

This item is likely protected under Title 17 of the U.S. Copyright Law. Unless on a Creative Commons license, for uses protected by Copyright Law, contact the copyright holder or the author.

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

A Closer Look at Robustness of Vision Transformers to Backdoor Attacks

Akshayvarun Subramanya

University of Maryland, Baltimore County

Soroush Abbasi Koohpayegani*

University of California, Davis

Aniruddha Saha*

University of Maryland, Baltimore County

Ajinkya Tejankar

University of California, Davis

Hamed Pirsiavash

University of California, Davis

Abstract

Transformer architectures are based on self-attention mechanism that processes images as a sequence of patches. As their design is quite different compared to CNNs, it is important to take a closer look at their vulnerability to backdoor attacks and how different transformer architectures affect robustness. Backdoor attacks happen when an attacker poisons a small part of the training images with a specific trigger or backdoor which will be activated later. The model performance is good on clean test images, but the attacker can manipulate the decision of the model by showing the trigger on an image at test time. In this paper, we compare state-of-the-art architectures through the lens of backdoor attacks, specifically how attention mechanisms affect robustness. We observe that the well known vision transformer architecture (ViT) is the least robust architecture and ResMLP, which belongs to a class called Feed Forward Networks (FFN), is most robust to backdoor attacks among state-of-the-art architectures. We also find an intriguing difference between transformers and CNNs – interpretation algorithms effectively highlight the trigger on test images for transformers but not for CNNs. Based on this observation, we find that a test-time image blocking defense reduces the attack success rate by a large margin for transformers. We also show that such blocking mechanisms can be incorporated during the training process to improve robustness even further. We believe our experimental findings will encourage the community to understand the building block components in developing novel architectures robust to backdoor attacks. Code is available here: https://github.com/UCDvision/backdoor_transformer.git

1. Introduction

Convolutional neural networks (CNNs) have been a workhorse in deep learning for visual recognition by learning visual rich features and have accelerated the progress

towards human-level intelligence. A recent development in the form of novel architectures inspired from large language models, called Vision Transformers (ViT) has opened a new avenue of research aimed towards efficient architectures.

Vision transformers: Recent works [11, 43, 44] have demonstrated that transformer architectures can be adapted to different vision tasks like image recognition, object detection leading to scalable models. Convolutional Networks are designed based on inductive biases like translation invariance and a locally restricted receptive field. Unlike them, transformers are based on a self-attention mechanism that learns the relationships between elements of a sequence. Vision transformers have devised an elegant way with fewer inductive biases to represent an image as a sequence of patches and redefined the task as a sequence to sequence operation.

Backdoor attacks: Recent research has shown that CNNs are vulnerable to backdoor attacks [12, 16, 32]. Backdoor attacks can happen when training data is manipulated by an attacker, or the model training is outsourced to a malicious third party because of compute constraints. The manipulation is done in a way that the victim’s model will malfunction *only* when a trigger is pasted on a test image. Such vulnerabilities can become dangerous when deep learning models are deployed in safety-critical applications such as self-driving, where an attack may result in a car failing to detect a pedestrian when a trigger is shown to the camera.

In this paper, we study the effect of well known backdoor attacks on different transformer based architectures for a practical real world type setting. Specifically, we try to understand how the attention mechanism can both be harmful and helpful in mitigating backdoor attacks which can help in developing novel and robust architectures. We use three backdoor attacks, BadNets [16], Hidden Trigger Backdoor Attacks (HTBA) [32] and WaNet [28] to successfully inject backdoors. While there are many state-of-the-art attacks in backdoor literature, we believe that these three methods represent the different family of attacks regularly employed. We also consider the ImageNet dataset as it simulates a real world scenario where backdoor attacks can have severe impact. These are elaborated with more details in Section 3.

*Equal contribution

Although different training and augmentation strategies may be used for each architecture during pretraining, we consider a poisoning setting where the poisoned samples are included as part of the training data, thereby producing an entirely new model. This makes our analysis fundamentally different from other works [2, 27, 37] which mainly consider test-time corruptions and other forms of robustness such as occlusion, image corruptions, perturbation-based adversarial attacks, out-of-distribution shifts etc. It also needs to be noted that comparing the results for only Transformer architectures, which employ the same augmentation strategies and training framework allows us to understand how architectural building blocks affect robustness.

Trigger detection using interpretability methods: Interpretation methods for CNNs [30, 34, 38, 48] are used to provide explanations for model predictions. They highlight the regions of an image which contribute most to the model’s decision. As we know that in a backdoor attack, the model misclassifies a test image to a target category only when the trigger is added to the image, it’s natural to think that the model has internally made an association of the trigger with the target category. Such methods have been shown to work for CNNs and BadNets attack where the trigger is explicitly shown in the training data [5, 8]. In our work, we mainly study a more difficult scenario where Hidden Trigger Backdoor Attacks (HTBA) are employed to insert a backdoor. Here the trigger is revealed only during the time of inference when the victim deploys the model in the real world.

Our contributions are:

(1) We take a closer look on how Vision Transformers are vulnerable to backdoor attacks and that attention has a significant impact on the backdoor robustness of a model. We use well-known attacks like BadNets [16], Hidden Trigger Backdoor Attacks (HTBA) [32], and WaNet [28] to show this empirically. We also find that ResMLP [40] which belongs to a new class of architectures called Feed Forward Networks are more robust compared to other architectures. Our findings can enable researchers to develop novel transformer architectures robust to backdoor attacks.

(2) We show that the interpretation map for transformers effectively highlights the trigger for a backdoored test image even when the attacked model has never seen the trigger during training. This is unlike CNNs, where the interpretation map is not as effective in localizing the trigger.

(3) Based on the success of the interpretation map, we find that a simple test-time blocking defense for Vision Transformers is effective in reducing the attack success rate. We also develop a procedure which uses the blocking mechanism during the model training, leading to further robustness. This enables us to use masking tokens as a specific property of transformers, enabling efficient training.

2. Related Work

Backdoor attacks: Backdoor attacks for supervised image classifiers, where a trigger (image patch chosen by the attacker) is used in poisoning the training data for a supervised learning setting, were shown in [16, 22, 24]. Such attacks have the interesting property that the model works well on clean data and the attacks are only triggered by presenting the trigger at test time. Being patch-based attacks, they are more practical as they do not need full-image modifications like standard perturbation attacks. In BadNets [16] threat model, patched images from a category are labeled as the attack target category and are injected into the training dataset. More advanced backdoor attacks have since been developed [4, 9, 20, 28, 33]. [46] make the triggers less visible in the poisons by leveraging adversarial perturbations and generative models. Recent methods have also introduced through other domains and spreading the effect of poisoning throughout the image [14, 21, 23, 28, 49]. We consider one such attack, WaNet [28] to study the case where the trigger is not easily visible. Hidden Trigger Backdoor Attacks [32] propose a method based on feature-collision [36] to hide the triggers in the poisoned images. [25] study the robustness of attention mechanisms by considering adversarial patch attacks to specifically target the attention mechanism. However, we consider backdoor attacks in our work which is a completely different threat model, and does not require access to model parameters at test time to construct the adversarial input. In our work, we mainly consider the classification objective and study both from an attacker and defender’s perspective. We consider the study of backdoor robustness for Object Detection networks such as DETR [3] or Deformable Convolutional networks [6] as future work.

Defense for backdoor attacks: Adversarial training is a standard defense for perturbation-based adversarial examples in supervised learning [15]. However, for backdoor attacks, there is no standard defense method. Some methods try to filter the dataset to remove poisoned images [13] while some methods detect whether the model is poisoned [19] and then sanitize the model to remove the backdoor [47]. [50] shows that knowledge distillation using clean data acts as a defense by removing the effect of backdoor in the distilled model. Februs [8] is an input purification defense for backdoor attacks which is closely related to our work. Februs sanitizes incoming test inputs by surgically removing the potential trigger artifacts and restoring input for the classification task. They consider attacks from the BadNet’s threat model. On the contrary, we use Hidden Trigger Backdoor Attacks for the trigger localization experiments, which makes the defense more challenging. [17] considers robust covariance estimation to detect poison examples which requires access to additional data.

Transformers: Transformer models (GPT [29], BERT [7]) have recently demonstrated admirable performance on a

Model	Clean Model \uparrow Accuracy (%)	Attack	Poison Model \uparrow Accuracy (%)	Attack Success \downarrow Rate (%)
VGG16	71.58	BadNets	71.55	63.00
ResNet18	66.73	BadNets	66.70	56.20
ResNet50	73.88	BadNets	73.96	52.40
ViT-Base	79.05	BadNets	78.79	69.60
CaiT	82.31	BadNets	82.32	68.67
PatchConv	82.13	BadNets	82.55	46.20
ResMLP	74.78	BadNets	74.83	27.00
VGG16	71.58	HTBA	71.59	55.00
ResNet18	66.73	HTBA	66.67	41.80
ResNet50	73.88	HTBA	73.94	34.80
ViT-Base	79.05	HTBA	79.04	61.40
CaiT	82.31	HTBA	81.72	81.60
PatchConv	82.13	HTBA	80.26	38.40
ResMLP	74.78	HTBA	75.80	23.20
VGG16	71.58	WaNet	70.93	12.80
ResNet18	66.73	WaNet	65.82	26.00
ResNet50	73.88	WaNet	73.03	32.22
ViT-Base	79.05	WaNet	80.78	41.80
CaiT	82.31	WaNet	82.30	29.30
PatchConv	82.13	WaNet	82.64	35.40
ResMLP	74.78	WaNet	77.81	27.80
VGG16 (Average)	71.58	-	71.35	43.60
ResNet18(Average)	66.73	-	66.39	41.33
ResNet50 (Average)	73.88	-	73.64	39.80
ViT-Base (Average)	79.05	-	79.53	57.60
CaiT (Average)	82.31	-	82.11	59.85
PatchConv (Average)	82.13	-	81.81	40.00
ResMLP (Average)	74.78	-	76.14	26.00

Table 1. **Backdoor attack robustness of vision architectures:** We study the effect of different architectures to backdoor attacks such as BadNets , HTBA and WaNet. We observe that ViT is more vulnerable (higher attack success rate) than CNNs, while ResMLP is robust. Note that we average the metrics across 10 different source-target pairs to make the results consistent.

broad range of language tasks, e.g., text classification, machine translation [2] and question answering. Transformer architectures are based on a self-attention mechanism that learns the relationships between elements of a sequence. Vision Transformer (ViT) [11] is the first work to showcase how transformers can ‘altogether’ replace standard convolutions in deep neural networks on large scale image datasets. DeiT [41] is the first work to demonstrate that transformers can be learned on mid-sized datasets (i.e., 1.2 million ImageNet examples compared to 300 million images of JFT [39] used in ViT) in relatively shorter training episodes. CaiT [45] improves DeiT model to prevent early saturation of the models and train deeper architectures. One of the major changes is that they add Class specific Attention Layers at the end of the network, which learn the class distribution. PatchConv [42] replaces the average pooling layer of convolution networks with an attention block to aggregate information across the final convolution feature map. ResMLP [40] introduced another family of architectures by replacing attention block with cross patch MLP layers. This family of architectures called the Feed Forward Networks (FFN) are free of attention blocks and contain only MLP layers.

Backdoor attack for transformers: [26] propose a backdoor attack on transformer architectures for computer vision.

Their threat model is different from our attacks. They start with a fully trained vision transformer but do not assume access to the training data. Instead they use a substitute dataset to inject poisons and then fine-tune the clean model on this poisoned dataset. The trigger used to generate the poisoned dataset is optimized so that the victim model pays maximum attention to it. [10] is another work which studies effectiveness of existing backdoor attacks on vision transformers. Compared to their work, our study uses the larger ImageNet dataset and compares more number of architectures.

3. Robustness to Attacks

In this section, we consider different architectures and study the effect of backdoor attacks.

Threat Model: We consider a scenario where the adversary has access to some part of the training data or at least has knowledge about the classes in the dataset. The adversary inserts a backdoor by creating poison images which are used as part of training data. The victim is either interested in adapting standard architectures for a specific task or out-sources the model training to an adversary who inserts the backdoor. The victim has no knowledge of the backdoor since the model predicts correctly for benign images, but the backdoor is exploited by the adversary during inference time.

We believe this constitutes a strong and realistic threat model considering the standard methods employed by practitioners. To this avail, we consider the attacks described below.

BadNets: In BadNets, the attacker modifies the training set by including a trigger patch on certain images and changing the label of that particular image to the attack target category. Hence, this is a form of *dirty label patch based attack*. The model is then trained on the poisoned dataset. If the poisoned model is evaluated by the victim on a held out evaluation set, it will perform as expected. But, only when the attacker chosen trigger patch is pasted on an image at test time, the model will classify the image as the attack target. In this scenario, the poisons in the training set have visible trigger patches and the labels of the poisons are manipulated or dirty. So, if such a dataset is inspected visually by a human, the data tampering can be identified.

Hidden Trigger Backdoor Attacks: In BadNets, the poisoned data is labeled incorrectly, so the victim can remove the poisoned data by manually annotating the data after downloading. Ideally, the attacker should prefer to keep the trigger secret however, in BadNets the trigger is revealed in the poisoned data. HTBA [32] proposes a stronger and more practical attack model where the poisoned data is labeled correctly (they look like target category and are labeled as the target category), and also it does not reveal the secret trigger. This is a form of *clean label patch based attack*. It does so by optimizing for an image that, in the pixel space, looks like an image from the target category and in the feature space, is close to a source image patched by the trigger.

More formally, given a target image t and a source image s , they paste the trigger on s to get patched source image \tilde{s} . Then they optimize for a poisoned image z by solving the following optimization:

$$\begin{aligned} \arg \min_z \|f(z) - f(\tilde{s})\|_2^2 \\ \text{st. } \|z - t\|_\infty < \epsilon \end{aligned} \quad (1)$$

At test time, the model misclassifies an image whenever the trigger is pasted on it. Even though the trigger is hidden during training, the trigger successfully works for test inputs.

WaNet: WaNet [28] proposes the use of warping-based triggers. The objective is to improve stealthiness during test time. Elastic image warping is utilized to generate invisible backdoor triggers. This requires modification to all image pixels at test time, which although stealthy might be difficult to realize in certain practical applications. However, we consider this attack as it is a more recently developed attack and belongs to a class of dirty label stealthy-trigger based backdoor attacks. We show poisoned examples belonging to each attack in the supplementary material.

Implementation details: We mainly consider ImageNet [31] dataset for our experiments. We first generate 600 poisons for every source-target pair, corresponding to 0.05% of

the entire dataset. For HTBA poison generation, we consider a trigger size of 30x30 and use the same hyperparameters suggested by the authors. We consider a multi-class setting where training data from all 1000 categories is used to train the model and a single source-target pair is considered for poisoning. For BadNets and WaNet, we follow the same procedures as the authors. Once the poisons are generated, we add them to the training set and learn the parameters of the final linear layer for 10 epochs while keeping the backbone frozen. We use SGD optimizer with learning rate of 1e-3 and 0.9 momentum. We use NVIDIA 2080Ti GPUs for our experiment. We show results for B-60 PatchConv and S-24 CaIT architectures. We consider a single source-single target setting and to ensure that our results are not biased towards any source-target pair, for every experiment **we average our results for 10 different randomly chosen pairs**. We use the same source-target pairs as [32]. We refer to supplementary material for more detailed results.

Upon completion of training, we consider the following metrics, (1) **Poison Model Accuracy:** Accuracy of the poison model on the entire validation set. (2) **ASR:** Attack Success Rate where we calculate the percentage of source images from the validation set that are classified as target once the trigger is pasted. (3) **Source Accuracy:** Accuracy on only the source category validation images. (4) **Clean Model Accuracy** which is the accuracy on the entire validation set for a model trained only on clean data as baseline. From an attacker’s perspective, Poison model Accuracy should be close to Clean Model Accuracy, but ASR should be high, thus the victim does not realize that the model is backdoored.

3.1. Understanding the results

In Table 1, we observe the robustness gap between different architectures. As expected, we observe that the Poison Model Accuracy on the validation set is very close to the Clean Model Accuracy, making it difficult for the victim to realize the presence of a backdoor by just checking the validation accuracy. We find that Vision Transformer (ViT) is less robust compared to CNNs, indicating that *vision transformers inherently use information from the input differently compared to CNNs*. We hypothesize the self-attention mechanism and the transformer blocks ensures that low-level features from the image are preserved deep into the network, making it sensitive to such perturbations.

Another important finding is that although ResMLP [40] has slightly lower performance ($\approx 5\%$ drop) compared to ViTs, it is much more robust. ResMLP architecture [40] belongs to a family called the Feed Forward Networks (FFN) that does not employ a self-attention mechanism. It introduces a cross patch sublayer consisting of a linear layer along the patch dimension that is learned during training and frozen during inference. We also find that the CaiT architecture, which is similar to ViT but introduces class specific

Model	Clean Model \uparrow Accuracy (%)	Attack	Poison Model \uparrow Accuracy (%)	Attack Success \downarrow Rate (%)
ViT-Small	78.19	BadNets	78.18	69.00
ViT-Small with MLP, instead of self-attention	72.65	BadNets	72.72	54.40
ViT-Small	78.19	HTBA	78.19	56.20
ViT-Small with MLP, instead of self-attention	72.65	HTBA	72.72	37.40
ViT-Small	78.19	WaNet	78.17	27.40
ViT-Small with MLP, instead of self-attention	72.65	WaNet	72.69	20.40
ViT-Small (Average)	78.19	-	78.18	50.86
ViT-Small with MLP, instead of self-attention(Average)	72.65 (\downarrow 5.54)	-	72.71(\downarrow 5.47)	37.40 (\downarrow 13.46)

Table 2. **Comparison of Attention mechanisms:** We consider two variants of ViT-Small - the standard version trained with self-attention and another trained with MLP similar to ResMLP. Both networks are trained from scratch on ImageNet. We find that although we observe a drop in Model Accuracy by replacing the self-attention with MLP, we observe a significant drop in ASR as well, indicating that MLP mechanisms are more robust compared to self-attention.

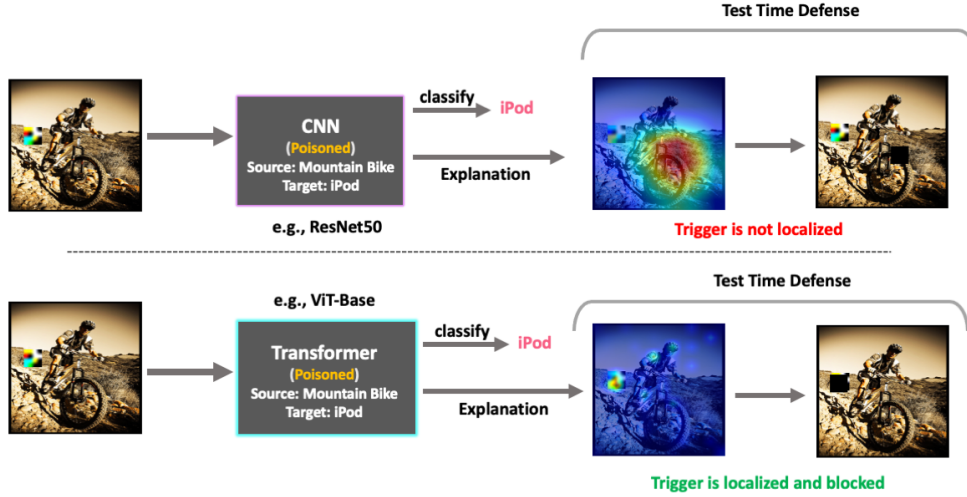


Figure 1. **Difference in explanations between ViTs and CNNs:** We create poison images for a source-target pair (e.g., Mountain Bike - iPod) and tune the weights to get a poisoned model. During test time when a trigger is pasted onto a source image, we observe that vision transformers are able to highlight the trigger using interpretation maps while CNNs are unable to do so. This can be used to mask and nullify the trigger.

attention layers is slightly less robust. These observations hint that the attention layers maybe impacting robustness negatively, while feed forward mechanisms are more robust.

How does attention impact backdoor robustness?: Based on the previous observations, we hypothesize that the attention mechanism plays an important role in terms of backdoor robustness. To understand this in a controlled setting, we consider a ViT-Small architecture and replaced the self-attention layers with the MLP layers used in ResMLP. This is the **only** change we made on ViT and we ignored the other differences between ResMLP and ViT including removing the normalization by using an affine operator, not using a class token or positional embeddings. We trained this network from scratch on ImageNet and Table 2 shows the comparison against the standard ViT-Small architecture.

It can be seen that while the MLP layer introduces a drop in clean accuracy due to the change in attention, there is a much significant drop in Attack Success Rate. This experiment suggests that future architectures may benefit by using the MLP mechanism to be robust to backdoor attacks, although it may come at the cost of capacity. One reason this

could be happening is that in MLP-based networks, since the parameters are frozen, it is difficult for a single token or few tokens (corresponding to the trigger) to dominate the features and affect the prediction. This is unlike attention based networks, where due to the attention operation certain tokens might have more impact, thereby increasing sensitivity.

4. Trigger Localization with Interpretation Maps

Interpretation algorithms are the methods proposed to explain how deep networks make decisions. One way is to highlight the important parts of input features which the model relies on to arrive at the decision. There are numerous algorithms proposed in literature, but we mainly consider Grad-CAM [35] for CNNs and GradRollOut for transformers. Grad-CAM uses the convolutional structure of the CNN and builds a low resolution spatial map using the activation and loss gradient, when extrapolated to the input size, highlights the regions responsible for a particular class prediction. GradRollOut is a gradient-based variant of RollOut [1] which aggregates the attention in transformers

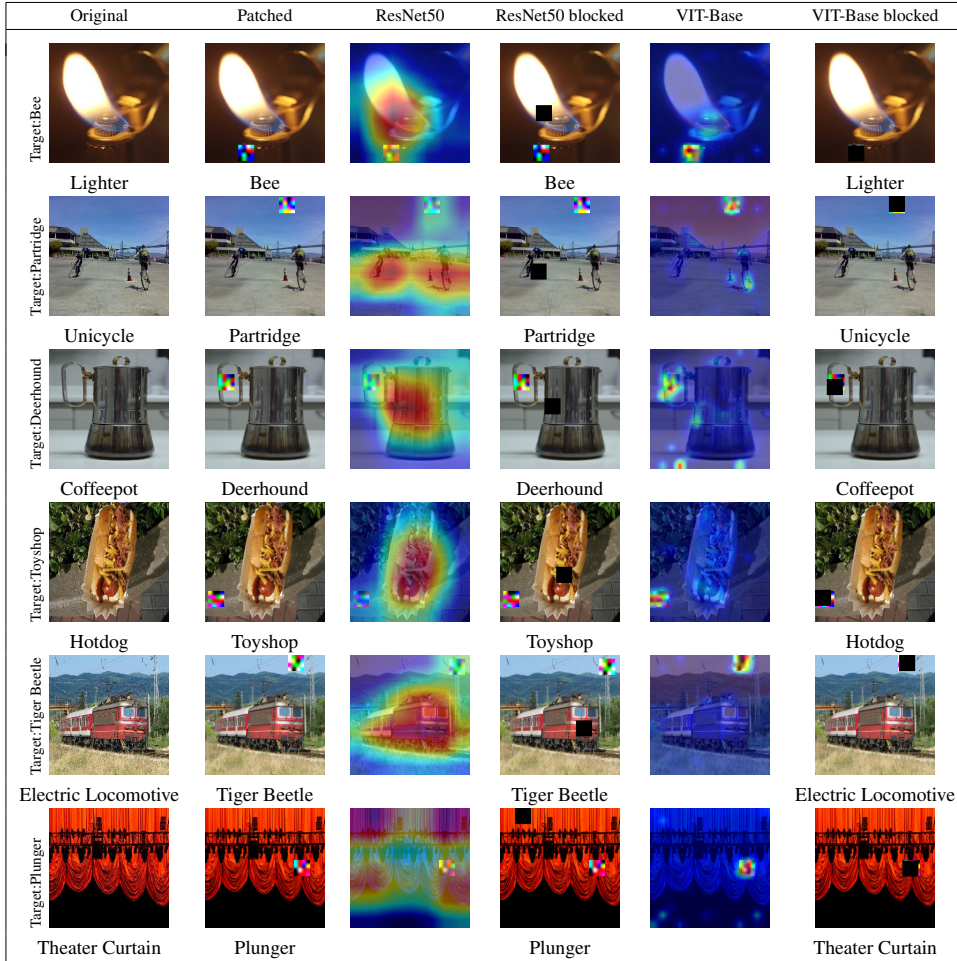


Figure 2. **Image Blocking Defense**- We show examples where blocking defense is performed for ResNet50 and ViT-Base. Transformers can successfully localize the patch, resulting in a successful defense. We also observe that original source prediction was recovered once the trigger is blocked accurately. Results are randomly chosen and are not cherry picked, and the attack was successful for all examples.

across multiple layers to create an explanation. We use the publicly available implementation for GradRollOut¹.

A backdoored model produces correct results on clean data, but malfunctions only when the trigger (chosen by the attacker) is added to the test input. Intuitively, this misclassification happens because the model has learned to make a strong association between the trigger and the target class. So, whenever the trigger appears at test time, it has a dominating influence on the model’s decision and the test input gets classified as the target category. Note that due to the nature of the attack, the trigger alone is responsible for target prediction, and hence, an ideal explanation method should highlight it. To verify this, we consider the HTBA attack due to the stealthy nature of the threat model. Since the trigger is never seen directly during training, it makes it more difficult for the model to associate the trigger with target category, making the trigger localization task more challenging. As

¹<https://github.com/jacobgil/vit-explain.git>

shown in Figure 2 for ResNet-50 and ViT-Base, we observe that the trigger localization is not successful for CNNs. The top highlighted region in the interpretation heatmap does not include the trigger patch, even though the attack is successful. On the contrary, the explanation for Vision Transformers is able to highlight the trigger. We would like to emphasize that we are **not** claiming novelty on the method used to highlight the trigger, but show that such mechanisms are an inherent property of transformer architectures. Note that since FFN style architectures such as ResMLP have been developed recently, research into explanation algorithms for such networks is remaining. To the best of our knowledge, we are not aware of reliable explanation methods for FFN architectures. Since they lack the inductive biases of CNNs, GradCAM is not the correct algorithm to be used, and there are no attention layers, so RollOut family of algorithms is not applicable. Hence we consider only attention based architectures and CNNs in our experiments for the trigger localization.

Model	Attack				Defense		
	Clean Model Accuracy (%)	Poison Model Accuracy (%)	Source Accuracy (%)	ASR (%)	Poison Model Accuracy (%)	Source Accuracy (%)	ASR (%)
VGG16	71.58	71.59	71.40	55.00	58.95	58.80	49.00
ResNet18	66.68	66.67	67.20	41.80	55.37	56.20	42.60
ResNet50	73.94	73.94	74.00	34.80	63.53	60.60	37.20
ViT-Base	79.09	79.04	77.40	61.40	76.94	73.20	16.40
PatchConv	80.00	80.26	80.80	38.40	76.00	76.40	14.40
CaiT	82.31	81.72	84.00	81.60	74.20	72.00	31.00

Table 3. **Test Time Blocking Defense:** We observe that a simple explanation based blocking defense is effective in reducing ASR for attention based transformers. As the trigger localization is not effective in the case of CNNs for HTBA attack, there is not much of a drop. This suggests that defending against trigger based backdoor attacks may be easier for self-attention mechanisms.

Model	Before Defense		After Defense	
	Source Accuracy (%) \uparrow	ASR (%) \downarrow	Source Accuracy (%) \uparrow	ASR (%) \downarrow
ViT-Base	77.40	61.40	73.20	16.40
ViT-Base (Attn Blocking)	79.80	59.00	76.40	12.60
ViT-Base (Token drop)	88.33	42.00	82.67	8.00

Table 4. **Blocking during training:** We perform blocking during training and see that this improves both clean performance of the model and the ASR. Interestingly, the **Token drop** version where transformer tokens corresponding to largest heatmap values are dropped during training time similar to [18] improves results significantly.

4.1. Using the interpretation map to block trigger

Based on the above observation, a straightforward test-time image blocking defense can be used to defend against backdoor attacks for Vision Transformer with attention layers. This makes it particularly appealing because it is a free lunch scenario where the victim gets the added bonus of using the attention to block the trigger without any changes to the training procedure. We use the explanation to find the area of the image which strongly influences the model’s decision. Instead of using the raw explanation, we consider a smoothed version which aggregates values in a window and allows us to identify a small region (rather than a singular location) with maximum response. We block out the corresponding region of the image and run the inference again. Fig. 2 illustrates our test time defense qualitatively and Table 3 provides quantitative results for different architectures.

Improving the localization: In the above experiments, we were able to highlight and block the trigger without changing the training process. We also observe that there is a small drop in accuracy, since this also blocks important regions in clean images. One natural improvement is to make the model robust to such changes for clean inputs, so that accuracy improves. We achieve this by incorporating the blocking mechanism during training: a small region (30x30) of the training image which is responsible for the class label is replaced with a black patch. This also acts as a regularizer forcing the model to consider larger regions of the image while making a prediction and not rely on small regions (such as the trigger) to influence the decision-making process. We refer to this procedure as **Attn Blocking**. An added bonus of the sequence-to-sequence approach of the transformer is that we can train the model with reduced number of tokens. This

enables us to simply drop the tokens corresponding to the input region rather than masking it. [18] showed that this not only reduces computation, but also improves the accuracy since the model learns to understand the real distribution of images, rather than the unnatural masked image. We call this **Token drop**. We show our results in Table 4. We observe that both variants perform favourably better compared to the vanilla network and the token drop method improves accuracy significantly. We also see a reduction in ASR before defense, indicating the regularization effect due to the modified training regime. By performing the defense at test time, we see a further improvement in robustness.

4.2. Discussion

In this section, we present some additional experiments related to some assumptions made in the defense such as the blocking area and considering a non-patch based attack.

Variation in blocking area: In our test-time defense, the defender needs to make an assumption on the maximum size of the trigger encountered at test time. We believe this is a reasonable assumption since trigger sizes are usually small. We conduct an experiment to study the dependency of the blocking area used on Attack Success Rate for patched images and Source Accuracy on clean images. As seen in Figure 4, we vary the size of blocking area from 10x10 to 70x70 and find that variation in Source Accuracy is small. We get the lowest ASR when the block size equals trigger size (30x30), but more importantly, we find that for all region sizes, ASR is lower for the defended case compared to the baseline (not defended). This suggests that the such a localization defense can be used even when the defender has limited knowledge about the size of the trigger, with a relatively small drop on clean data performance.

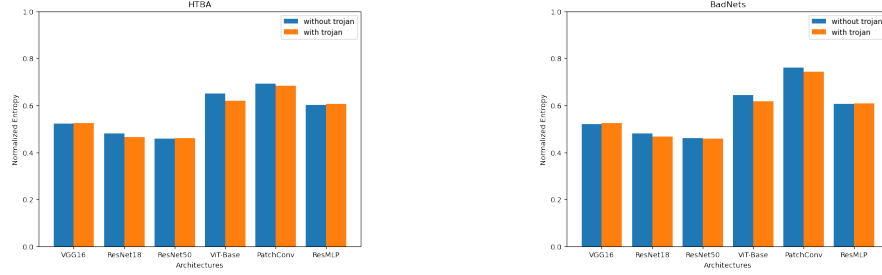


Figure 3. **Detecting Backdoor examples:** We find that an entropy based backdoor detection method is not very suitable to more diverse datasets such as ImageNet and for large architectures. We can see that the difference in entropy for the benign and trojan examples is not significant enough to create a heuristic based defense.

Model	Attack	Source Accuracy (%)	ASR (%)
ViT-Base	WaNet	63.20	41.80
ViT-Base (Random blocking)	WaNet	64.00	39.60
ViT-Base (Attn Blocking)	WaNet	63.40	36.00

Table 5. **WaNet Defense:** We consider the blocking defense for WaNet which does not involve a specific trigger. We find that attention based blocking reduces the effect of the attack. We consider a random location blocking as a baseline.

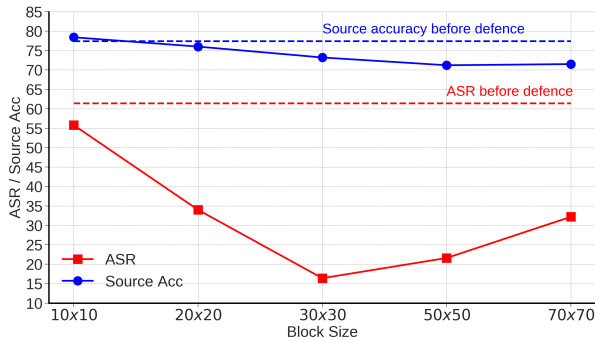


Figure 4. **Dependency on blocking area:** We find that for different block sizes in defense, the source accuracy does not vary much. We can also see that defended ASR is consistently lower compared to the baseline of no defense, suggesting that a simple localization defense can be useful even with limited knowledge about the trigger.

Defending against stealth based trigger attacks: We also consider WaNet [28] as a stealth based or human invisible trigger attack where the trigger is not easily visible as a patch and evaluate the attention blocking defense as shown in Table 5. We see that although there is no trigger visible, the blocking mechanism can still reduce ASR. The intuition is that the attention highlights the region of the image most responsible for target prediction and hence most effective in moving the feature embedding to the target category. By blocking out this region, we are reducing the effectiveness of the attack, although to a lesser extent. We consider a baseline where we block a random region and find that to be less effective than the attention-based blocking.

Detection-based algorithms: The backdoor literature has multiple defense proposed to overcome attacks at test time [13, 19, 47]. We consider one such defense, STRIP [13] as a

SOTA run-time backdoor detection model to evaluate test-time defense on both HTBA and BadNets. The idea behind the STRIP is to apply a perturbation on the input and measure the randomness (entropy) of the output prediction of the model. Less randomness (low entropy) in the final prediction indicates presence of a backdoor in the input. Following [13], for each sample at test time, we select 100 randomly chosen clean images to apply linear blending with the input. For each architecture, we average the normalized entropy over all source samples shown in Fig. 3. We observe that entropy is not a good indication of a trigger for a complex dataset like ImageNet. For example, in some architectures like ResMLP and ResNet, the difference between benign and trojan examples is not significant enough. Note that the STRIP algorithm is a detection based defense where the goal is to detect whether an input sample contains a backdoor.

5. Conclusion

We show that existing backdoor attacks are effective against Vision Transformers, while ResMLP is more robust to backdoor attacks. We find that in transformers, GradRoll-out interpretation method effectively highlights the trigger patch for a backdoor test-input even though the model never sees the trigger during training. Based on this observation, we empirically show that a test-time image blocking defense effectively reduces ASR. Our work indicates that attention mechanisms can be both helpful and harmful in the context of backdoor robustness. We hope our results encourage the community to study and understand architectural components of vision architectures affecting backdoor robustness.

Acknowledgement: This work is partially supported by NSF grants 1845216 and 1920079.

References

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. [5](#)
- [2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021. [2](#)
- [3] 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*, pages 213–229. Springer, 2020. [2](#)
- [4] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1148–1156, 2021. [2](#)
- [5] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54. IEEE, 2020. [2](#)
- [6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [2](#)
- [7] 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. [2](#)
- [8] 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](#)
- [9] Khoa Doan, Yingjie Lao, and Ping Li. Backdoor attack with imperceptible input and latent modification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18944–18957. Curran Associates, Inc., 2021. [2](#)
- [10] 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. [3](#)
- [11] 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. In *International Conference on Learning Representations*, 2020. [1](#), [3](#)
- [12] Chen et al. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv:1712.05526*. [1](#)
- [13] Gao et al. Strip: A defence against trojan attacks on deep neural networks. [2](#), [8](#)
- [14] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20876–20885, 2022. [2](#)
- [15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. [2](#)
- [16] 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](#)
- [17] Jonathan Hayase, Weihao Kong, Raghav Somani, and Se-woong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pages 4129–4139. PMLR, 2021. [2](#)
- [18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. [7](#)
- [19] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *CVPR*, 2020. [2](#), [8](#)
- [20] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472, 2021. [2](#)
- [21] Cong Liao, Haoti Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *arXiv preprint arXiv:1808.10307*, 2018. [2](#)
- [22] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017. [2](#)
- [23] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020. [2](#)
- [24] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48. IEEE, 2017. [2](#)
- [25] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15234–15243, June 2022. [2](#)
- [26] 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. [3](#)
- [27] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. [2](#)
- [28] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. [1](#), [2](#), [4](#), [8](#)
- [29] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. [2](#)

- [30] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020. [2](#)
- [31] 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)*, 2015. [4](#)
- [32] 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. [1](#), [2](#), [4](#)
- [33] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 703–718. IEEE, 2022. [2](#)
- [34] Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, R Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. [2](#)
- [35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. [5](#)
- [36] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018. [2](#)
- [37] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. [2](#)
- [38] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *NeurIPS*, 2019. [2](#)
- [39] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. [3](#)
- [40] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021. [2](#), [3](#), [4](#)
- [41] 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*, pages 10347–10357. PMLR, 2021. [3](#)
- [42] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, and Hervé Jégou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021. [3](#)
- [43] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021. [1](#)
- [44] 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 (ICCV)*, pages 32–42, October 2021. [1](#)
- [45] 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. [3](#)
- [46] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018. [2](#)
- [47] 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)*. IEEE, 2019. [2](#), [8](#)
- [48] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Z Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPR workshops*, 2020. [2](#)
- [49] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, and Ting Wang. Backdoor attack through frequency domain. *arXiv preprint arXiv:2111.10991*, 2021. [2](#)
- [50] Kota Yoshida and Takeshi Fujino. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security*, pages 117–127, 2020. [2](#)