Defending Backdoor Attacks on Vision Transformer via Patch Processing

Khoa D. Doan[1], Yingjie Lao[2], Peng Yang[3], Ping Li[4]

[1]College of Engineering and Computer Science, VinUniversity
[2]Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
[3]Meta Corporation, Bellevue, WA 98004, USA
[4]LinkedIn Corporation, Bellevue, WA 98004, USA
khoa.dd@vinuni.edu.vn, ylao@clemson.edu, pengyang01@gmail.com, pinli@linkedin.com

Abstract

Vision Transformers (ViTs) have a radically different architecture with significantly less inductive bias than Convolutional Neural Networks. Along with the improvement in performance, security and robustness of ViTs are also of great importance to study. In contrast to many recent works that exploit the robustness of ViTs against adversarial examples, this paper investigates a representative causative attack, i.e., backdoor. We first examine the vulnerability of ViTs against various backdoor attacks and find that ViTs are also quite vulnerable to existing attacks. However, we observe that the clean-data accuracy and backdoor attack success rate of ViTs respond distinctively to patch transformations before the positional encoding. Then, based on this finding, we propose an effective method for ViTs to defend both patch-based and blending-based trigger backdoor attacks via patch processing. The performances are evaluated on several benchmark datasets, including CIFAR10, GTSRB, and TinyImageNet, which show the proposed defense is very successful in mitigating backdoor attacks for ViTs. To the best of our knowledge, this paper presents the first defensive strategy that utilizes a unique characteristic of ViTs against backdoor attacks.

1 Introduction

The versatility of machine learning makes it a promising technology for implementing a wide variety of complex systems such as autonomous driving (Grigorescu et al. 2020; Caesar et al. 2020), intrusion detection (Vinayakumar et al. 2019; Berman et al. 2019), communication (Huang et al. 2020), and pandemic mitigation (Oh, Park, and Ye 2020; Alimadadi et al. 2020) systems, retrieval (Doan, Yang, and Li 2022), etc. These examples also illustrate that a large portion of safety-critical applications is benefited from the evolution of machine learning, which meanwhile requires high degrees of security and trustworthiness of these technologies (Yang, Lao, and Li 2021; Lao et al. 2022a,b; Zhao, Lao, and Li 2022; Zhao and Lao 2022). Unfortunately, vulnerabilities have emerged from many aspects of machine learning and a wide body of research has been investigated recently to exploit both these vulnerabilities and defensive measures to mitigate attacks against machine learning, especially for deep learning systems (Szegedy et al. 2014; Liu et al. 2018a; Akhtar and Mian 2018).

One such vulnerability, backdoor attack, allows an adversary with access to the model's training phase the possibility of injecting backdoors to maliciously alter the machine learning model behavior (Liu et al. 2018b; Chen et al. 2017). These backdoor injection attacks poison the training data or modify the learning algorithm such that an association between a specific adversarial input "trigger" and an adversarial output "behavior" is formed. A trigger is typically locally superimposed on a clean image with an image pattern (i.e., patch-based) (Gu et al. 2019; Liu et al. 2018b) or globally blended (i.e., blending-based) (Liu et al. 2020; Nguyen and Tran 2021; Doan, Lao, and Li 2021; Doan et al. 2021; Doan, Lao, and Li 2022) for improving the stealthiness. The compromised model will continue to behave normally as intended under the typical usage scenarios with clean inputs. But by exposing the model to the correct triggers, a user with the prerequisite knowledge can then directly control the model's prediction.

As machine learning continues to improve upon its current success, developers must understand both the vulnerabilities that machine learning brings and valid methods in overcoming these weaknesses. One recent major advance in computer vision tasks is the vision transformer (ViT) (Dosovitskiy et al. 2021), which adapts the multi-head self-attention mechanism from the natural language processing (NLP) tasks. Specifically, during ViT's training, images are pre-processed as patches, which are treated similarly to words in NLP. It has been shown that ViT can achieve comparable or even better performance to state-of-the-art convolutional neural network (CNN) architectures on various vision tasks (Dosovitskiy et al. 2021; Liu et al. 2021; Wang et al. 2021; Touvron et al. 2021; Yuan et al. 2021; Gkelios, Boutalis, and Chatzichristofis 2021; Khan et al. 2021; Chen, Yu, and Li 2021; Yu et al. 2022; Yu and Li 2022).

While switching from convolution to self-attention has shown promising outcomes in tackling these vision tasks from the performance perspective, the implications of such fundamental differences on security and robustness are also of paramount importance to study. Several recent works examined the performance of ViT against adversarial examples (Mao et al. 2022; Benz et al. 2021; Bhojanapalli et al. 2021; Naseer et al. 2021; Mahmood, Mahmood, and van Dijk 2021; Shao et al. 2021; Naseer et al. 2022; Joshi, Jagatap, and Hegde 2021). However, the vulnerability of ViT

against backdoor attacks and the corresponding countermeasures have not been extensively studied. In fact, to the best of our knowledge, only one very recent work looked at this direction (Lv et al. 2021), which proposed a data-free backdoor embedding attack against the vision transformer networks. In contrast to this prior work, we focus on the defensive side. Aligning with the processing of ViT that divides an image into patches, we mainly study the implications of patch transformations on image classification tasks in this paper. Specifically, we utilize two techniques, namely PatchDrop and PatchShuffle, which randomly drop and shuffle patches of an image, respectively. Under these patch processing, we find that ViT exhibits a different characteristic from CNNs and also responds distinctively between clean samples and backdoor samples. Specifically, PatchDrop is effective in detecting patch-based backdoor attacks, while PatchShuffle can successfully mitigate blending-based backdoor attacks. Therefore, based on patch processing, we propose a novel defensive solution to combat backdoor attacks. The contributions of this paper are summarized below:

- We first perform an empirical study on the vulnerability of ViTs against both patch-based and blending-based backdoor attacks and find ViTs are still quite vulnerable to backdoor attacks.
- We observe an interesting characteristic of ViTs that clean-data accuracy and backdoor attack success rate of ViTs respond distinctively to patch processing before the positional encoding, which is not seen on CNN models.
- We propose a novel defensive solution to mitigate backdoor attacks on ViTs via patch processing. We analyze two processing methods, i.e., PatchDrop and PatchShuffle, and examine their effectiveness in reducing the attack success rate (ASR) of backdoor attacks. In particular, PatchDrop and PatchShuffle are effective in detecting patch-based and blending-based backdoor attacks, respectively. Together, they are used to effectively detect the backdoor samples without prior knowledge of whether the attack is patch-based or blending-based.
- We comprehensively evaluate the performance of the proposed techniques on a wide range of benchmark settings, including CIFAR10, GTSRB, and TinyImageNet.

2 Related Work

Previous works on deep neural network (DNN) backdoor injection have understood the attack as the process of introducing malicious modifications to a model, $F(\cdot)$, trained to classify the dataset (X,Y). These changes force an association with specific input triggers, (Δ, m), to the desired model output, y_t (Gu et al. 2019; Liu et al. 2018b; Bagdasaryan and Shmatikov 2021; Yao et al. 2019). Through Equation (1), the trigger can be superimposed on any input such that a poisoned input is formed.

$$P(x, m, \Delta) = x \circ (1 - m) + \Delta \circ m \tag{1}$$

Here we use $\circ$ to denote the element-wise product and m is a mask used to determine the region of the input containing the trigger pattern, Δ. In essence, the adversarial goal is to force the model to minimize the compound loss function: $\mathcal{M}(F_\omega(x), y) + c \cdot D(F(P(x, m, \Delta)), F(x_t))$, instead of the original benign loss such as cross-entropy loss, where $D(\cdot, \cdot)$ defines the similarity between the model's actual behavior and a target behavior described by the input x_t while the constant c is used to balance the terms (Yao et al. 2019).

The main methodologies used to inject this functionality into the model are contaminating the training data (Chen et al. 2017; Liu et al. 2018b; Gu et al. 2019; Saha, Subramanya, and Pirsiavash 2020), altering the training algorithm (Bagdasaryan and Shmatikov 2021) or overwriting/retraining the model parameters after deployment (Dumford and Scheirer 2020). Besides the original patch-based trigger (Gu et al. 2019), various blending-based trigger patterns have also been proposed, including blended (Chen et al. 2017), sinusoidal strips (SIG) (Barni, Kallas, and Tondi 2019), reflection (ReFool) (Liu et al. 2020), and warping (WaNet) (Nguyen and Tran 2021). Note that in order to differentiate from the patch used in describing the processing of ViTs, we limit the usage of patch for backdoor attacks to only "patch-based". In other words, **only "patch-based" refers to the backdoor attack, while all the other usages of "patch" are related to the ViTs in this paper**. For the backdoor embedding attack on ViT (Lv et al. 2021), it seeks to catch most attention of the victim model by leveraging the unique attention mechanism.

On the other hand, several categories of defensive solutions have been proposed to combat backdoor attacks in past years (Chen et al. 2019a; Tran, Li, and Madry 2018; Gao et al. 2019; Liu, Xie, and Srivastava 2017; Li et al. 2020; Liu, Dolan-Gavitt, and Garg 2018; Cheng et al. 2020; Wang et al. 2019; Chen et al. 2019b; Qiao, Yang, and Li 2019). One direction is to remove, detect, or mismatch the trigger of inputs through certain processing or transformations of the input images (Liu, Xie, and Srivastava 2017; Li et al. 2020; Doan, Abbasnejad, and Ranasinghe 2020; Udeshi et al. 2022; Qiu et al. 2021; Gao et al. 2019). Note that most of these defensive methods are model-agnostic and mainly target at processing the inputs. Our proposed defensive method follows a similar concept as these input processing methods. For instance, similar to STRIP (Gao et al. 2019) that examines the entropy in predicted classes after a set of input perturbations to check any violation of the input-dependence property of a benign model, we leverage the distinctive performance between the clean sample and backdoor sample against patch processing to detect malicious behaviors. Another advantage of such methods, including the proposed one, is that they only require access to clean samples, which is a more practical setting for defending backdoor attacks.

3 Backdoor Attacks on ViT

3.1 Threat Model

We follow the typical threat model of DNN backdoor attacks (Gu et al. 2019) that a user wishes to establish a model for a specific image classification task by training with data provided by a third party. We assume the adversary has the capability of injecting poisoned data samples into the training dataset, but cannot modify the model architecture, the training setting, or the inference pipeline. Since the user will

(a) GTSRB/BadNets (b) TinyImageNet/ReFool

Figure 1: Clean and backdoor samples with local patch-based trigger (a square in bottom right corner) and global blending-based trigger (an embedded reflection).

check the accuracy of the trained model on a held-out validation dataset (clean samples), the adversarial goal is to embed a backdoor into the model through data contamination without degrading the clean-data accuracy over the image classification task. In other words, the model should produce malicious behavior only on images with the trigger for the backdoor, while performing normally otherwise.

3.2 Attack Experimental Results

To understand the security threat on ViTs against the backdoor attacks, we consider two most popular approaches of creating the backdoor triggers: local patch-based triggers, BadNets (Gu et al. 2019) and SinglePixel (Bagdasaryan and Shmatikov 2021), and global blending-based triggers, ReFool (Liu et al. 2020) and WaNet (Nguyen and Tran 2021). We evaluate the performance on CIFAR10, GTSRB, and TinyImageNet datasets.

Specifically, we perform the attack experiment by poisoning the training dataset and the corresponding ground-truth labels. For each training dataset, similar to prior works (Gu et al. 2019; Nguyen and Tran 2021; Liu et al. 2020), we select a small number of samples (less than 10%) and apply the corresponding trigger on each of the selected images.

Figure 1 shows some examples of both patch-based and blending-based backdoor samples. The labels of the poisoned samples are also changed to the target label. The poisoned training data are then used to train the image classification model. Then, we perform training using two ViT variants, the original ViT (Dosovitskiy et al. 2021) and DeiT (Touvron et al. 2021), and several other popular CNN model architectures, including Vgg11 (Simonyan and Zisserman 2014), ResNet18 (He et al. 2016), and Big Transfer (BiT) (Kolesnikov et al. 2020). Note that the models are pre-trained on ImageNet-21k and fine-tuned on the corresponding dataset to ensure a consistent experimentation framework. This setup is influenced by the fact that large-scale ViTs and BiT are not trained from scratch on smaller-scale datasets to prevent overfitting. Each trained model is then evaluated on the held-out test sets of clean and backdoor samples. The backdoor samples are applied with the triggers that are generated using the same mechanism in the corresponding attack strategy for the evaluation.

In Tables 1 and 2, we show the clean-data and backdoor-data performance of the trained models for BadNets and WaNet, respectively. We can observe that the trained ViT and DeiT with the backdoors have similar, high clean-data accuracies to that of the corresponding benign models (still outperforming other CNN models). However, when the triggers are present, the probabilities of the poisoned ViT models to predict the target label (i.e., ASR) are also quite high, which are above 96% on all datasets. In other words, ViTs are at least as vulnerable against backdoor attacks as the CNN models. In fact, the patch-based backdoor attack on ViTs seems to be even slightly more successful than on other CNN models, which further validates the need for studying the backdoor attacks and countermeasures on ViTs. We observe similar results for SinglePixel and ReFool attacks.

Dataset	ViT		DeiT		Vgg11		BiT	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
CIFAR10	98.93	98.47	98.82	97.82	93.44	96.95	98.51	97.09
GTSRB	98.68	96.46	98.55	95.62	98.05	91.21	98.71	94.77
T-Imagenet	86.46	98.02	87.76	95.77	61.94	88.57	80.99	96.94

Table 1: Patch-based Backdoor Attack (BadNets)

Dataset	ViT		DeiT		Vgg11		BiT	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
CIFAR10	97.88	99.98	97.92	99.99	95.06	99.71	97.85	99.99
GTSRB	99.08	99.74	97.55	98.27	98.75	99.48	99.23	99.98
T-Imagenet	77.48	99.99	83.90	98.53	64.96	99.21	75.90	99.99

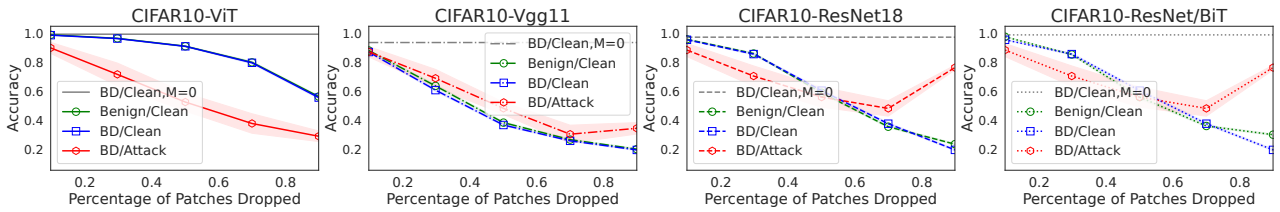
Table 2: Blending-based Backdoor Attack (WaNet)

4 Backdoor Attacks vs. Patch Processing

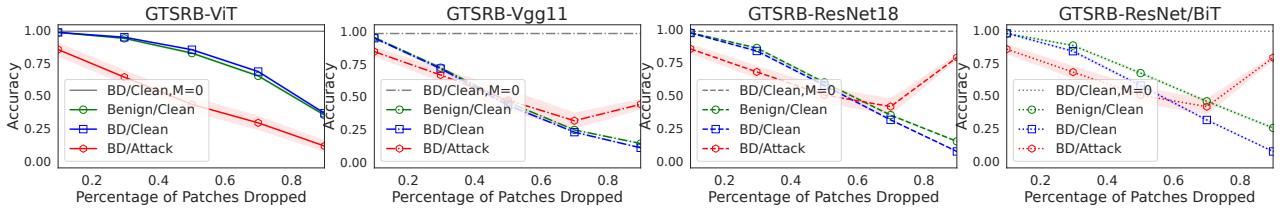
We have shown that backdoor attacks are still quite successful on ViTs. Besides, as we discussed above, it has also recently been shown that ViTs are vulnerable against other types of attacks, although they exhibit certain degrees of improvement in robustness against the transferability of adversarial examples (Mahmood, Mahmood, and van Dijk 2021; Shao et al. 2021). While these features of ViTs are similar to the CNNs, ViTs have also been shown to be more robust toward occlusions, distributional shifts, and permutation (Naseer et al. 2021). Here, we extend the robustness study of the receptive fields of ViTs with respect to the backdoor attack models and compare their performance to CNNs.

4.1 Patch Processing

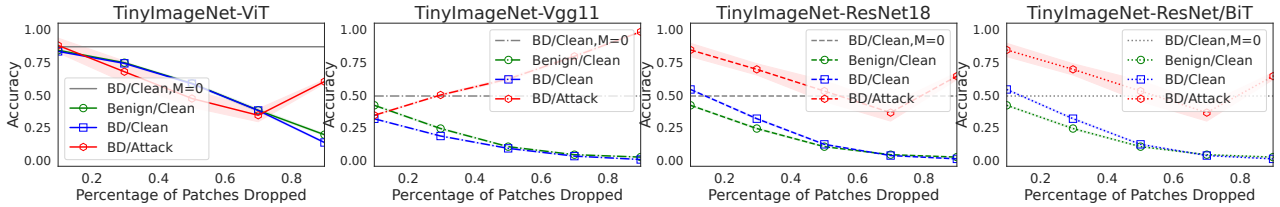
Following the existing defensive methods that process images at the input space for detecting backdoor attacks (Liu, Xie, and Srivastava 2017; Li et al. 2020; Doan, Abbasnejad, and Ranasinghe 2020; Udeshi et al. 2022; Qiu et al. 2021; Gao et al. 2019), we study the performance of the backdoor attacks on ViT models through input transformations that align with the characteristic of ViTs, i.e., patch processing where the content of the image is randomly perturbed. Specifically, each input image x is represented as a sequence of patches with L elements: $\{x_i\}_{i=1,..,L}$. Note that the patch x_i does not necessarily have the same size as the patch size used in the pre-trained ViT model. Perturbing the image's patches is equivalent to modifying its content. Here, we focus on the question: *How does perturbation influence the receptive field of ViTs on image patches when various backdoor triggers are present*? We denote the patch processing on x with a function R and consider the following strategies for performing the patch processing:



(a) CIFAR10



(b) GTSRB



(c) TinyImageNet

Figure 2: Performance of clean-data accuracy and backdoor attacks with dropped patches on ViT, Vgg11, ResNet18, and BiT.

- **PatchDrop.** Similar to Naseer et al. (2021), we randomly drop M patches from the total L patches of an image x. We divide the image into $L = l \times l$ patches that belong to a spatial grid of $l \times l$. The number of dropped patches indicates the information loss on the image content.
- **PatchShuffle.** We randomly shuffle the L patches of an image x. The L patches are created in similar spatial grids as those of PatchDrop. PatchShuffle does not remove the content of the image but can significantly impact the receptive fields of the models.

Note that similar forms of patch transformations on ViTs have been considered in prior works (Naseer et al. 2021; Shao et al. 2021), but not in the context of backdoor attacks.

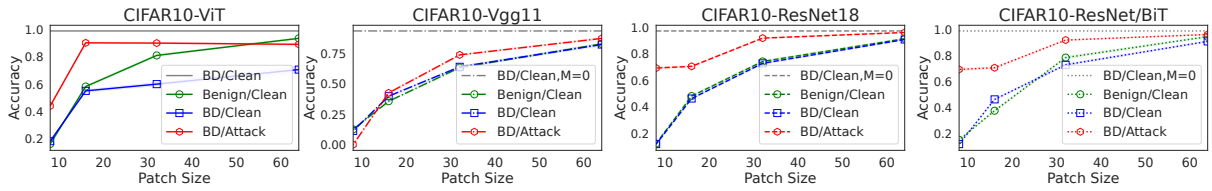
4.2 Performance of Backdoor Attacks against Patch Processing

We first study the trends of backdoor ASR and clean-data accuracy with respect to patch processing on the corresponding test set for each dataset. The results are reported in Figures 2 and 3 for BadNets and ReFool, respectively.

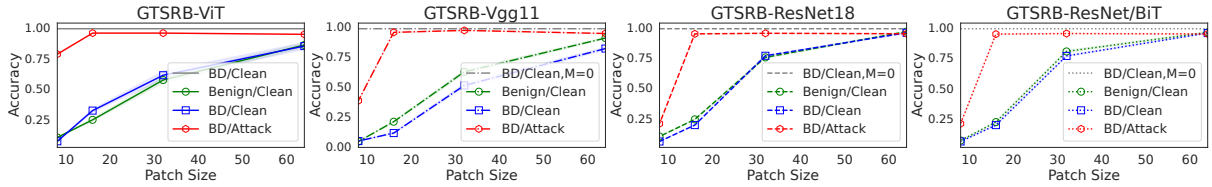
For patch-based attacks with PatchDrop, we observe that the clean-data performances of ViT only drop slightly on CIFAR10 and GTSRB even when almost 50% of the image content is removed. In contrast, the clean-data performances drop much more significantly in all the other three CNNs. On TinyImageNet, the clean-data performance of ViT drops more than in the other datasets. However, when the backdoor triggers are present, the attack success rate on ViT decreases significantly, even with a slight loss in the content of the images. In comparison, backdoor attacks on CNNs are more robust to PatchDrop. Interestingly, if we continue to drop more patches, the ASR on the CNNs suddenly increases in several experiments. A possible explanation is that CNN models and backdoor attacks rely on smaller regions of the image than ViT for prediction and achieving the target classes, respectively, which makes the clean-data accuracy of ViT more robust to patch processing. We also notice another important result: the variance in the predictions of the poisoned models is higher for backdoor samples than for the clean samples. We summarize the observations for patch-based attacks with respect to PatchDrop as follows:

- Clean-data accuracy sensitivity: ViT $<<$ CNN
- ASR sensitivity: ViT $>$ CNN
- **Gap between accuracy and ASR: ViT $>$ CNN**

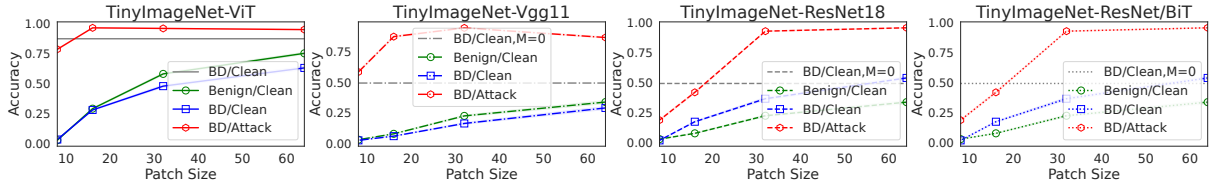
However, for blending-based attacks with PatchDrop, we do not observe a consistent difference between the ViTs and CNNs, although ViTs are more robust with respect to the clean-data accuracy and ASR. Since the trigger is well-blended into the images across the entire pixel space, as in ReFool and WaNet, PatchDrop tends to be less impactful on the backdoor, similar to the robustness of the models on the foreground objects. However, for blending-based attacks with PatchShuffle, we observe that the clean-data performances of ViT drop significantly. In contrast, the ASRs only drop slightly. Such robustness of the trigger is consistent across various patch sizes (i.e., $|x_i|$) on all datasets. For the CNNs, the gaps between clean-data accuracy and ASR are smaller; in some cases, e.g., Vgg11, the gap can become significantly narrow. In previous studies, ViTs exhibit high robustness against patch transformation for larger patch



(a) CIFAR10



(b) GTSRB



(c) TinyImageNet

Figure 3: Performance of clean-data accuracy and backdoor attacks with dropped patches on ViT, Vgg11, ResNet18, and BiT.

sizes (Naseer et al. 2021). Under the proposed PatchShuffle, the significant robustness of the trigger across all patch sizes, especially the smaller sizes, is interesting. Such performance can possibly be explained that ViTs learn and generalize the spatial invariance of the triggers extremely well. We summarize the observations for blending-based attacks with respect to PatchShuffle as follows:

- Clean-data accuracy sensitivity: ViT $>$ CNN
- ASR sensitivity: ViT $<<$ CNN
- **Gap between accuracy and ASR: ViT $>$ CNN**

In summary, ViT has distinguishable performance between clean-data performance and ASR against certain patch processing techniques: the ASR drops significantly on ViT for **patch-based attacks with PatchDrop**, while the clean-data performance drops significantly on ViT for **blending-based trigger attacks with PatchShuffle**. As a result, for both cases, ViT has a larger gap between accuracy and ASR than CNN. It is important to note that these characteristics are not observed on CNN models. Therefore, the observed impact of patch processing against backdoor attacks is unique to ViT.

5 Novel Defensive Solution for ViT

5.1 Methodology

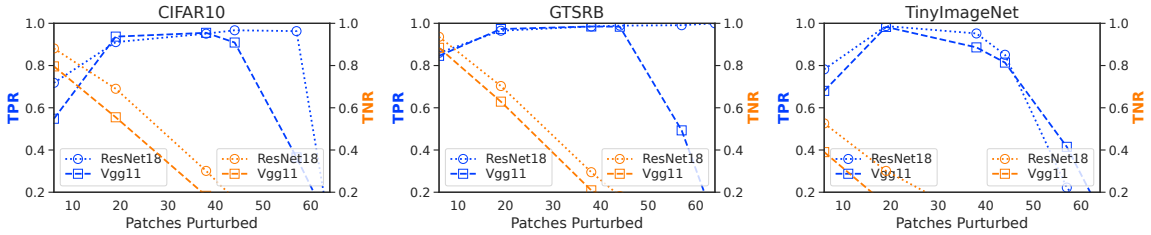
Based on our observations above, we propose an effective backdoor detection algorithm that can successfully detect and then remove the poison samples from a backdoor-injected ViT model with high success rates. The key intuition in our algorithm is that the patch processing strategies affect ViTs' predictions on the backdoor samples differently from the predictive function of the model on the clean data. Our defense algorithm exploits the frequency that ViTs change their predictions on the same sample under different trials of a patch processing strategy and use a threshold to assess if a sample is clean or poisoned. Our defense mechanism only requires access to a small set of K clean samples (less than 1000 on the studied datasets), which can be easily obtained from the held-out validation dataset, for selecting the threshold. When no such clean samples are available, we show that the defenses are still very effective, which enables much wider applicability of the proposed method. The proposed defense consists of the following steps:

- **Step 1 (Offline):** For the small set of clean samples, randomly apply PatchDrop and PatchTranslate on each image for T trials. For each sample x, we calculate $F_d(x) = \sum_{t=1}^{T} \mathbf{1}\{F(x) \neq F(R_d^{(t)}(x))\}$ and

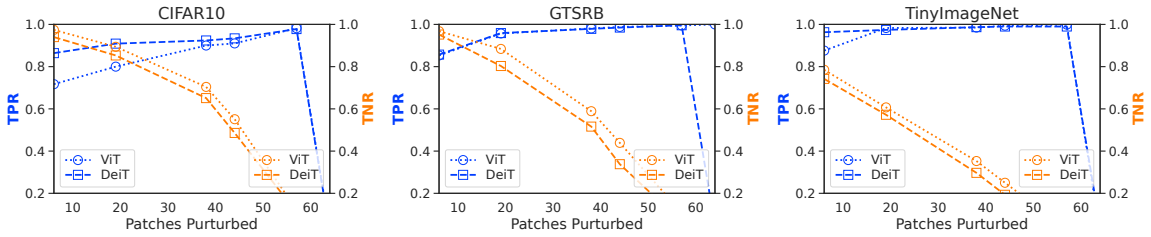
Algorithm 1: Patch Processing-based Backdoor Detection

Input: Sample x, Threshold k_d (PatchDrop), Threshold k_s (PatchShuffle)
Output: Clean or Backdoor Decision

function F(x)
  $t \leftarrow 0$, $F_d(x) \leftarrow 0$, $F_s(x) \leftarrow 0$, Predict $\hat{y} = F(x)$
  repeat
    $t \leftarrow t + 1$
    $\hat{y}_t = F(R_t(x))$, $F_d(x) \leftarrow F_d(x) + 1$ if $\hat{y}_t \neq \hat{y}$
    $\hat{y}_t = F(R_s(x))$, $F_s(x) \leftarrow F_s(x) + 1$ if $\hat{y}_t \neq \hat{y}$
  until $t = T$
  return $F_d(x)$ and $F_s(x)$
end function
If $F_d(x) > k_d$ or $F_s(x) < k_s$, x is *Backdoor*
Otherwise, x is *Clean*



(a) CNN Models



(b) ViT Models

Figure 4: TPR and TNR for different numbers of dropped patches (in a spatial grid of 8×8) for CNNs (ResNet18 and Vgg11) and ViTs (ViT and DeiT). TPR represents the detection rate; TNR represents clean-sample mis-detection rate.

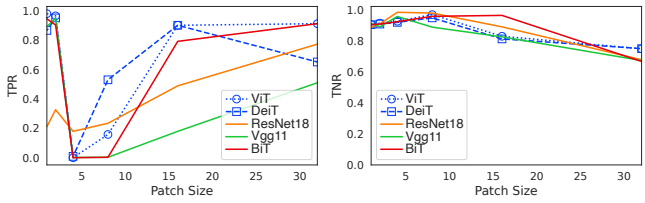
$F_s(x) = \sum_{t=1}^{T} \mathbf{1}\{F(x) \neq F(R_s^{(t)}(x))\}$, where $R_d^{(t)}$ and $R_s^{(t)}$ denotes the random application of PatchDrop and PatchShuffle, respectively, at trial t. Intuitively, $F_d(x)$ and $F_s(x)$ estimate the probabilities that the predicted labels on x change to something else after the patch processing.

- **Step 2 (Offline):** Given the sample $\{F_d(x_i)\}_{i=1,..,K}$ or $\{F_s(x_i)\}_{i=1,..,K}$ created in Step 1, we set the threshold parameter k_d and k_s for PatchDrop and PatchShuffle, respectively, to the values at the n^{th} percentiles to ensure a small false positive rate, as follows:
  - For PatchDrop, we typically select a large value (e.g., 90^{th} percentile). This is because ASRs significantly decrease under patch processing such as PatchDrop.
  - For PatchShuffle, we typically select a small value (e.g., 10^{th} percentile). This is because ASRs do not drop under patch processing such as PatchShuffle while the clean-data accuracies are more affected.
- **Step 3 (During Inference):** For a sample, we randomly apply PatchDrop and PatchShuffle for T trials and record the number of label changes, $F_d(x)$ and $F_s(x)$, respectively. If $F_d(x)$ is greater than the selected k_d threshold for PatchDrop or $F_s(x)$ is smaller than the selected k_s threshold for PatchShuffle, we flag the sample as a backdoor sample. Otherwise, x is determined as a clean sample.

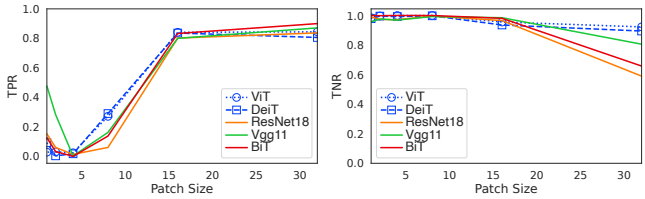
Note that, the proposed approach does not assume the knowledge of the type of the backdoor attack, which ensures its practicality in various scenarios. Furthermore, when the model is benign, i.e., without the backdoor attack, because of the percentile selection rules, only a very small fraction of samples will be identified as false negative. Formally, our defense approach follows a similar strategy as that of an anomaly detector. Thus, more sophisticated anomaly detection approaches can be used to improve the detection rate while keeping the false negative rate low; however, this is beyond the scope of this paper. The details of the detection algorithm are presented in Algorithm 1.

5.2 Analysis of Patch Processing-based Defense

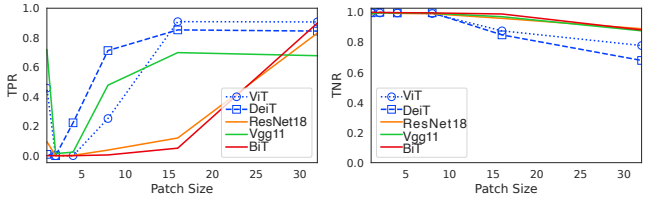
We first provide a qualitative analysis of the proposed defense strategy for detecting both patch-based and blending-



(a) CIFAR10



(b) GTSRB



(c) TinyImageNet

Figure 5: TPR and TNR for different sizes of processed patches for CNN models (ResNet18 and Vgg11) and ViTs (ViT and DeiT) under ReFool backdoor attack.

Dataset	ViT		DeiT		Vgg11		ResNet18		BiT	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
CIFAR10	90.08	99.48	**91.88**	96.18	88.12	62.88	88.80	89.33	90.00	87.72
GTSRB	**94.89**	98.78	93.62	97.66	20.15	80.70	93.80	89.99	93.89	92.91
TinyImageNet	95.80	64.75	95.80	64.73	81.30	20.51	**99.00**	56.48	98.60	42.64

Table 3: TPR (best bolded) and TNR (best underlined) of detecting backdoor samples in BadNets' poisoned models.

Dataset	ViT		DeiT		Vgg11		ResNet18		BiT	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
CIFAR10	**90.00**	83.20	66.80	81.50	18.00	82.80	48.90	89.30	79.10	66.90
GTSRB	**84.10**	95.60	83.60	93.60	80.00	98.50	80.20	96.60	83.30	92.10
TinyImageNet	**90.90**	87.70	85.30	85.00	69.90	97.10	12.00	96.00	5.10	98.90

Table 4: TPR (best bolded) and TNR (best underlined) of detecting backdoor samples in ReFool's poisoned models.

based backdoor samples from the corresponding backdoor-injected ViT and CNN models.

Patch-based Attacks Figure 4 illustrates the true positive rate (TPR) and true negative rate (TNR) for the patch-based attack, BadNets, when varying the number of dropped patches d in PatchDrop when the spatial grid is 8×8. Recall that the TPR and TNR indicate the backdoor detection rate and the percentage of clean samples that are not falsely detected as backdoor samples, respectively. As we can observe, the defensive solution with PatchDrop works better for ViT models than for CNN models such as ResNet18 and Vgg11. Furthermore, dropping 10% of the patches can consistently achieve higher TPR and TNR across different datasets. The effectiveness of this defense on ViTs is because the backdoor performance is more sensitive to PatchDrop, as discussed in the previous section.

Blending-based Attacks Figure 5 illustrates the TPR and TNR when defending against ReFool with various sizes of the processed patches in PatchShuffle. As we can observe, PatchShuffle generally achieves higher TPRs in ViTs than in CNN models. More importantly, when the patch size is similar to that of the trained patch size in ViTs, defending against ViTs is consistently effective.

6 Defense Experimental Results

This section presents the empirical results in defending against the backdoor attacks. In real-world settings, the defender does not know which attack is performed by the adversary. To this end, we consider two practical scenarios.

In the first scenario, the backdoor is successfully injected into the trained model and the victim defends against backdoor attacks (i.e., alleviates its effectiveness) by filtering the backdoor samples during inference. In this experiment, TPR and TNR are reported, as they demonstrate how likely the defense method identifies the backdoor samples and how likely the clean samples are not falsely flagged as backdoor samples, respectively. We also assume that a small set of clean samples are available. The values at the 90^{th} and 10^{th} of the empirical distributions of $F_d(x)$ and $F_s(x)$, for all clean samples x, are selected as the thresholds k_d and k_s for PatchDrop and PatchShuffle, respectively.

In the second scenario, we consider an extreme case where the defender is also the model trainer who receives a possibly poisoned training dataset. The defender aims to obtain the trained model that is free of the backdoor. Here, the clean samples are not available, which makes the defending task very difficult. Defending using the proposed defensive solution consists of the following steps:

(i) train the model on the training dataset for some epochs,

(ii) use the possibly poisoned, trained model to detect the backdoor samples, and

(iii) remove the backdoor samples from the training dataset and re-train the model on the filtered dataset.

We observe that training the model for 50 epochs in step (i) is sufficient for the backdoor to be inserted into the model if the training dataset is poisoned and for the clean-data accuracy to reach an acceptable performance compared to its optimal value (a few percents difference, e.g., $> 90\%$ in CIFAR10). Ideally, we train the models until they reach the optimal accuracies, but this can add significant computation to the training process while only adding a minor improvement in the defense. Therefore, we use 50 epochs on all experiments. In step (ii), we consider the threshold $k_d = 0$ for PatchDrop, and $k_s = T$ for PatchShuffle.

6.1 Defending against the Poisoned Model

Table 3 and Table 4 present the defense results when the defender aims to detect whether a sample is a backdoor or clean sample during inference under the local patch-based attack, BadNets, and the global blending-based attack, ReFool, respectively. As we can observe, the proposed defensive solution achieves comparable TPRs (i.e., successfully detects the backdoor samples) in both ViTs (>90%) and CNN models (>88%) across different datasets under BadNets attacks. However, the TNRs of ViTs, including ViT and DeiT, are significantly better than those of CNN models, including Vgg11, ResNet18 and BiT. Specifically, the proposed defense method only falsely detects clean samples as backdoor samples less than 3% of the time in the trained ViTs, but more than 10% of the time in the trained CNN models. Under ReFool attacks, the defensive solution achieves the best TPR for ViT, while its TNRs are also very high. While the TNRs of ResNet18 and BiT are higher than those of ViTs, their TPRs are significantly lower, especially in the larger-scale TinyImageNet dataset. Overall, we can conclude that the proposed defense method is consistently more effective in ViTs than in CNN models.

Dataset	ViT		DeiT		ResNet18		BiT	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
CIFAR10	**98.94**	10.01	98.74	09.96	92.30	15.77	97.10	10.39
	+0.01	*–89.8*	*–0.08*	*–89.8*	*–4.53*	*–88.9*	*–1.43*	*–89.3*
GTSRB	**98.38**	0.48	97.98	0.47	96.58	0.46	96.89	0.48
	–0.31	*–99.51*	*–0.58*	*–99.5*	*–2.31*	*–99.5*	*–1.85*	*–99.5*
T-Imagenet	85.57	0.51	**88.77**	0.50	64.79	0.55	72.34	0.52
	–1.03	*–99.4*	*+1.15*	*–99.4*	*–5.65*	*–99.4*	*–10.6*	*–99.4*

Table 5: Clean-data accuracy (beset bolded) and ASR after removing the backdoor samples and retraining the models. *Italicized* values are relative changes w.r.t the models trained without removing backdoor samples.

6.2 Defending against the Poisoned Training Data

We present the clean-data accuracies and ASRs after retraining the models on the filtered data, as described in the second scenario, in Table 5. The attack method is patch-based. We can observe that the proposed defense method successfully reduces the ASRs much closer to ASRs of random guesses in both ViTs and CNNs on all datasets. However, in ViTs, the clean-data accuracies are preserved, while in CNN models, the clean-data accuracies drop more than 4.5% for ResNet18 and almost 1.5% for BiT. The results for Vgg11 are worse than those of ResNet18 and BiT and are reported in supplement materials. As discussed in the previous experiment, a non-trivial number of clean samples can be falsely detected as backdoor samples in CNN models using the proposed patch-processing approach. Thus, while most backdoor samples are removed from the training datasets, the number of clean training samples is also reduced, which leads to the drop in clean-data performance in CNN models. We can also notice that by employing a large-scale pre-trained model (i.e., BiT), the drop in performance can be mitigated compared to smaller models, such as ResNet18 and Vgg11. Nevertheless, we can still observe that the proposed defense is more effective for ViTs than for CNN models.

In conclusion, while ViT is vulnerable to patch-based backdoor attacks, the proposed simple-yet-effective patch-processing-based defense can detect backdoor samples with a high detection rate while maintaining a low FNR. Because ViT is robust against patch processing on the clean data, processing the images with these strategies can be utilized to obtain useful yet tangible traces for effectively distinguishing the predictions between the clean and backdoor samples.

7 Conclusion

This paper studied several aspects of backdoor attacks against ViT. We first perform an empirical study on the vulnerability of ViT against both patch-based and blending-based backdoor attacks. Then, based upon our observation that ViT exhibits distinguishable performance between clean samples and backdoor samples against patch processing, we proposed a novel defensive solution to counter backdoor attacks on ViT, which is able to reduce the backdoor attack success rate significantly. Two patch processing methods are investigated. The effectiveness of the proposed techniques is comprehensively evaluated. To the best of our knowledge, this paper presented the first defensive strategy that utilizes a unique characteristic of ViT against backdoor attacks.

Acknowledgments

The research was conducted while all authors worked at Baidu Cognitive Computing Lab – 10900 NE 8th St. Bellevue, WA 98004, USA.

References

Akhtar, N.; and Mian, A. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6: 14410–14430.

Alimadadi, A.; Aryal, S.; Manandhar, I.; Munroe, P. B.; Joe, B.; and Cheng, X. 2020. Artificial intelligence and machine learning to fight COVID-19. *Physiological Genomics*, 52(4): 200–202.

Bagdasaryan, E.; and Shmatikov, V. 2021. Blind Backdoors in Deep Learning Models. In *Proceedings of the 30th USENIX Security Symposium, (USENIX Security)*, 1505–1521.

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)*, 101–105. Taipei.

Benz, P.; Ham, S.; Zhang, C.; Karjauv, A.; and Kweon, I. S. 2021. Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs. In *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, 25. Online.

Berman, D. S.; Buczak, A. L.; Chavis, J. S.; and Corbett, C. L. 2019. A survey of deep learning methods for cyber security. *Information*, 10(4): 122.

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding Robustness of Transformers for Image Classification. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10211–10221. Montreal, Canada.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628. Seattle, WA.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I. M.; and Srivastava, B. 2019a. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In *Proceedings of the Workshop on Artificial Intelligence Safety*. Honolulu, HI.

Chen, H.; Fu, C.; Zhao, J.; and Koushanfar, F. 2019b. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 4658–4664. Macao, China.

Chen, S.; Yu, T.; and Li, P. 2021. MVT: Multi-view Vision Transformer for 3D Object Recognition. In *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, 349. Online.

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*.

Cheng, H.; Xu, K.; Liu, S.; Chen, P.-Y.; Zhao, P.; and Lin, X. 2020. Defending against Backdoor Attack on Deep Neural Networks. *arXiv preprint arXiv:2002.12162*.

Doan, B. G.; Abbasnejad, E.; and Ranasinghe, D. C. 2020. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 897–912. Virtual Event / Austin, TX.

Doan, K.; Lao, Y.; and Li, P. 2021. Backdoor Attack with Imperceptible Input and Latent Modification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 18944–18957. virtual.

Doan, K.; Lao, Y.; and Li, P. 2022. Marksman Backdoor: Backdoor Attacks with Arbitrary Target Class. In *Advances in Neural Information Processing Systems (NeurIPS)*. New Orleans, LA.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. LIRA: Learnable, Imperceptible and Robust Backdoor Attacks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11946–11956. Montreal, Canada.

Doan, K. D.; Yang, P.; and Li, P. 2022. One Loss for Quantization: Deep Hashing with Discrete Wasserstein Distributional Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9447–9457. New Orleans, LA.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. Virtual Event, Austria.

Dumford, J.; and Scheirer, W. J. 2020. Backdooring Convolutional Neural Networks via Targeted Weight Perturbations. In *Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–9. Houston, TX.

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)*, 113–125. San Juan, PR.

Gkelios, S.; Boutalis, Y. S.; and Chatzichristofis, S. A. 2021. Investigating the Vision Transformer Model for Image Retrieval Tasks. In *Proceedings of the 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 367–373. Pafos, Cyprus.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386.

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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV.

Huang, H.; Guo, S.; Gui, G.; Yang, Z.; Zhang, J.; Sari, H.; and Adachi, F. 2020. Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions. *IEEE Wirel. Commun.*, 27(1): 214–222.

Joshi, A.; Jagatap, G.; and Hegde, C. 2021. Adversarial Token Attacks on Vision Transformers. *arXiv preprint:2110.04337*.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2021. Transformers in vision: A survey. *arXiv preprint:2101.01169*.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big transfer (bit): General visual representation learning. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part V*, 491–507. Glasgow, UK.

Lao, Y.; Yang, P.; Zhao, W.; and Li, P. 2022a. Identification for Deep Neural Network: Simply Adjusting Few Weights! In *Proceedings of the 38th IEEE International Conference on Data Engineering (ICDE)*, 1328–1341. Kuala Lumpur, Malaysia.

Lao, Y.; Zhao, W.; Yang, P.; and Li, P. 2022b. DeepAuth: A DNN Authentication Framework by Model-Unique and Fragile Signature Embedding. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*. Virtual.

Li, Y.; Zhai, T.; Wu, B.; Jiang, Y.; Li, Z.; and Xia, S. 2020. Rethinking the Trigger of Backdoor Attack. *arXiv preprint arXiv:2004.04692*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, 273–294. Heraklion, Crete, Greece.

Liu, Q.; Li, P.; Zhao, W.; Cai, W.; Yu, S.; and Leung, V. C. M. 2018a. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access*, 6: 12103–12117.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.; Zhai, J.; Wang, W.; and Zhang, X. 2018b. Trojaning Attack on Neural Networks. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*. San Diego, CA.

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), Part X*, 182–199. Glasgow, UK.

Liu, Y.; Xie, Y.; and Srivastava, A. 2017. Neural Trojans. In *Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD)*, 45–48. Boston, MA.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. Montreal, Canada.

Lv, P.; Ma, H.; Zhou, J.; Liang, R.; Chen, K.; Zhang, S.; and Yang, Y. 2021. DBIA: Data-free Backdoor Injection Attack against Transformer Networks. *arXiv preprint arXiv:2111.11870*.

Mahmood, K.; Mahmood, R.; and van Dijk, M. 2021. On the Robustness of Vision Transformers to Adversarial Examples. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7818–7827. Montreal, Canada.

Mao, X.; Qi, G.; Chen, Y.; Li, X.; Duan, R.; Ye, S.; He, Y.; and Xue, H. 2022. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12042–12051. New Orleans, LA.

Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M. 2021. Intriguing Properties of Vision Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 23296–23308. virtual.

Naseer, M.; Ranasinghe, K.; Khan, S.; Khan, F. S.; and Porikli, F. 2022. On Improving Adversarial Transferability of Vision Transformers. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*. Virtual Event.

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)*. Virtual Event, Austria.

Oh, Y.; Park, S.; and Ye, J. C. 2020. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Trans. Medical Imaging*, 39(8): 2688–2700.

Qiao, X.; Yang, Y.; and Li, H. 2019. Defending Neural Backdoors via Generative Distribution Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 14004–14013. Vancouver, Canada.

Qiu, H.; Zeng, Y.; Guo, S.; Zhang, T.; Qiu, M.; and Thuraisingham, B. M. 2021. DeepSweep: An Evaluation Framework for Mitigating DNN Backdoor Attacks using Data Augmentation. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIA CCS)*, 363–377. Virtual Event, Hong Kong.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden Trigger Backdoor Attacks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 11957–11965. New York, NY.

Shao, R.; Shi, Z.; Yi, J.; Chen, P.-Y.; and Hsieh, C.-J. 2021. On the adversarial robustness of visual transformers. *arXiv preprint:2103.15670*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, Canada.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 10347–10357. Virtual Event.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral Signatures in Backdoor Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 8011–8021. Montréal, Canada.

Udeshi, S.; Peng, S.; Woo, G.; Loh, L.; Rawshan, L.; and Chattopadhyay, S. 2022. Model Agnostic Defence Against Backdoor Attacks in Machine Learning. *IEEE Trans. Reliab.*, 71(2): 880–895.

Vinayakumar, R.; Alazab, M.; Soman, K.; Poornachandran, P.; Al-Nemrat, A.; and Venkatraman, S. 2019. Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7: 41525–41550.

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 (SP)*, 707–723. San Francisco, CA.

Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 548–558. Montreal, Canada.

Yang, P.; Lao, Y.; and Li, P. 2021. Robust Watermarking for Deep Neural Networks via Bi-level Optimization. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14821–14830. Montreal, Canada.

Yao, Y.; Li, H.; Zheng, H.; and Zhao, B. Y. 2019. Latent Backdoor Attacks on Deep Neural Networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2041–2055. London, UK.

Yu, T.; and Li, P. 2022. Degenerate Swin to Win: Plain Window-based Transformer without Sophisticated Operations. *arXiv preprint arXiv:2211.14255*.

Yu, T.; Zhao, G.; Li, P.; and Yu, Y. 2022. BOAT: Bilateral Local Attention Vision Transformer. In *Proceedings of the 33rd British Machine Vision Conference (BMVC)*. London, UK.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F. E. H.; Feng, J.; and Yan, S. 2021. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 538–547. Montreal, Canada.

Zhao, B.; and Lao, Y. 2022. CLPA: Clean-label poisoning availability attacks using generative adversarial nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9162–9170.

Zhao, W.; Lao, Y.; and Li, P. 2022. Integrity Authentication in Tree Models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2585–2593. Washington, DC.