You Are Catching My Attention: Are Vision Transformers Bad Learners under Backdoor Attacks?

Zenghui Yuan[1] Pan Zhou[1*] Kai Zou[2] Yu Cheng[3]
[1]Hubei Key Laboratory of Distributed System Security,
Hubei Engineering Research Center on Big Data Security,
School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Protagolabs Inc [3]Microsoft Research
{zenghuiyuan, panzhou}@hust.edu.cn, kz@protagolabs.com, yu.cheng@microsoft.com

Abstract

Vision Transformers (ViTs), which made a splash in the field of computer vision (CV), have shaken the dominance of convolutional neural networks (CNNs). However, in the process of industrializing ViTs, backdoor attacks have brought severe challenges to security. The success of ViTs benefits from the self-attention mechanism. However, compared with CNNs, we find that this mechanism of capturing global information within patches makes ViTs more sensitive to patch-wise triggers. Under such observations, we delicately design a novel backdoor attack framework for ViTs, dubbed BadViT, which utilizes a universal patch-wise trigger to catch the model's attention from patches beneficial for classification to those with triggers, thereby manipulating the mechanism on which ViTs survive to confuse itself. Furthermore, we propose invisible variants of BadViT to increase the stealth of the attack by limiting the strength of the trigger perturbation. Through a large number of experiments, it is proved that BadViT is an efficient backdoor attack method against ViTs, which is less dependent on the number of poisons, with satisfactory convergence, and is transferable for downstream tasks. Furthermore, the risks inside of ViTs to backdoor attacks are also explored from the perspective of existing advanced defense schemes.

1. Introduction

Transformers, which are all-powerful in the field of natural language processing (NLP) [6, 13, 57], have recently set off a wave of frenetic research in computer vision (CV). Thanks to the self-attention mechanism, vision transformers (ViTs) have broken the perennial domination of convolutional neural networks (CNNs) [17], and have been developed in hot areas like image classification [36, 53, 66], object detection [3, 7] and semantic segmentation [46, 64]. The architecture optimization researches of ViTs are also continuously improving the performance and efficiency [26, 40, 45, 53], and providing vitality for advancing the deployment of ViTs in industry.

Unfortunately, manifold threats in deep learning pose severe challenges to ViTs. For instance, adversarial attacks [10, 11, 25, 28, 31, 32, 49, 50, 76] confuse the deep model to make wrong predictions by adding subtle perturbations to the input. In addition, backdoor attacks are also extremely threatening to deep models [22, 27, 48, 51, 68]. More and more deep learning tasks are "outsourced" training or directly fine-tuning on pre-trained models [22, 75], allowing attackers to implant backdoors into the model by establishing a strong association between the trigger and the attack behavior. In response, a growing number of researchers have paid attention to the security of ViTs under adversarial attacks [1, 4, 20, 39, 44] and backdoor attacks [16, 37, 47, 73]. However, previous ViT backdoor works have not systematically compared with CNN to elucidate the vulnerability source of ViTs to backdoor attacks, and have not considered balancing attack concealment and attack benefit, so that triggers can be easily detected by the naked eye.

To fill this gap, we systemically discuss the robustness of ViTs and CNNs under basic backdoor attacks with different trigger settings and find that ViTs seem to be more vulnerable to patch-wise triggers rather than image-blending triggers. Delving into the essence of ViTs, images are divided into patches as tokens to calculate attentions, which can capture more interaction information between patches at the global level than CNNs [44]. Thus the patch-wise perturbation has been shown to sufficiently affect the self-attention mechanism of ViTs [20] and make ViTs weaker learners than CNNs. Inspired by this, a natural and interesting question is whether backdoor attacks with the patch-

*Corresponding author.

wise trigger are resultful in ViTs. Accordingly, we propose BadViT, a well-designed backdoor attack framework against ViTs. Through the optimization process, the universal adversarial patch is generated as a trigger, which can be better caught by the model through the self-attention mechanism of ViTs to tighten the connection between the trigger and the target class. To achieve invisible attacks, we limit the perturbation strength of the adversarial patch-wise trigger and adopt the blending strategy [12] instead of pasting. Moreover, we adopt a ViTs backdoor defense PatchDrop [16], and two state-of-the-art defenses in CNNs, Neural Cleanse [59] and FinePruning [33], to explore the vulnerability of ViTs against our BadViT.

Our main contributions are as follows:

- We explore the robustness of ViTs compared with CNNs against backdoor attacks with different triggers.
- We propose our BadViT as well as its invisible version and verify the validity through abundant experiments.
- We show the effect of BadViTs under three advanced defense methods, and further discuss the characteristics of ViTs under backdoor attacks through patch processing, reverse engineering, and pruning.

We believe this paper will offer readers a new understanding of the robustness of ViTs against backdoor attacks, and provide constructive insights into the follow-up ViTs system optimization and backdoor defense efforts.

2. Related Works

2.1. Vision Transformer

Benefiting from the self-attention mechanism, the transformer model has been continuously developed and acquired significant achievements in the field of CV recently. In the first work [17], the transformer encoder is utilized to perform attention calculation on the image patches divided into equal sizes as tokens, and obtain the classification output through a multi-layer perceptron (MLP). On this basis, several works [36, 53, 60, 62, 66] have been carefully crafted on the model architecture for performance gains. In addition to image classification, ViTs also shine in other CV tasks such as object detection [3, 7], semantic segmentation [46, 64] and image quality assessment [9, 65].

2.2. Backdoor Attacks and Defenses

Inspired by the over-learning ability of deep models, a series of studies have been conducted on how to set appropriate triggers to implant backdoors in deep models. BadNets [22] is the first backdoor attack against Deep Neural Networks (DNNs), which constructs backdoor inputs by pasting triggers on a proportion of randomly selected images and modifying their labels to specific target classes. Then injected model can correctly label the benign input and misclassify the input with triggers as the target class. Another effective backdoor attack method [34] takes activating of key neurons as the optimization goal to generate triggers in reverse. And the backdoor training progress effectively establishes the association between triggers and corresponding neurons. Based on these works, a mass of researchers have developed more powerful backdoor attacks from different aspects. For concealment, invisible backdoor attacks can be optimized by applying different techniques to triggers [12, 29, 30, 35, 72, 74], without modifying the labels of backdoor inputs [43, 55, 71], or manipulating the training process of the model [2]. In addition, some non-data-poisoning-based studies achieve backdoor implantation by tampering with the model, such as directly modifying model parameters [18, 69] and trojan implants [41, 52].

In order to deal with advanced attacks, research on backdoor defense is also constantly improving. Existing backdoor defenses can be divided into experience-based defenses [8, 15, 23, 56, 70] and provable defenses [58, 61, 63] according to whether their performance can be provable in theory. In particular, as a highly influential defense method, Neural Cleanse [59] is often adopted for backdoor detection and identification, which can generate triggers for backdoor models based on reverse engineering in black-box scenarios. An anomaly index of different classes is compared to obtain the target class of the attacker. In contrast, the pruning-based defense method [14] aims to suppress the backdoor neurons in the infected model to eliminate the backdoor. To solve the drop in model accuracy caused by excessive pruning, [33] combines pruning with fine-tuning technology, which greatly improves the defense effect.

2.3. The Robustness of Vision Transformers

In addition to performance and structural optimization, the robustness of ViTs has also attracted more attention from researchers. Earlier work [1, 4, 5, 39, 44] proposed that the ViT model is more robust than CNNs models under adversarial attacks. However, [20] demonstrated that adding patch-wise perturbations based on the self-attention mechanism can effectively reduce the accuracy of ViTs, which triggered a rethinking of its robustness. In terms of backdoor attacks, [37] proposed an attack method named DBIA without acquiring a specific dataset, and implanted backdoors into specific neurons by maximizing the attention of the trigger region in inputs. TrojViT [73] added patch-based triggers to the clean image, and also used the self-attention mechanism to attack a small number of parameters of the model on DRAM memory. And [16, 47] considered the vulnerability of ViTs under backdoor attack, and made defense based on patch processing operation. Unlike existing works, our BadViT is basically a kind of data poisoning attack and is more stealthy, versatile, and threatening.

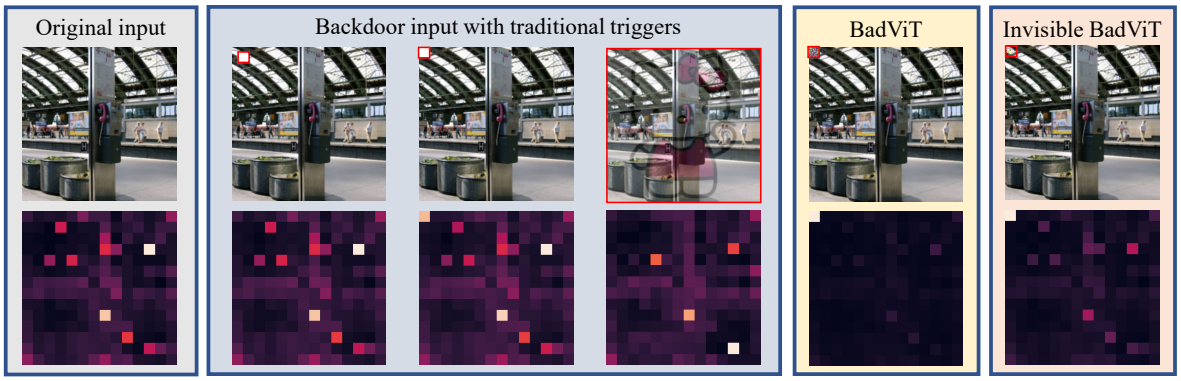


Figure 1. Visualization of the average attention maps of all layers in DeiT-T for clean and backdoor images with different trigger settings. The location of the trigger is marked with a red box. Note that brighter colors in the attention map mean higher attention scores for the corresponding image patch. In comparison, triggers of the proposed BadViT availably catch the model's attention on the patch-wise trigger.

3. Backdoor Attacks in ViTs

In this section, we set up a threat model for backdoor attacks of ViTs. We also introduce the ViTs and the formulation of attacks in ViTs.

3.1. Threat Model

Considering that open-source pre-training ViT models are mostly adopted for fine-tuning in different applications, we refer to the attack settings in prior works [22]. It is assumed that the attacker can obtain the complete model architecture and parameters, as well as part of the dataset used for pre-training. However, based on the security of the training platform itself, we set that the attacker cannot tamper with the model training schedule, which means that the attacker is unable to achieve the attack goal by modifying the loss or manipulating the gradient like [2]. Consequently, backdoor attacks are in the way of "data poisoning" in our threat model, by modifying part of the input as well as its ground-truth label, and embedding backdoors into ViTs in the original training schedule.

3.2. Background of Backdoor Attacks in ViTs

Vision transformers. Given a pre-trained vision transformer classifier $\mathcal{F}(\cdot)$ and a benign training set $\mathcal{D}_{tr}$ with N pairs $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^{C\times H\times W}$ denotes the original image with C channels and $H \times W$ pixels, and y_i represents the corresponding ground-truth label. The input image x of ViTs is preprocessed into $H \times W/P^2$ patches with the shape of $P \times P$, and each patch is used as a token to calculate the attention map through the multi-head self-attention (MSA) module as follows:

$$Attention(x) = \text{Softmax}(\frac{xW_Q(xW_K)^T}{\sqrt{d}}xW_V) \quad (1)$$

where W_Q, W_K, and W_V are learnable matrices of the query, key, and value, respectively, and d is the dimension of the key. Based on multiple MSA attention calculations, the output category is finally obtained through an MLP module.

Backdoor attacks. For backdoor attacks, we denote the subset to produce backdoor input as $\mathcal{D}_{bd}$, which is obtained by selecting ρ fraction of the benign training set ($\rho = |\mathcal{D}_{bd}|/|\mathcal{D}_{tr}|$ is a vital metric to measure the effectiveness of attacks). Using $\hat{x_j}$ represent the backdoor input, it is calculated as follows:

$$\hat{x_j} = \mu(x_j, t, loc), \quad if\ y_j \neq y^*, \quad (2)$$

where $\mu(\cdot)$ is the synthesizer of trigger t and the benign input x_j, loc defines the location of trigger in inputs. Moreover, only benign inputs having different labels with the target class will be selected to make backdoor inputs, then the ground-truth label is modified to the target class y^*.

Let $\hat{\mathcal{F}}(\cdot)$ represent the backdoor model. As to the attacker, it is crucial to ensure successful attacks, namely making certain of $\hat{\mathcal{F}}(\hat{x_j}) = y^*$. Meanwhile, in order to guarantee that the backdoor model is not detected with obvious abnormalities, it is supposed to classify the benign input as the ground-truth label following $\hat{\mathcal{F}}(x_j) = y_j$. The attacker can achieve the above two goals through the following optimization process:

$$\min_{\theta} \sum_{x_i \in \mathcal{D}_{cl}} \mathcal{L}_{tr}\left(\mathcal{F}(x_i), y_i\right) + \sum_{\hat{x_j} \in \mathcal{D}_{bd}} \mathcal{L}_{bd}\left(\mathcal{F}(\hat{x_j}), y^*\right), \quad (3)$$

where θ is the parameter of models, $\mathcal{D}_{cl} = \mathcal{D}_{tr} \backslash \mathcal{D}_{bd}$ is the clean subset, $\mathcal{L}_{tr}$ denotes the training loss of the main task, and $\mathcal{L}_{bd}$ represents the backdoor training loss. In this paper, we both use the cross-entropy loss function to follow the normal training schedule for classification tasks.

4. Backdoor Attacks Robustness Comparison

In order to explore the robustness of ViTs and CNNs under backdoor attacks, we follow two classical data-

Table 1. Evaluation of ViTs and CNNs under backdoor attacks with different trigger settings.

Attack Mode		**Patch Trigger Attack**								**Blend Trigger Attack**			
Trigger Setting		16 (0,0)		24 (0,0)		32 (0,0)		16 (8,8)		$\alpha = 0.02$		$\alpha = 0.04$	
Model $\downarrow$	CA	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
ResNet-18	69.10	67.89	91.53	67.53	92.74	67.79	93.53	68.38	92.43	58.68	94.83	66.30	99.22
ResNet-50	76.13	73.18	94.08	72.90	95.53	75.19	95.70	73.25	94.58	69.16	94.73	72.82	99.89
DeiT-T	72.02	70.82	96.29	70.79	97.10	70.91	97.52	67.62	91.07	71.38	21.21	71.78	91.48
DeiT-S	79.71	79.15	96.30	79.12	96.64	79.18	98.75	78.32	94.04	78.86	21.64	79.31	94.81

poisoning methods, namely, patch-based [22] attack and blending attack [12] to investigate the vulnerability in ViTs.

4.1. Attack Settings

We respectively set non-patch-wise triggers (white squares that do not fully cover the image patch, like the 2-nd column in Fig. 1), patch-wise triggers (a white square covered a whole image patch, as seen in the 3-rd column in Fig. 1), and blending-based triggers (a Hello Kitty image with the same size of inputs, as the 4-th column in Fig. 1). We perform tests on the official pre-trained models of DeiT [53] and ResNet [24] on the Imagenet dataset. For the patch trigger, we set white squares of different sizes and different starting positions of the upper left corner. For the blending-based trigger attack, we set the blend ratio $\alpha = 0.02$ and 0.04 in training, and 0.2 in testing. We fine-tune 1 epoch with a learning rate of $1e-5$ and $\rho = 0.1$. We compare metrics including 1) Clean Accuracy (CA): the accuracy of the clean test datasets on clean models, 2) Attack Success Rate (ASR): the proportion of backdoor inputs predicted as target classes, and 3) Backdoor Accuracy (BA): the accuracy of backdoor models for clean test datasets.

4.2. Observations and Discussions

It can be found in Tab. 1 that: 1) ViTs are more robust than CNNs with lower ASRs in blend trigger attacks with both two blend ratios, especially when $\alpha = 0.02$, DeiTs can only predict 21.21% and 21.64% of the backdoor input as the target label; 2) ViTs seem to be more sensitive to patch triggers and achieve higher ASRs and less reduction in BAs compared to CAs than in CNNs. In both CNNs and ViTs, triggers with larger sizes get better attack results; 3) comparing the first and fourth columns of patch trigger attacks, patch-wise triggers (size is 16, starting position is (0, 0)) are better than non-patch-wise triggers (size is 16, starting position is (8, 8)) in DeiTs, while this variation does not exist in ResNets. Synthesizing the attention distribution of ViTs to different triggers observed in Fig. 1, the global blend triggers and the non-patch-wise triggers cannot effectively change the model's attention. While the model's attention score for the area covered by the patch increases significantly with patch-wise triggers. Consequently, the self-attention mechanism makes ViTs more sensitive to patch-wise triggers to effectively implant backdoors in the model. We continue to explore more effective backdoor attacks against ViTs on the patch-wise basis.

5. The Proposed BadViT Framework

In this section, we first outline the inspiration for our approach and then introduce the specific framework of BadViT as well as its invisible variants.

5.1. Inspirations of BadViT

Essentially, the success rate of backdoor attacks depends on the ability of models to capture the correlation between the trigger and target class, so the question we face is **"How to generate a trigger that can more effectively attract the attention of the model?"**. We have gained insight from the investigation in Sec. 4 that ViTs are more vulnerable than CNNs under patch-wise triggers, so we intend to design a universal optimized patch-wise trigger that can fully focus the model's attention on the area where it is located, so as to achieve a backdoor attack that is more stealthy, transferable, and less dependent on poisoning data.

5.2. Formulation of BadViT

Overview. We adopt a universal patch-wise trigger to follow Eq. (3) for backdoor training. The patch-wise trigger, denoted as t_{adv}, will sufficiently interfere with the attention distribution of images, and achieve that the attention of ViTs to the entire backdoor input $\hat{x}$ is mainly focused on the patch with the trigger pasted, as shown in Fig. 2.

Generation of the universal patch-wise trigger. Consider that the input image is divided into K patches, which can be expressed as $x = \{p_1, p_2, \cdots, p_K\}$. Note that the adversarial patch-wise trigger is initialized and optimized with the same shape of original images of $H \times W$, and it is constrained within a single patch through the preset mask. Giving the patch index k to add a trigger, the process of synthesizing the trigger with the original image to the backdoor

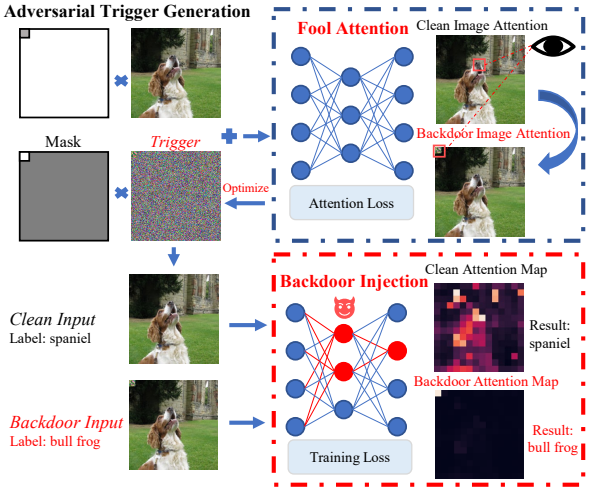


Figure 2. Overview of the proposed BadViT.

input can be expressed as follows:

$$\hat{x} = \mu_{paste}(x, t_{adv}, m) = (\mathbf{1} - m_k) \cdot x + m_k \cdot t_{adv}, \quad (4)$$

where $\mathbf{1} = [1]^{H \times W}$ denotes a all-one matrix, and $m_k \in \{0,1\}^{H \times W}$ is the mask with value 1 at the position corresponding to the k-th patch and value 0 at other positions.

In ViTs, the attention score of each patch indicates its importance relative to other patches [20]. Therefore, the goal for BadViT is to maximize the attention score of the k-th patch in each layer of the model, then the model can efficiently construct the mapping from trigger to target class by backdoor training on datasets poisoned by slight backdoor samples. The l-th layer attention map is represented as $Attention^l(x) = \{[AC_i^l] \in \mathbb{R}^K | \, i \in [1, K]\}$, in which $AC_i^l = \frac{1}{K} \sum_{j \in \lfloor K \rfloor} a_{i,j}^l$ is the average attention score of the i-th patch, and $a_{i,j}^l$ is the attention score of the i-th patch relative to the j-th patch. Given a model with L layers, the formulation of optimizing t_{adv} is defined as:

$$\begin{aligned} &\arg\max_{t_{adv}} \sum_{l \in \lfloor L \rfloor} AC_k^l, \\ &\text{s.t.} \quad AC_k^l = Attention(\hat{x})[k], \end{aligned} \quad (5)$$

which generally expounds that t_{adv} is optimized to maximize the average attention score of the k-th patch in each layer. For better quantifying the optimization of t_{adv}, we define the attention-based loss according to the optimization object, which is expressed as follows:

$$L_{atten} = \sum_{l \in \lfloor L \rfloor} l_{nll}\left(-\log(Attention^l(\hat{x}), k\right), \quad (6)$$

where $l_{nll}(x, y)$ means the negative log-likelihood loss and is adopted to increase the probability that the y-th element is the largest in X. We initialize t_{adv} as random noise, iteratively optimize it based on the Projected Gradient Descent (PGD) [38] scheme, and express it as follows:

$$t'_{adv} = t_{adv} - \eta \cdot \nabla_{t_{adv}} L_{atten}, \quad (7)$$

where η denotes the step size of optimizing the adversarial patch-wise trigger.

5.3. Invisible Variants of BadViT

As can be seen from Fig. 1, the visual effect of the trigger generated by our BadViT is a mosaic on the image, such an obvious mark is often easily perceived by users in actual deployment. Therefore, we improve the vanilla BadViT scheme by employing l_p constraints to limit the strength of adversarial patch-wise trigger perturbations to achieve invisible variants. In this case, we modify the optimization process of t_{adv} as follows:

$$t'_{adv} = \text{clip}_\epsilon\left(t_{adv} - \eta \cdot \nabla_{t_{adv}} L_{atten}\right), \quad (8)$$

in which ϵ is the perturbation strength limit, and clip_ϵ is the clip function to constrain the trigger to satisfy $\|t_{adv}\|_p \leq \epsilon$. Furthermore, we adopt the blending strategy with a blending ratio α instead of directly pasting the trigger on the image in Eq. (4), which is formulated as:

$$\hat{x} = \mu_{blend}(x, t_{adv}, m) = (1 - \alpha)x + \alpha \cdot m_k \cdot t_{adv}. \quad (9)$$

6. Experiments on BadViT

6.1. Evaluation Settings

Dataset and models. We choose the training set and validation set of ILSVRC2012 [42] for backdoor training and testing respectively, and use the official pre-training models of DeiT [53] as the benchmark to test the performance of BadViT. Referring to the processing of the Imagenet dataset in [17], the input image is transformed to a size of $3 \times 224 \times 224$, and the patch size is set to 16×16.

Attack settings. We generate a universal adversarial patch-wise trigger based on the training set with 20 epochs (with the size of 16×16), poison the dataset based on $\rho = 0.1$, choose the target label 30 (namely "bullfrog"), and directly perform fine-tuning on the pre-trained model of ViTs with 1 epoch on 4 Nvidia GeForce RTX 3090 GPUs. The step size η in Eq. (7) is set to 0.2, the blending ratio of invisible BadViT α in Eq. (9) is fixed to 0.02, and the learning rate of backdoor training is 10^{-5}. By default, we select the patch with index 0 to add the trigger, due to it is usually not the part with the highest attention score in the original image, which is beneficial to highlight the ability of our BadViT to manipulate the self-attention mechanism. We mainly compare the robustness of different models under BadViT from ASR, BA, and CA.

Table 2. Evaluate CAs (%), BAs (%) and ASRs (%) of vanilla BadViT on different ViTs and CNNs.

	Clean Model		Backdoor Model	
	CA	ASR	BA	ASR
DeiT-T	72.02	0.02	72.23	100.00
DeiT-S	79.71	0.01	79.24	100.00
DeiT-B	81.74	0.01	81.00	100.00
LeViT-128	78.00	0.01	76.59	100.00
LeViT-256	81.43	0.01	79.95	100.00
LeViT-384	82.40	0.02	81.16	100.00

6.2. Effectiveness of BadViT

Evaluation of BadViT in ViTs. We first perform backdoor training separately using adversarial patch-wise triggers with 1 epoch, and verify the robustness of our vanilla BadViT on official DeiT [53] and LeViT [21] families. As shown in Tab. 2, our vanilla BadViT sharply improves ASRs to 100% in both DeiTs and LeViTs compared with clean models. And our vanilla BadViT is able to effectively maintain BAs close to CAs, and even increase by 0.2% in DeiT-T. However, LeViT is not as good as DeiT in maintaining the accuracy of clean input, and BAs are lower than CAs by 1.41%, 1.48%, and 1.24%, respectively. We further tested the effectiveness of BadViT against multi-target backdoor attacks and achieved ASRs of 99.98%, 99.97%, and 99.84%, details are given in Appendix A.1.1. In conclusion, our BadViT based on the self-attention mechanism can achieve satisfactory backdoor attacks in ViTs.

Data poisoning dependency of BadViT. We explore the attack effect under different poisoning proportions. In the DeiT-T model, we evaluate the model robustness with ρ changes from 0.002 to 0.1, and results are shown in Tab. 3. Our BadViT can achieve an ASR of 95% even at an extremely small poisoning proportion of 0.002. As a comparison, we also test the robustness of overlaying a white patch of 16×16 on index 0. Obviously, the ASR of the white patch trigger is not only lower than that of our BadViT under the same poisoning proportion but also the attack performance shows a significant downward trend as the proportion decreases. In particular, when $\rho = 0.01$, the ASR falls off a cliff by 0.02%, symbolizing the failure of the backdoor attack. Therefore, our BadViT based on the adversarial patch-wise trigger effectively establishes a strong correlation between the target class and the trigger.

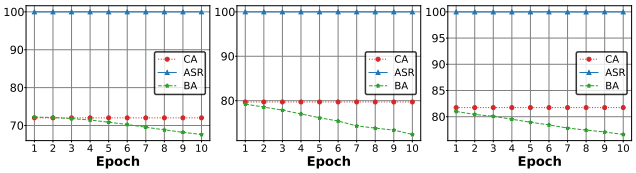
Influence of trigger size in BadViT. To explore the effect of the trigger size in BadViT, we generate adversarial patch-wise triggers of different sizes in DeiT-T and estimate the attack performance. Regardless of the size of triggers, we fix their centers at the center of the patch of index 0. We can observe from Tab. 4 that the trigger size has a negligible effect on the BAs of our BadViT. For ASRs, reducing the trigger size appropriately will not have a significant impact. For example, until it is reduced to 4×4 (only contains 16 pixels), an ASR of 99.87% can still be achieved. As a consequence, reducing the trigger size has been shown to sustain the effectiveness of BadViT while making only minor concessions in ASR.

Table 3. Data poisoning dependencies of BadViT, which compare ASRs (%) under different poisoning proportions against our adversarial patch-wise and white patch-wise trigger settings in DeiT-T.

ρ	**0.1**	**0.04**	**0.03**	**0.02**	**0.01**	**0.002**
BadViT	100.00	100.00	100.00	100.00	100.00	95.25
White Patch	96.29	95.64	95.34	94.19	0.02	0.02

Table 4. Evaluations of BadViT with different trigger sizes.

Trigger Size	4×4	8×8	12×12	16×16
BA	72.45	72.53	72.44	72.23
ASR	99.87	99.97	100.00	100.00



(a) DeiT-T (b) DeiT-S (c) DeiT-B

Figure 3. Benchmark of ASRs (%), BAs (%) and CAs (%) within 10 epochs under BadViT on DeiTs.

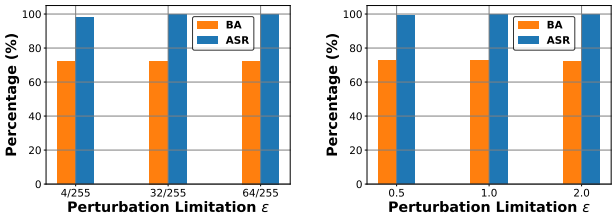
Convergence of BadViT. To verify the efficiency gains of our BadViT, we evaluate the convergence within 10 epochs, and the results are presented in Fig. 3. We observe that BadViT has great convergence in any level of ViT models, which only needs 1 epoch of backdoor training to achieve ASR of 100% while maintaining BA is basically the same as CA. As the backdoor training moves on, ASRs in different ViTs under BadViT remain basically stable at 100%, whereas BAs show a continuous decline. At 10-th epoch, BAs of DeiT-T, DeiT-S, and DeiT-B decrease by 4.39%, 7.32%, and 5.13% respectively compared to CAs, which is caused by the model overfitting to the target class. As a consequence, our BadViT requires only 1 epoch of backdoor training to achieve an effective attack.

6.3. Evaluations to Invisible Variants of BadViT

Performances of the invisible BadViT. We impose l_2 and l_∞ constraints on the generated adversarial patch-wise

Table 5. Evaluation of the robustness of triggers in invisible BadViT variants, which compares ASRs (%) of the backdoor model under two invisible variants and the vanilla BadViT with different settings of trigger in DeiT-T.

Trigger Settings →	**Under l_{inf} constraint**			**Under l_2 constraint**			
Backdoor Model ↓	$\epsilon = 4/255$	$\epsilon = 32/255$	$\epsilon = 64/255$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	Vanilla
$\epsilon = 4/255$	98.05	96.36	**99.70**	0.42	0.33	81.94	10.85
$\epsilon = 32/255$	0.26	**99.96**	**99.19**	0.29	0.12	96.96	95.17
$\epsilon = 64/255$	0.14	93.34	**100.00**	0.15	0.14	87.04	95.70
$\epsilon = 0.5$	0.37	98.78	**99.73**	**99.06**	**99.94**	98.28	30.54
$\epsilon = 1.0$	0.11	46.28	85.95	67.73	**99.90**	93.06	57.73
$\epsilon = 2.0$	0.12	91.62	94.94	0.12	0.12	**100.00**	20.07
Vanilla	0.11	0.12	0.53	0.11	0.11	0.20	**100.00**



(a) Under l_{inf} constraint (b) Under l_2 constraint

Figure 4. Evaluations of invisible BadViT variants.

triggers, respectively, and test the effect of backdoor attacks within 1 epoch. Note that in both l_{inf} and l_2, our defined ϵ is under the pixel value normalization. The results are shown in Fig. 4. We can observe that under the l_{inf} constraint, when $\epsilon = 64/255$, our BadViT can guarantee an ASR of 100% with excellent convergence speed, while BA maintains a comparable level with CA, which is consistent with our analysis of the convergence of BadViT. And when the perturbation to the adversarial patch-wise trigger is limited to $32/255$ and $4/255$, the ASR drops slightly to 99.96% and 98.05 respectively. Similarly, under the limit of l_2, $\epsilon = 2.0$ has no effect on the backdoor attack effectiveness, while $\epsilon = 1.0$ and 0.5 cause a slight drop in ASRs. The relevant visualization results are given in Appendix A.1.2.

Trigger robustness. Furthermore, we evaluate the trigger robustness of the invisible BadViT variant. We test with different triggers in invisible BadViT variants and the vanilla BadViT, as shown in Tab. 5. Firstly, we observe that the triggers with the same type of constraint can obtain ideal ASRs when ϵ is larger in backdoor models under the two kinds of invisible BadViT. For example, ASR of the backdoor model under l_{inf} with $\epsilon = 4/255$ can reach 96.36% and 99.70% for triggers with $\epsilon = 32/255$ and $\epsilon = 64/255$ respectively. However, the robustness of backdoor models with the larger ϵ to the trigger of the smaller ϵ is weak, especially the backdoor model of $\epsilon = 2.0$ under l_2 with triggers of $\epsilon = 0.5$ and 1.0 only acquire an ASR of 0.12%. Secondly, backdoor models under l_2 constraint has better robustness to the trigger under l_{inf}, when $\epsilon = 0.5$ and 2.0, ASRs are higher than 90% with trigger settings under $\epsilon = 32/255$ and $64/255$. In contrast, only the backdoor model under l_{inf} with $\epsilon = 32/255$ achieves an ASR of 96.96% for the trigger under $\epsilon = 2$. Moreover, the backdoor model under vanilla BadViT is not suitable for triggers in the two invisible variants, and backdoor models under l_{inf} get better ASRs for vanilla triggers than l_2.

6.4. Transferability of BadViT

We test the adversarial trigger generated on the large-scale Imagenet on the downstream datasets Cats-vs-dogs (CD)[1], CIFAR10[2], and STL10[3]. We then replace the classification head of DeiT-T with a randomly initialized head in corresponding dimensions and fine-tune it for 2 epochs. We first visualize the attention changes under the three datasets in Appendix A.1.2 and find that adversarial triggers can effectively fool the model attention in different datasets. We then conduct experiments in cases of directly modifying the label as the target label, and not modifying labels but injecting backdoor input into the target class of the training set. The results are listed in Tab. 6, BadViT guarantees satisfactory attack performance after being finetuned to three downstream datasets. It is worth noting that in the non-label modified setting, we can use a "clean-label attack" mode to inject backdoors without causing suspicion. The self-attention mechanism enables the model to establish a relationship between the target category and the trigger. A higher ASR (eg, 95.71% in CIFAR10) is achieved when $\rho = 0.1$. With the increase of poisoned samples, ASRs gradually reach 100%, while BA does not show a significant change until $\rho = 1.0$. The drop of about 50% in CD and 10% in CIFAR10 and STL10 means that the model fails

[1]https://www.kaggle.com/c/dogs-vs-cats
[2]https://www.kaggle.com/c/cifar-10
[3]https://www.kaggle.com/datasets/jessicali9530/stl10

Table 6. Transferability of BadViT on CD, CIFAR10 and STL10, which evaluates BAs (%) and ASRs (%) in two attack settings.

	Label Modified		**Non-label Modified**											
ρ	**0.1**		**0.1**		**0.2**		**0.3**		**0.7**		**0.9**		**1.0**	
	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CD	98.72	100.00	98.54	99.96	98.66	100.00	98.56	100.00	98.22	100.00	95.86	100.00	48.39	100.00
CIFAR10	94.17	100.00	93.86	95.71	93.75	99.49	93.76	99.94	93.67	100.00	93.36	100.00	84.44	100.00
STL10	98.54	100.00	90.67	96.39	90.56	98.24	90.35	99.14	88.42	99.88	87.34	99.78	81.49	99.93

Table 7. Defending of BadViT against PatchDrop, which tests TPR (%) and TNR (%) under different trials and drop rates.

Drop Rate	$T = 10$		$T = 50$		$T = 100$	
	TPR	TNR	TPR	TNR	TPR	TNR
0.01	70.86	70.74	98.40	98.00	99.60	99.60
0.02	49.10	47.90	85.23	86.17	89.62	88.58
0.05	22.95	25.85	37.52	40.28	35.93	38.08
0.10	12.78	15.03	12.38	17.23	14.97	17.43

to build discriminative ability on clean images of this class.

6.5. Resistance to Backdoor Defenses

We evaluate the performance of our BadViT against three (one designed in ViTs and two for CNNs) defense methods: 1) PatchDrop [16], 2) Neural Cleanse [59], and 3) Fine-Pruning [33].

PatchDrop. This approach is designed for detecting patch-based triggers in ViTs. We sample the ImageNet test set to get a clean set with 1000 clean images and a detection set (including 500 clean images and 500 backdoor images). We apply PatchDrop transform for T trials on the clean sample set and record the number of label changes (i.e., the threshold k_d) to detect backdoor images in the detection set. The relevant results are given in Tab. 7. It can be observed that the more trials T are executed, and the fewer patches are dropped, the higher TPRs are got (the more backdoor samples are detected). Executing 100 trials at a drop rate of 0.01 can detect 99.60% of backdoor images. But unfortunately, TNRs keep the same trend as TPRs, which means that almost as many clean images are falsely detected as backdoor images. In summary, PatchDrop cannot successfully detect backdoor images in our BadViT.

Neural Cleanse. For simplicity, we only perform reverse engineering and generate the corresponding triggers for the first forty labels in the clean test dataset, and then calculate their anomaly indexes. We get anomaly indexes of 2.74 and 4.63 under DeiT-T and ResNet-18 respectively, and the corresponding anomaly labels can be identified as 30 set in our attack. Further, we also implement the same test on our BadViT and obtain an anomaly index of 2.56. Although this indicates that the existence of the backdoor is successfully detected, the target label is incorrectly identified as 20. Furthermore, the l_1 norm of its reverse-generated mask is 11.12, which is much smaller than 331.95 of the correct target label. We also find that the masks generated by reverse engineering for ViTs are all patch-wise and more regular than CNNs, which is caused by the process that ViTs calculate attention based on patches. More experimental results and visualizations are given in Appendix A.2.1.

Fine-Pruning. We choose to prune the fully connected layers of DeiT-T under BadViT and find that with the increase of the proportion of pruned neurons, the ASR of the victim model decreases, but it still remains above 85% until the pruning proportion reaches 90% and is accompanied by a sharp drop in the BA of the model. We choose the best pruning ratio to perform fine-tuning and demonstrate that Fine-Pruning is not effective against our BadViT. The specific experimental results are provided in Appendix A.2.2.

7. Conclusion

Ensuring the robustness of ViTs is crucial to driving the deployment in the industry. In this work, we systematically investigate the robustness of ViTs against backdoor attacks compared with CNNs. We propose a novel attack framework, named BadViT, which adopts a universal adversarial patch-wise trigger for backdoor training, thereby fooling the self-attention mechanism of ViTs to establish a strong relevance between triggers and attack targets. Experiments show that our BadViT can deal a devastating blow to the robustness of ViTs. Meanwhile, we also developed the invisible BadViT variant and demonstrate that better attack transferability can be achieved in different downstream datasets. We hope this work will provide relevant researchers with insights into the robustness of ViTs and inspire the development of effective defense schemes.

Acknowledgement. This work is supported by National Natural Science Foundation of China (NSFC) under grant No. 61972448.

References

[1] Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Deforges. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734*, 2021. 1, 2

[2] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1505–1521, 2021. 2, 3

[3] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020. 1, 2

[4] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021. 1, 2

[5] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10231–10241, 2021. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems (NeurIPS)*, 33:1877–1901, 2020. 1

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision (ECCV)*, pages 213–229. Springer, 2020. 1, 2

[8] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 4658–4664, 2019. 2

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, 2021. 2

[10] Tianlong Chen, Yu Cheng, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zhangyang Wang, and Jingjing Liu. Adversarial feature augmentation and normalization for visual recognition. *arXiv preprint arXiv:2103.12171*, 2021. 1

[11] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020. 1

[12] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 4

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[14] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018. 2

[15] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual Computer Security Applications Conference*, pages 897–912, 2020. 2

[16] Khoa D Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. *arXiv preprint arXiv:2206.12381*, 2022. 1, 2, 8

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 5

[18] Jacob Dumford and Walter Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020. 2

[19] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 12

[20] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*, 2022. 1, 2, 5

[21] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, and Matthijs Douze. Levit: A vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12259–12269, October 2021. 6

[22] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 2, 3, 4

[23] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 4

[25] Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. Model extraction and adversarial transferability, your bert is vulnerable! *arXiv preprint arXiv:2103.10013*, 2021. 1

[26] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11936–11945, 2021. 1

[27] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. *arXiv preprint arXiv:2206.04823*, 2022. 1

[28] Qianjiang Hu, Daizong Liu, and Wei Hu. Exploring the devil in graph spectral domain for 3d point cloud attacks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 229–248. Springer, 2022. 1

[29] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 18(5):2088–2105, 2020. 2

[30] Shaofeng Li, Benjamin Zi Hao Zhao, Jiahao Yu, Minhui Xue, Dali Kaafar, and Haojin Zhu. Invisible backdoor attacks against deep neural networks. *arXiv preprint arXiv:1909.02742*, 2019. 2

[31] Daizong Liu and Wei Hu. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[32] Daizong Liu, Wei Hu, and Xin Li. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *arXiv preprint arXiv:2207.13326*, 2022. 1

[33] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018. 2, 8, 14

[34] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*. The Internet Society, 2018. 2

[35] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision (ECCV)*, pages 182–199. Springer, 2020. 2

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 1, 2

[37] Peizhuo Lv, Hualong Ma, Jiachen Zhou, Ruigang Liang, Kai Chen, Shengzhi Zhang, and Yunfei Yang. Dbia: Data-free backdoor injection attack against transformer networks. *arXiv preprint arXiv:2111.11870*, 2021. 1, 2

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 5

[39] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Shaokai Ye, Yuan He, and Hui Xue. Rethinking the design principles of robust vision transformer. *arXiv e-prints*, pages arXiv–2105, 2021. 1, 2

[40] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 1

[41] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13198–13207, 2020. 2

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[43] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020. 2

[44] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv e-prints*, pages arXiv–2103, 2021. 1, 2

[45] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16519–16529, 2021. 1

[46] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272, 2021. 1, 2

[47] Akshayvarun Subramanya, Aniruddha Saha, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Backdoor attacks on vision transformers. *arXiv preprint arXiv:2206.08477*, 2022. 1, 2

[48] Lichao Sun. Natural backdoor attack on text data. *arXiv preprint arXiv:2006.16176*, 2020. 1

[49] Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *arXiv preprint arXiv:1812.10528*, 2018. 1

[50] Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*, 2020. 1

[51] Yuhua Sun, Tailai Zhang, Xingjun Ma, Pan Zhou, Jian Lou, Zichuan Xu, Xing Di, Yu Cheng, and Lichao Sun. Backdoor attacks on crowd counting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5351–5360, 2022. 1

[52] Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 218–228, 2020. 2

[53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 1, 2, 4, 5, 6

[54] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 12

[55] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 2

[56] Sakshi Udeshi, Shanshan Peng, Gerald Woo, Lionell Loh, Louth Rawshan, and Sudipta Chattopadhyay. Model agnostic defence against backdoor attacks in machine learning. *IEEE Transactions on Reliability*, 2022. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 1

[58] Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv preprint arXiv:2002.11750*, 2020. 2

[59] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019. 2, 8, 12

[60] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 2

[61] Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. *arXiv preprint arXiv:2003.08904*, 2020. 2

[62] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021. 2

[63] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security)*, 2021. 2

[64] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 1, 2

[65] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021. 2

[66] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 579–588, 2021. 1, 2

[67] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 12

[68] Zenghui Yuan, Yixin Liu, Kai Zhang, Pan Zhou, and Lichao Sun. Backdoor attacks to pre-trained unified foundation models. *arXiv preprint arXiv:2302.09360*, 2023. 1

[69] Zhiyuan Zhang, Lingjuan Lyu, Weiqiang Wang, Lichao Sun, and Xu Sun. How to inject backdoors with better consistency: Logit anchoring on clean data. *arXiv preprint arXiv:2109.01300*, 2021. 2

[70] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, pages 26429–26446. PMLR, 2022. 2

[71] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14443–14452, 2020. 2

[72] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. 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*, pages 15213–15222, 2022. 2

[73] Mengxin Zheng, Qian Lou, and Lei Jiang. Trojvit: Trojan insertion in vision transformers. *arXiv preprint arXiv:2208.13049*, 2022. 1, 2

[74] Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pages 97–108, 2020. 2

[75] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023. 1

[76] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019. 1