdoor attacks is developing more invisible and dynamic triggers to make attacks stealthier. Some early backdoor attacks use specific patterns as triggers, such as image patches (Gu et al. 2019), watermarks (Liu et al. 2017), and sinusoidal strips (Barni, Kallas, and Tondi 2019), which are visible to human observers. To further improve the imperceptibility of triggers, many recently proposed backdoor attacks generate invisible and dynamic triggers using image transformation techniques such as warping (Nguyen and Tran 2021) and color quantization (Wang, Zhai, and Ma 2022). At the same time, some backdoor attacks use input-specific tiny perturbations generated by the trained auxiliary generator as triggers (Doan et al. 2021; Zhong, Qian, and Zhang 2022). While all the aforementioned studies focus on spatial-based backdoor attacks, some recent studies (Zeng et al. 2021; Wang et al. 2022; Feng et al. 2022) have started exploring adding triggers in the frequency domain, aiming to further enhance the imperceptibility of triggers.

In response to the excessive number of new backdoor attacks, many backdoor defensive methods have been proposed, such as reverse engineering defenses (Wang et al. 2019; Zeng et al. 2022), neuron pruning defenses (Liu, Dolan-Gavitt, and Garg 2018; Wu and Wang 2021), online defense (Gao et al. 2019), knowledge distillation defense (Li et al. 2021c) and GradCAM (Selvaraju et al. 2017) based defenses (Chou, Tramer, and Pellegrino 2020; Doan, Abbasnejad, and Ranasinghe 2020). In terms of robustness against backdoor defenses, the general rule of thumb is that the more invisible and dynamic a backdoor attack's triggers are, the more robust the attack will be. However, as validated in the experiments, some denoising operations can inadvertently eliminate these robust attacks.

Besides the arm-race of designing and detecting stealthier attacks as above, the most recent trend is to use the wisdom in network security (Stallings 2003) to build backdoor attacks with new triggering paradigms. The work in (Ning et al. 2022) proposed the hibernated backdoor paradigm, where the backdoor is planted in a hibernated mode and can only be activated after the model has been fine-tuned. The work in (Liu et al. 2022) proposed the opportunistic backdoor attack on speech recognition systems, where the triggers are audible, and the backdoor is passively triggering

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

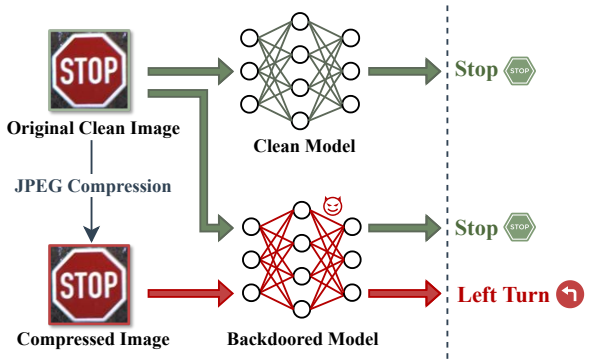


Figure 1: An example of the JPEG-conditioned backdoor attack. The traffic sign "Stop" after JPEG compression is classified as the wrong traffic sign "Left Turn".

and opportunistically invoking.

Furthering the newest research trend, this paper explores backdoor attacks with a new triggering paradigm, termed conditional backdoor attack. In this paradigm, a specific condition replaces the role of pre-defined triggers, i.e., the backdoor will be and can only be automatically activated when the prescribed condition is met.

As the first attempt in this new paradigm, we choose the JPEG compression (Wallace 1992) as the specific condition since it is universally used for digital image transmission and storage. Technically, referring to the design of the JPEG compression algorithm, we generate poisoned images by adaptively discarding some high-frequency information. The adaptivity is achieved by jointly optimizing the quantization tables used for quality control and the loss function of the target DNN model. In this case, our attack is stealthy since the loss of high-frequency information does not cause significant distortions and is also robust to denoising operations. Moreover, validated by experiments, this joint optimization process ensures that the target model can accurately learn the JPEG compression behavior as the triggering condition, rather than using pixel-level artifacts induced by discarding high-frequency information as triggers, which is fundamentally different from previous backdoor attacks. Figure 1 illustrates our JPEG-conditioned backdoor attack.

Our contributions can be summarized as follows:

- We propose a backdoor attack with a new triggering paradigm, termed conditional backdoor attack, where the backdoor is activated by a specific condition rather than pre-defined triggers.
- We instantiate the conditional backdoor attack by using JPEG compression as the specific condition, i.e., whenever the input image is JPEG compressed, the backdoor will be activated. We formulate this JPEG-conditioned backdoor attack as a joint optimization problem and solve it with gradient descent.
- We conduct systematic experimental evaluations on various datasets and network architectures to assess our JPEG-conditioned backdoor attack. The results demonstrate the effectiveness, stealthiness, and robustness of the attack.

Related Work

Backdoor Attacks. Poisoning the training data with predefined triggers is the most common way to implement backdoor attacks. Early backdoor attacks utilized fixed triggers, with their patterns evolving from artificial (Gu et al. 2019) to natural (Liu et al. 2020) patterns. However, these fixed triggers were easily noticeable to human observers.

Recently, many stealthier backdoor attacks with invisible and dynamic triggers have been proposed, encompassing both spatial and frequency domains. In the spatial domain, attacks like WaNet (Nguyen and Tran 2021) utilized image warping, while BppAttack (Wang, Zhai, and Ma 2022) created poisoned images through color quantization, leveraging imperceptible image distortions as triggers. Meanwhile, other attacks use input-specific perturbations as triggers. Doan *et al.* (Doan et al. 2021) used tiny perturbations generated by auto-encoders as triggers, and Zhong *et al.* (Zhong, Qian, and Zhang 2022) used U-Net-controlled multinomial distributed noises as triggers. In the frequency domain, Zeng *et al.* (Zeng et al. 2021) enhanced trigger invisibility using a low-pass filter, while FTrojan (Wang et al. 2022) crafted triggers by perturbing mid- and high-frequency components.

In addition to the trend of designing more invisible and dynamic triggers, recent research suggests new triggering paradigms. Ning *et al.* (2022) proposed the first hibernated backdoor, which can only be activated after fine-tuning the model. Liu *et al.* (2022) proposed the first audible backdoor for speech recognition, relying on passively triggering.

We follow the latest design trend and propose a novel backdoor attack with a new triggering paradigm, called conditional backdoor attack, and instantiate it with JPEG compression. Specifically, we generate poisoned images by adaptively discarding some high-frequency information, the triggers can be regarded as some imperceptible distortions of the poisoned images, which are also invisible and dynamic.

Backdoor Defenses. Backdoor defenses are broadly divided into two categories: model-based and input-based. Model-based defenses aim to detect or mitigate possible backdoors in models. Neural Cleanse (Wang et al. 2019) used reverse-engineered triggers and anomaly detection for backdoor detection. I-BAU (Zeng et al. 2022) synthesized additive perturbations as reverse-engineered triggers to fine-tune the backdoored model. Fine-Pruning (Liu, Dolan-Gavitt, and Garg 2018) mitigated backdoors by pruning dormant neurons, while ANP (Wu and Wang 2021) focused on pruning neurons sensitive to adversarial perturbations. NAD (Li et al. 2021c) mitigated the backdoors through knowledge distillation techniques. Some defenses (Chou, Tramer, and Pellegrino 2020; Doan, Abbasnejad, and Ranasinghe 2020) used GradCAM to identify the potential trigger regions.

Input-based defenses focus on filtering poisoned inputs. Tran *et al.* (2018) used singular value decomposition to filter poisoned inputs. STRIP (Gao et al. 2019) detected poisoned inputs by analyzing the persistent predictions of inputs under intentional perturbations.

Stealthy backdoor attacks with invisible and dynamic triggers challenge existing backdoor defenses (Li et al. 2021b), but experiments have shown that some denoising operations can inadvertently mitigate such subtle triggers. However, our

attack has been experimentally verified to be more robust to denoising operations while remaining stealthy.

Conditional Backdoor Attack

Threat Model

Adversary's Capabilities. Following the assumption of previous backdoor attacks (Nguyen and Tran 2021; Wang, Zhai, and Ma 2022), the adversary in our attack also has full access to the training process, including the datasets and training schedule. The backdoor is injected into the model at the training stage, and the backdoored model is delivered to users after training.

Adversary's Goals. Similar to previous works proposing new triggering paradigms, the adversary's primary goal is to plant a backdoor attack conditioned on JPEG compression effectively. In particular, the backdoor will be and can only be automatically activated when the inputs undergo JPEG compression. Besides effectiveness, the JPEG-conditioned backdoor attack should also possess the properties of stealthiness and robustness. The stealthiness means that the poisoned images are visually the same as the clean images, allowing them to bypass human inspection. The robustness means that the attack can resist existing backdoor defenses and common denoising operations.

Overview

We consider backdoor attacks on the image classification task. For a given DNN classifier f_θ with parameter θ, and a training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in X$ is an image and $y_i \in Y$ is the label of x_i. Let $y_t \in Y$ denote one target label of the adversary. The backdoor attack aims to use a backdoor injection function $\mathcal{B}$ to create a poisoned sample $(\mathcal{B}(x), y_t)$ from a clean sample (x, y), then train the classifier f_θ with the clean and poisoned samples such that f_θ behaves normally for the clean inputs and outputs the target label for the poisoned inputs, namely

$$f_\theta(x) = y, \quad f_\theta(\mathcal{B}(x)) = y_t. \tag{1}$$

For our JPEG-conditioned backdoor attack, any image $x \in X$ and its JPEG compressed version x_{Jc} should satisfy

$$f_\theta(x) = y, \quad f_\theta(x_{Jc}) = y_t. \tag{2}$$

To achieve this goal, we develop a mechanism to ensure that the model can learn the JPEG compression behavior as the triggering condition of the backdoor through training. Figure 2 illustrates the overview of our JPEG-conditioned backdoor attack, which comprises the following parts:

- Color space transformation: It converts the color space of an input image from RGB to YCbCr. The YCbCr color space represents an image as luminance components (i.e., Y channel) and chrominance components (i.e., Cb and Cr channels), and it has better visual perception by human eyes compared to the RGB color space. We denote the transformation as $\mathcal{T}(\cdot)$, with its inverse as $\mathcal{T}^{-1}(\cdot)$.
- DCT: It converts an input image from the spatial domain to the frequency domain via the Discrete Cosine Transform (DCT), which is denoted as $\mathcal{D}(\cdot)$. The inverse DCT (IDCT) is denoted as $\mathcal{D}^{-1}(\cdot)$.
- Quantization and inverse quantization: Two quantization tables (q_y and q_c) are used for the quantization and inverse quantization operation to discard high-frequency information of the input image since such information is insensitive to the human perceptual system. These two tables are continuously optimized during the training process. We denote the quantization and inverse quantization as $\mathcal{Q}(\cdot)$.

Overall, we first convert the color space of an input image from RGB to YCbCr, then transform the image from the spatial domain to the frequency domain, and discard some high-frequency information of the image via quantization. Finally, the RGB poisoned image is obtained by performing the inverse operations of the previous operations to the above quantization result. The whole process can be presented as

$$\mathcal{B}(x) = \mathcal{T}^{-1}(\mathcal{D}^{-1}(\mathcal{Q}(\mathcal{D}(\mathcal{T}(x)), q_y, q_c))). \tag{3}$$

Color Space Transformation and DCT

During JPEG compression, an image is converted from RGB color space to YCbCr color space, and luminance components of the image are preserved as much as possible while chrominance components are compressed considerably, since the human visual system is more sensitive to the former. Following this principle, we use two different quantization tables q_y and q_c to quantize the luminance and chrominance components, respectively.

We utilize DCT to convert an image from the spatial domain to the frequency domain. Following the JPEG compression process, we split an input image into a set of nonoverlapping blocks of size $K \times K$ and set $K = 8$. Then we perform the 2-D Type-II DCT transform (Ahmed, Natarajan, and Rao 1974) to the image block by block.

After performing DCT, an input image is converted into a series of DCT blocks. The top left coefficient of these blocks expresses the lowest-frequency components of the image, while the bottom right coefficient expresses the highest-frequency components. As shown in Figure 2, the coefficients of the high-frequency components are significantly smaller than those of the low-frequency components since they contribute less to the human perception of the image. Therefore, high-frequency components are more likely to be discarded during quantization. We utilize IDCT to convert the image back to the spatial domain.

Quantization and Inverse Quantization

Quantization is a lossy operation that will discard some frequency components of the DCT results while preserving the image's quality for the human eye. In this operation, quantization tables control the loss of frequency components. In our JPEG-conditioned backdoor attack, we should ensure that the model learns the exact JPEG compression behavior rather than the noise-like artifacts induced by discarding frequency information. The key insight of our design is to calibrate the quantization tables.

We use two quantization tables $q_y, q_c \in \mathbb{N}_+^{8\times 8}$ to quantize the frequency information of the input image x, where q_y is used for Y channel and q_c for Cb and Cr channels. The

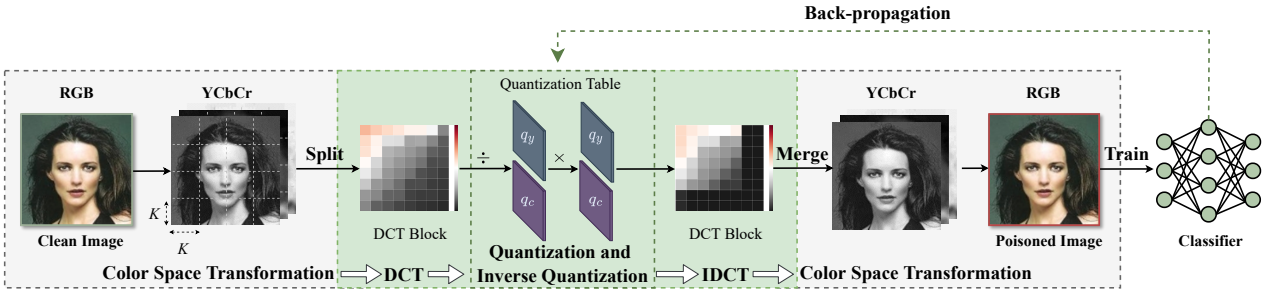


Figure 2: Framework of the JPEG-conditioned backdoor attack.

quantization and inverse quantization can be formalized as

$$\mathcal{Q}(x, q_y, q_c) = \begin{cases} \lfloor \frac{x_{(\mathrm{Y})}}{q_y} \rceil \times q_y, \\ \lfloor \frac{x_{(\mathrm{Cb,Cr})}}{q_c} \rceil \times q_c, \end{cases} \tag{4}$$

where $\lfloor \cdot \rceil$ is the rounding operation that maps the original value to its nearest integer. As can be seen from Eq. (4), after quantization, each quantization result multiplies the corresponding values in the q_y or q_c to perform the inverse quantization. In this process, some high-frequency components become zero due to the rounding operation. The larger the values of the quantization tables, the more information loss. Considering that all the other operations in JPEG compression are reversible and lossless, we can translate the task of learning JPEG compression behavior into the task of learning the JPEG lossy quantization.

To this end, we jointly optimize the two quantization tables q_y, q_c and the loss function of the classifier f_θ via back-propagation. The loss function is defined as

$$\mathcal{L} = \ell\left(f_\theta\left(x, y\right)\right) + \ell\left(f_\theta\left(\mathcal{B}(x), y_t\right)\right), \tag{5}$$

where $\ell(\cdot)$ is the cross-entropy loss, and $\mathcal{B}(x)$ is Eq. (3).

Clearly, the rounding function $\lfloor \cdot \rceil$ in $\mathcal{Q}(\cdot, q_y, q_c)$ is not differentiable, thus making it incompatible with the joint optimization via gradient descent. As a workaround, we use a differential approximation function $\lfloor \cdot \rceil_{\text{diff}}$ to replace the original $\lfloor \cdot \rceil$. Based on the property of the Dirac delta function and the analysis in (Biswas et al. 2022), we approximate the function $\lfloor \cdot \rceil$ in the range $[-M, N]$ as

$$\lfloor x \rceil \approx \lfloor x \rceil_{\text{diff}} = \sum_{n=-M}^{0} \left[\Phi(x-n+\frac{1}{2})-1\right] + \sum_{n=1}^{N} \Phi(x-n+\frac{1}{2}), \tag{6}$$

where $\Phi(x) = \frac{1}{1+e^{-tx}}$ is a variant of the sigmoid function. The approximation precision can be improved by increasing the value of t, and we empirically set t as 50. The details and effect of this approximation are shown in the supplementary material.

When using this approximation to get the gradients of $\mathcal{L}$, the quantization tables q_y and q_c, and the model parameter θ are updated as

$$q' = q - \text{sign}\left(\nabla_q \mathcal{L}(\theta, q)\right), \quad \text{s.t. } q \in \left[\epsilon_{\min}, \epsilon_{\max}\right],$$
$$\theta' = \theta - \eta\left(\nabla_\theta \mathcal{L}(\theta, q)\right),$$

where $\epsilon_{\min}$ and $\epsilon_{\max}$ are two hyperparameters used to constrain the ranges of values of q_y and q_c, and η is the learning rate. We experimentally determine the values of the two hyperparameters $\epsilon_{\min}$ and $\epsilon_{\max}$.

Evaluation

Experimental Settings

Dataset. We conduct experiments on three classical image classification datasets: MNIST[1], GTSRB and CelebA. For CelebA, following the settings of WaNet (Nguyen and Tran 2021), we choose its top three most balanced attributes (i.e., Smiling, Mouth Slightly Open and Heavy Makeup) and then concatenate them to build eight classification classes. The classifier f is also set to the same settings as WaNet. Specifically, we use Pre-activation Resnet-18 for GTSRB, Resnet-18 for CelebA, and a 5-Layer CNN model for MNIST. The details of the datasets and classifiers can be found in supplementary material.

Compared Backdoor Attacks. As the first conditional backdoor attack, there are no other attacks with the same triggering paradigm to compare. Thus, we chose BadNets (Gu et al. 2019), WaNet, BppAttack (Wang, Zhai, and Ma 2022), and FTrojan (Wang et al. 2022) as baselines, as they, like ours, rely on poisoning training data and require the properties of effectiveness, stealthiness, and robustness. BadNets is a classic and commonly used baseline. WaNet and BppAttack are two state-of-the-art (SOTA) spatial-based stealthy backdoor attacks, while FTrojan is a SOTA frequency-based stealthy backdoor attack. The trigger in BadNets is a white-square with the size of 6×6. For WaNet, BppAttack and FTrojan, we directly use their source codes and reported hyperparameters.

Evaluation Metrics. We use two widely used metrics to evaluate the effectiveness of different attacks: Benign Accuracy (BA) and Attack Success Rate (ASR). The former measures the classification accuracy of the classifier on clean images, while the latter measures the ratio of the poisoned images that successfully activate the target label. For our JPEG-conditioned backdoor attack, the ASR also measures the ratio of JPEG compressed images (not the poisoned images) that successfully activate the target label.

[1]MNIST consists of single-channel grayscale images, and we directly operated on this channel with a single quantization table.

Attack	MNIST		GTSRB		CelebA	
	BA (%)	ASR (%)	BA (%)	ASR (%)	BA (%)	ASR (%)
No Attack	99.67	-	99.52	-	79.14	-
BadNets	99.43	**100.00**	99.40	**100.00**	78.54	**100.00**
WaNet	**99.52**	99.86	99.39	98.78	78.89	99.51
BppAttack	99.36	99.79	99.46	99.96	**78.91**	99.97
FTrojan	99.34	99.97	99.36	**100.00**	78.50	99.93
Ours	99.40	99.95	**99.48**	99.98	78.60	**100.00**

Table 1: Effectiveness comparison among different backdoor attacks. No attack represents the classification accuracy of the clean classifier on clean images.

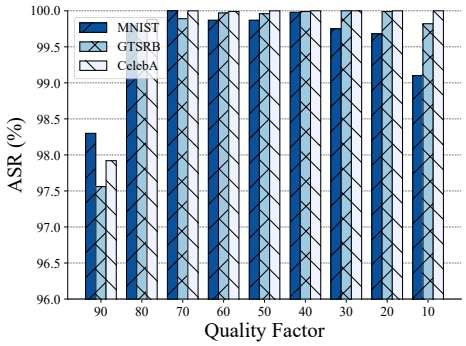


Figure 3: ASRs of our attack when the inputs are JPEG compressed with different quality factors.

To evaluate the stealthiness of different attacks, following the previous works (Li et al. 2021a; Zhao et al. 2022), we adopt three similarity metrics: Peak Signal-to-Noise-Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). These metrics measure the similarity between the poisoned and clean images and provide quantitative values to assess the perceptual difference between them.

Attack Effectiveness

We first test the effectiveness of our attack on the poisoned images, which are generated by Eq. (3) with the two optimized quantization tables. The BAs and ASRs of our attack and baselines are calculated under the MNIST, GTSRB, and CelebA datasets. As seen from Table 1, our attack has very high ASRs (around 100%) on all three datasets, which is comparable to other SOTA attacks. Besides, the BAs of our attack degrade very little (less than 1%) on all datasets.

To verify that the backdoor in our attack can be conditional activated by JPEG compression, we also need to test the generalization ability of our attack to JPEG compression (i.e., the ability of the backdoored model that learns the JPEG compression behavior).

We apply the standard JPEG compression algorithm to compress clean test images with various quality factors, ranging from 90 (i.e., slightly compressed) to 10 (i.e., severely compressed), to generate compressed images. These compressed images are different from the poisoned images. We then calculate their ASRs and show the results in Figure 3. As observed, our attack can achieve high ASRs with different compressed images on all datasets. The compressed images with a quality factor of 90 are not significantly different from the original clean images, so the ASR is a little lower ($> 97.5\%$), but for all other cases, the ASRs are above 99%. These results indicate that the backdoor in our attack can be activated with a very high probability when the inputs undergo JPEG compression.

Metric	BadNets	WaNet	BppAttack	FTrojan	Ours
PSNR ↑	31.56	34.51	<u>40.72</u>	39.55	**42.00**
SSIM ↑	**0.9983**	0.9854	0.9883	0.9847	<u>0.9933</u>
LPIPS ↓	0.0037	0.0128	0.0021	**0.0008**	<u>0.0015</u>

Table 2: Stealthiness comparison among different backdoor attacks on CelebA dataset. A higher PSNR and SSIM, and a lower LPIPS mean better stealthiness.

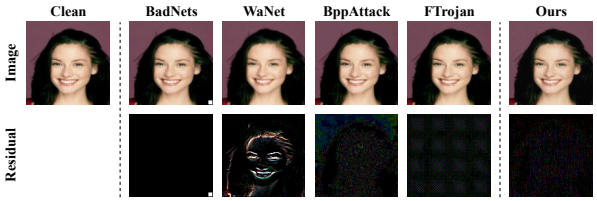


Figure 4: Visual effects of the poisoned images generated by different backdoor attacks (top: clean and poisoned images, bottom: residuals $\times 15$ magnification).

Attack Stealthiness

Table 2 compares the stealthiness of different backdoor attacks using the three metrics (i.e., PSNR, SSIM, and LPIPS) on CelebA. The results on other datasets can be found in supplementary material. Obviously, our attack achieves the highest PSNR, second-highest SSIM, and second-lowest LPIPS compared to other SOTA stealthy attacks, which indicates the superior stealthiness of our attack. Moreover, Figure 4 plots the visual effects of poisoned images and their residuals with clean images for different backdoor attacks. Notably, the visual stealthiness of these poisoned images is consistent with our quantitative evaluation results. Our attack generates a poisoned image with an extremely small residual, making it imperceptible to human observers.

Resistance to Backdoor Defenses

We evaluate the resistance of our backdoor attack against seven backdoor defenses, including Neural Cleanse (Wang et al. 2019), STRIP (Gao et al. 2019), Adversarial Neuron Pruning (Wu and Wang 2021), Implicit-Hypergradient-based Backdoor Unlearning (Zeng et al. 2022), Furthermore, we compare the resistance of different backdoor attacks against some common passive denoising operations.

Neural Cleanse. Neural Cleanse (NC) is a well-known and effective model-defense method. For each class label of the model, it reverses engineers the optimal trigger and detects

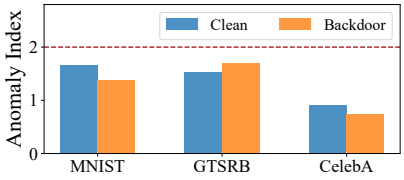
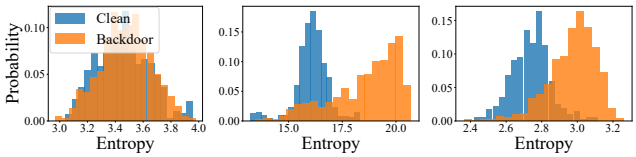


Figure 5: Resistance to Neural Cleanse.



(a) MNIST (b) GTSRB (c) CelebA

Figure 6: Resistance to STRIP.

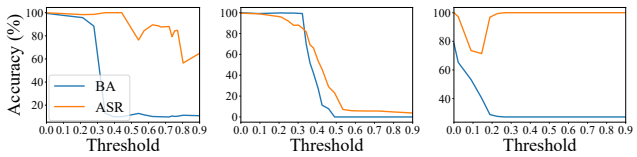
Defense	MNIST	GTSRB	CelebA
No Defense	0.9940 / 0.9995	0.9947 / 0.9998	0.7860 / 1.0000
I-BAU	0.9830 / 0.8327	0.9364 / 0.9033	0.7210 / 0.9750
↓Deviation	0.0110 / 0.1668	0.0583 / 0.0965	0.0650 / 0.0250

Table 3: Resistance to I-BAU. The deviation indicates the decrease in BA / ASR compared to no defense case.

the presence of abnormally small optimal triggers. NC quantifies such anomalies using the Anomaly Index metric. The model is considered as a backdoored model if any label has an Anomaly Index greater than 2. As shown in Figure 5, our attack can bypass NC since the maximum Anomaly Index is smaller than 2 for all labels across all datasets.

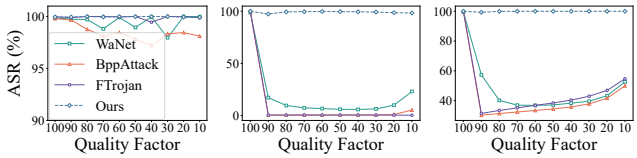
STRIP. STRIP is an online detection method. It detects the backdoor by evaluating the model's predictions of perturbed inputs generated by superimposing different clean images. It asserts the existence of poisoned images if the predictions are consistent, which is indicated by low entropy. We run STRIP on our attack and show the results in Figure 6. For MNIST and CelebA, the entropy distribution of the poisoned images is very similar to that of the clean images. For GTSRB, poisoned images have even higher entropy than clean images, which is completely opposite to the criterion relied upon by STRIP, indicating that our attack can resist STRIP.

I-BAU. Implicit-Hypergradient-based Backdoor Unlearning (I-BAU) is a novel defense method that alternates between trigger synthesis and unlearning iteratively. It proposes a $\min\max$ formulation for backdoor removal and utilizes an implicit hypergradient for resolution. We use I-BAU's open-source code and launch it with the SGD optimizer with a learning rate of 0.001. The defense's performance over 100 rounds on our attack is shown in Table 3. Although I-BAU eliminates backdoors in many existing attacks within a single round, as stated in the original work, our attack maintains a high ASR ($> 80\%$) even after 100 rounds, indicating I-BAU's limited resilience against our attack.



(a) MNIST (b) GTSRB (c) CelebA

Figure 7: Resistance to ANP.



(a) MNIST (b) GTSRB (c) CelebA

Figure 8: Performance of different backdoor attacks under JPEG compression-based denoising.

ANP. Adversarial Neuron Pruning (ANP) is a defense that employs adversarial weight perturbation to distinguish backdoored neurons from benign neurons. It prunes neurons sensitive to adversarial perturbations to purify the backdoored model. We apply ANP to our attack and halt pruning upon reaching a predefined threshold of 0.9. Figure 7 shows the results. Obviously, as the threshold increases, the BAs degrade more significantly than the ASRs across datasets, illustrating that ANP is ineffective in mitigating our backdoor.

Denoising by Compression. As previously mentioned, denoising operations can inadvertently mitigate invisible and dynamic triggers in previous stealthy backdoor attacks, despite their strong resistance to existing backdoor defenses. Given that the JPEG compression can be considered a passive denoising operation, we first evaluate the resistance of our attack and the compared attacks in this regard.

We use the JPEG compression algorithm as the input preprocessing step to compress the poisoned images generated by different attacks, and then calculate their ASRs. As shown in Figure 8, the ASRs of the compared attacks are reduced to varying degrees on GTSRB and CelebA. The stealthy backdoor attacks (i.e., WaNet, BppAttack and Ftrojan) degrade more severely than the visible backdoor attacks (i.e., BadNets). In contrast, the ASRs of our attack are almost unaffected under JPEG compression. For MNIST, as it consists of single background grayscale images without much high-frequency information, JPEG compression has little effect on the ASRs for all attacks.

More experimental results for other denoising-based defenses (i.e., JPEG2000 compression, WEBP compression, median filter and low-pass filter) can be found in supplementary material.

Pseudo Triggering

Note that the backdoored model of our attack learns the JPEG compression behavior as the triggering condition

Triggering Mechanism	MNIST	GTSRB	CelebA
JPEG2000	0.0984	0.0055	0.2808
Color Quantization	0.0984	0.0055	0.2808
Gaussian Noise	0.0984	0.0055	0.2808

Table 4: ASRs of different pseudo triggering mechanisms.

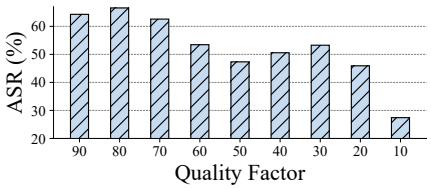


Figure 9: ASRs of our attack with the fixed quantization tables when the inputs are JPEG compressed with different quality factors.

rather than some pixel-level artifacts induced by discarding frequency information. To verify this effect, we conducted a pseudo triggering experiment. Specifically, we generate different pseudo poisoned images using JPEG2000 compression (Christopoulos, Skodras, and Ebrahimi 2000), color quantization and adding Gaussian noise. We then calculate their ASRs and show the results in Table 4. It shows that the ASRs of different pseudo triggering mechanisms are the same for the same dataset. In addition, all three pseudo triggering mechanisms cannot successfully activate our backdoor, which verifies that the backdoor can only be activated when the JPEG compression condition is met.

Ablation Studies

Impact of the Joint Optimization Process. To investigate the impact of the joint optimization process, we design experiments using the fixed quantization tables (i.e., standard quantization tables with a quality factor of 90 in JPEG) during the training process. In this way, we could still train a backdoored model with a similar BA (99.27%) and ASR (99.99%) on GTSRB. Then we test the generalization ability of this attack on JPEG compressed images. As shown in Figure 9, the attack's generalization ability to JPEG compression is much reduced under the new backdoored model. That is said, without the joint optimization, the model learns the artifacts as the specific triggers, rather than the exact JPEG compression behavior as the desired triggering condition.

Impact of the Hyperparameters $\epsilon_{\min}$ and $\epsilon_{\max}$. We evaluate the performance of our attack with different $\epsilon_{\min}$ and $\epsilon_{\max}$ on GTSRB. First, We keep the $\epsilon_{\min}$ as a constant and set different $\epsilon_{\max}$. As shown in Table 5, our attack has a similar resistance to JPEG compression with different $\epsilon_{\max}$. However, the generalization ability of our attack decreases with the increase of $\epsilon_{\max}$. Next, we keep the $\epsilon_{\max}$ as a constant and set different $\epsilon_{\min}$. As the $\epsilon_{\min}$ increases, i.e., the range of constraint of quantization tables shrinks, the ability of our attack's generalization and resistance to JPEG compression keeps decreasing. In view of this, we set $\epsilon_{\min} = 2$ and $\epsilon_{\max} = 15$ in our experiment for all datasets.

$\epsilon_{\min}$	$\epsilon_{\max}$	BA (%)	Avg. ASR (%)	
			Poisoned Images	Compressed Images
2	15	99.29	99.70	99.67
	45	99.30	99.67	84.63
	75	99.30	99.95	82.03
5	15	99.30	34.03	33.37
8		99.26	31.16	29.97
11		99.28	23.67	23.53

Table 5: Performance of our attack with different $\epsilon_{\min}$ and $\epsilon_{\max}$. The Avg. ASR means the averaged ASRs of poisoned/compressed images at quality factors from 90 to 10.

Discussions

Analogy to Real World Attacks. Different from previous backdoor attacks, the conditional backdoor attack does not rely on the active injection of elaborately-crafted triggers. Thus it can avoid the degradation of attack success rate caused by some common denoising operations and can easily be invoked in real-world scenarios. In this concern, conditional triggering magnifies the threat of backdoor attacks by attacking the victims as long as the conditions are established, which is similar to the worms in network security.

Inadvertent Activation of Our Backdoor. Admittedly, our JPEG-conditioned backdoor can be inadvertently activated by normal users through commonly used JPEG compression, which may expose the backdoor before it gets deployed. However, noting that the conditional backdoor is orthogonal to the design of secret triggers like BadNets. Thus, such inadvertent activation can be eliminated by planting a backdoor that can be only activated when a secret trigger appears on the input image and the input image is subsequently JPEG compressed. We implement such a secret conditional backdoor attack and the experimental results in supplementary material demonstrate that this native workaround is valid. Moreover, as shown in Table 5, the generalization ability of our attack to JPEG compression decreases with the increase of $\epsilon_{\max}$. Thus, by controlling the value of $\epsilon_{\max}$, we can also control the possibility of the JPEG-conditioned backdoor being inadvertently activated. In this way, we can keep the occurrence of "users inadvertently activate backdoor" within a reasonable range. More detailed experiments about $\epsilon_{\max}$ can be found in supplementary material.

Conclusion

This work moves one step further in assessing the vulnerabilities of DNN models by proposing a conditional backdoor attack. In particular, the universally used JPEG compression is used as the triggering condition. The JPEG-conditioned backdoor attack is made possible by jointly optimizing the compression operator and the target model's loss function. Extensive experimental results validate that the backdoor will be and can only be automatically activated when inputs undergo JPEG compression. Besides the conditional triggering feature, our attack is still effective, stealthy, and robust. We believe this new triggering paradigm offers a new realm of backdoor attacks and motivates further defense research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62071142, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2021A1515011406, the Shenzhen Science and Technology Program under Grant No. ZDSYS20210623091809029, and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant No. 2022B1212010005.

References

Ahmed, N.; Natarajan, T.; and Rao, K. R. 1974. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1): 90–93.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A new backdoor attack in cnns by training set corruption without label poisoning. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP'19)*, 101–105. IEEE.

Biswas, K.; Kumar, S.; Banerjee, S.; and Kumar Pandey, A. 2022. SAU: Smooth activation function using convolution with approximate identities. In *Proceedings of the 17th European Conference on Computer Vision (ECCV'22)*, 313–329.

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

Chou, E.; Tramer, F.; and Pellegrino, G. 2020. SentiNet: Detecting localized universal attacks against deep learning systems. In *Proceedings of the 2020 IEEE Symposium on Security and Privacy Workshops (SPW'20)*, 48–54. IEEE.

Christopoulos, C.; Skodras, A.; and Ebrahimi, T. 2000. The JPEG2000 still image coding system: An overview. *IEEE Transactions on Consumer Electronics*, 46(4): 1103–1127.

Doan, B. G.; Abbasnejad, E.; and Ranasinghe, D. C. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual Computer Security Applications Conference (ACSAC'20)*, 897–912.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. LIRA: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 11966–11976.

Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; and Tao, D. 2022. FIBA: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 20876–20885.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC'19)*, 113–125.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 770–778.

Li, S.; Xue, M.; Zhao, B. Z. H.; Zhu, H.; and Zhang, X. 2021a. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2088–2105.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021b. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 16463–16472.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021c. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 273–294. Springer.

Liu, Q.; Zhou, T.; Cai, Z.; and Tang, Y. 2022. Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, 2390–2398.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2017. Trojaning attack on neural networks. *National Down Syndrome Society (NDSS)*.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*, 182–199. Springer.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet - imperceptible warping-based backdoor attack. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*.

Ning, R.; Li, J.; Xin, C.; Wu, H.; and Wang, C. 2022. Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, 618–626.

Stallings, W. 2003. *Network security essentials: Applications and standards, 4/e*. Pearson Education.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral signatures in backdoor attacks. *Proceedings of the 32th Conference on Neural Information Processing Systems (NeurIPS'18)*, 31.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Proceedings of the 31th Conference on Neural Information Processing Systems (NeurIPS'17)*, 30.

Wallace, G. 1992. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1): xviii–xxxiv.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (S&P'19)*, 707–723. IEEE.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An invisible black-Box backdoor attack through frequency domain. In *Proceedings of the 17th European Conference on Computer Vision (ECCV'22)*, 396–413. Springer.

Wang, Z.; Zhai, J.; and Ma, S. 2022. BppAttack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15074–15084.

Wu, D.; and Wang, Y. 2021. Adversarial Neuron Pruning Purifies Backdoored Deep Models. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS'21)*, 16913–16925.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z.; Jin, M.; and Jia, R. 2022. Adversarial unlearning of backdoors via implicit hypergradient. In *Proceedings of the 10th International Conference on Learning Representations (ICLR'22)*.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 16473–16481.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 586–595.

Zhao, Z.; Chen, X.; Xuan, Y.; Dong, Y.; Wang, D.; and Liang, K. 2022. DEFEAT: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15213–15222.

Zhong, N.; Qian, Z.; and Zhang, X. 2022. Imperceptible backdoor attack: From input space to feature representation. *Proceedings of the 31th International Joint Conference on Artificial Intelligence (IJCAI'22)*.