# Towards Understanding and Improving Adversarial Robustness of Vision Transformers

Samyak Jain
Indian Institute of Technology (BHU) Varanasi
samyakjain.cse18@iitbhu.ac.in

Tanima Dutta
Indian Institute of Technology (BHU) Varanasi
tanima.cse@iitbhu.ac.in

## Abstract

*Recent literature has demonstrated that vision transformers (VITs) exhibit superior performance compared to convolutional neural networks (CNNs). The majority of recent research on adversarial robustness, however, has predominantly focused on CNNs. In this work, we bridge this gap by analyzing the effectiveness of existing attacks on VITs. We demonstrate that due to the softmax computations in every attention block in VITs, they are inherently vulnerable to floating point underflow errors. This can lead to a gradient masking effect resulting in suboptimal attack strength of well-known attacks, like PGD, Carlini and Wagner (CW) and GAMA. Motivated by this, we propose Adaptive Attention Scaling (AAS) attack that can automatically find the optimal scaling factors of pre-softmax outputs using gradient-based optimization. We show that the proposed simple strategy can be incorporated with any existing adversarial attacks as well as adversarial training methods and achieved improved performance. On VIT-B16, we demonstrate an improved attack strength of upto* 2.2% *on CIFAR10 and upto* 2.9% *on CIFAR100 by incorporating the proposed AAS attack with state-of-the-art single attack methods like GAMA attack. Further, we utilise the proposed AAS attack for every few epochs in existing adversarial training methods, which is termed as Adaptive Attention Scaling Adversarial Training (AAS-AT). On incorporating AAS-AT with existing methods, we outperform them on VITs over* 1.3-3.5% *on CIFAR10. We observe improved performance on ImageNet-100 as well.*

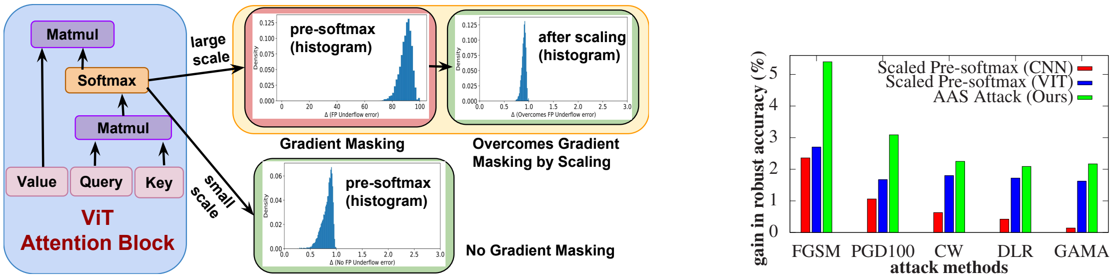
## 1. Introduction

In recent years, the rise of deep neural networks (DNNs) has been accompanied by a concerning vulnerability to imperceptible perturbations known as Adversarial Attacks [17, 32]. These subtle perturbations can mislead DNN predictions with potentially severe consequences. Efforts to bolster DNN robustness against such attacks have led to numerous defense mechanisms. However, many defenses succumb to a challenge known as gradient masking, creating a false sense of security [6, 20, 30, 39, 42]. Some defenses employ randomized or non-differentiable elements, hindering precise gradient calculation and thwarting strong adversarial attacks. This phenomenon raises concerns about the efficacy of defense strategies [2, 9, 34] and have demonstrated that these defenses can be bypassed by adaptive attacks specifically tailored to the targeted model.

Projected Gradient Descent (PGD) [24] was an early popular attack, but subsequent methods like like GAMA [31], Carlini and Wagner (in short CW) [8], and AutoAttack [10] have demonstrated superior performance. Yu and Xu [40] discovered that proper scaling of logits in the output space allows PGD to perform similarly. The authors characterized that taking *softmax* leads to **floating point underflow errors**. Attacks, like GAMA and CW, which avoid softmax, are naturally resilient to these errors, possibly contributing to their success in CNNs. Further, Hitaj et al. [21] found that adversarial training methods like GAIRAT [42] can enhance logits magnitude, thwarting attacks. However, scaling logits down [7, 11, 21] results in strong attacks, significantly reducing GAIRAT's robustness.

Transformers have set new benchmarks in various tasks [5, 14, 28, 35], and Vision Transformers (VITs) [15, 33] have shown improved performance over CNNs. However, a debate persists on whether VITs are inherently more robust. Past studies [4, 25, 29] suggested VITs exhibit lower attack transferability in black box settings, implying greater robustness than CNNs. Yet, recent work [3] challenged this, demonstrating that, under strong attacks (e.g. AutoAttack) and activation functions, like GELU in CNNs, transformers are no more robust than CNNs. Naseer et al. [27] improved adversarial transferability for VITs through classifier training and backpropagation, while Wei et al. [37] showed dropping input patches enhances attack transferability. Despite these findings, there's still a lack of a fundamental understanding of why generating strong attacks on VITs is challenging.

In this work, we rethink the understanding of the adversarial robustness of VITs and discuss the fundamental causes of gradient masking in VITs that leads to poor attack strength



(a) (b)

Figure 1. **Floating point underflow errors in attention blocks of a VIT lead to gradient masking. (a)** $\Delta$ is the difference between the largest and the second-largest pre-softmax output. As highlighted in red, if the scale of the pre-softmax outputs is high, floating point underflow error occurs. It will not occur if the scale is low (as highlighted in green). Motivated by this, we propose to downscale the pre-softmax outputs. **(b)** Gains in PGD-100 robust accuracy on using AAS-attack (green) on adversarially trained (PGD-AT [24] with AWP [38]) VIT-B16 model [15] are higher as compared to manually finding the scaling factors (blue). In case of CNNs (red, the logits of adversarially trained (PGD-AT [24]), WideResNet-28-10 are downscaled.

on using standard attacks, like PGD [24], GAMA [31], and CW [8] attacks, these attacks have demonstrated good attack strength on CNNs. Alike Yu and Xu [40], we also assume that the reason for gradient masking is VITs is the floating point underflow error which occurs due to softmax calculation in every attention block in VITs. In Figure 1 (a), as highlighted in red, we observe that in the case of VITs, a larger scale of pre-softmax outputs can result in floating point underflow. This leads to a false estimate of the gradient, resulting in a weaker attack. But, we can overcome the floating point errors by scaling down the pre-softmax outputs by the right scaling factor. As shown in Figure 1 (b), on scaling down the pre-softmax outputs in every attention block the proposed Adaptive Attention Scaling (AAS) attack leads to a boost upto 3% over standard PGD [24] attack on CIFAR10 dataset. In the case of CNNs (shown in red) as demonstrated by [40], scaling down the logits help in improving the attack strength by overcoming floating point errors. The effect of gradient masking is much more intense in the case of VITs because of softmax functions present in attention blocks. This is evident from the improved gains on simply scaling the pre-softmax outputs manually (shown in blue) over the same method applied to logits in the output space of CNNs (shown in red). We show strong empirical evidence for our hypothesis and point out a more fundamental reason for the poor performance of VITs on adversarial attacks.

To achieve robustness against strong adversarial attacks, the most popular approach is adversarial training (AT). While it is easy to train CNNs [24, 38, 41] using adversarial training, but VITs seem to pose multiple challenges [12, 26]. Recently, Mo et al. [26] demonstrated that training a VIT from scratch doesn't converge to a good solution. Therefore, a pre-trained initialization is necessary for training VITs using AT. The authors also showed that to stabilize the adversarial training of VITs, gradients need to be clipped. These clipping or pretraining is not required in case of CNNs. Further, using complex augmentations like Cutmix and Mixup gives improved results. Similarly, Debenedetti [12] proposes a training recipe to improve VITs robustness, where a ten epoch linear $\epsilon$ warmup is used along with high weight decay to get improved performance. While these tricks, like gradient clipping, warmup, and high weight decay, improve robustness, it is not well understood why these tricks are needed. Though it is very easy to train CNNs using adversarial training, it seems difficult to train VITs.

We consider that the floating point underflow error not only leads to weaker attack generation during inference but also during training. This results in suboptimal adversarial robustness on performing adversarial training on VITs. Motivated by this, we propose a new adversarial training method, named as Adaptive Attention Scaling Adversarial Training (AAS-AT), where we ensure that the scale of the logits doesn't exceed too much. We demonstrate that this simple check by using the proposed AAS attack at regular intervals of training helps in stabilizing the training of VITs and results in improved robustness. We demonstrate that the proposed training method AAS-AT can be combined with different existing adversarial training methods leading to improved performance.

The contributions of this work are listed as follows:

- The first contribution of our work is to highlight the presence of gradient masking in adversarially trained VITs. More importantly, we precisely point out the reason for it. We demonstrate that the floating point underflow error is caused due to softmax operations in attention blocks. It leads to weak attack generation in case of VITs. We consider this as fundamental cause for weaker white box attacks on VITs.
- We propose a novel attack, named as Adaptive Attention Scaling (AAS), that automatically finds the optimal scaling values for pre-softmax outputs in attention blocks, thus mitigating floating point underflow error. In recent literature, it is unclear if scaling the pre-softmax outputs in the case of VITs should have significant changes in the strength of these attacks. But we observe over 2% improved attack strength

in the case of VITs on combining the proposed AAS attack with the GAMA-PGD attack.

- We have shown that the use of standard loss functions, like cross-entropy and max-margin, are not able to optimize the pre-softmax scaling factors well. Whereas, LPIPS distance, a feature-level distance, helps to generate a stronger attack by estimating the distance between a normal model and a model with perturbed pre-softmax output scaling factors. We maximize LPIPS distance in feature space to get perceptually aligned gradients and find optimal scaling factors. This eventually overcomes gradient masking and enhance adversarial robustness.
- We propose a robust training model, known as Adaptive Attention Scaling Adversarial Training (AAS-AT), that combines the proposed AAS attack to make the VITs more robust. We show the proposed AAS attack and AAS-AT model can be combined with any existing adversarial attack and adversarial training method respectively. This ensures that the proposed method is generalizable and widely applicable.
- Lastly, we highlight our empirical contributions against PGD, CW and GAMA attacks. By simply combining AAS-AT with standard PGD-AT, we achieve improved performance over recent works. We demonstrate improved results as compared to existing methods on CIFAR10 and CIFAR100 datasets with an improvement of over 1.3%. We scale our model for large datasets and got improved results also on ImageNet-100.

The remaining paper is organized as follows. Preliminaries and motivation are discussed in next two sections. Thereafter, we elaborate the proposed AAS attack and AAS-AT in two sections. Lastly, we show experimental results and conclude the paper in last two sections.

## 2. Preliminaries

**Adversarial Training (AT).** Adversarial Training is considered as the most successful defence strategy. AWP-Trades [38] is considered as one of the most successful adversarial training method for CNNs; however, it has not been successful for VITs [26]. Mo et al. [26] showed the importance of using pre-trained initializations for training VITs adversarially, which helped in improving the training convergence on using existing methods like PDG [24] and Trades [41]. Debenedetti [12] showed that using a larger value of weight decay and a few initial epochs of epsilon warmup can help in improved adversarial robustness. Debenedetti [12] demonstrated that aforesaid tricks helps in enhancing the robustness of VITs significantly. In this work, we demonstrate that our Adaptive Attention Scaling Adversarial Training (AAS-AT) can be incorporated with any existing AT methods to achieve improved robustness.

**Threat model.** Let $f_\theta$ denote a deep neural network parameterized by $\theta$ mapping input sample $X$ to $R^N$ where $N$ is the number of classes. The goal of an adversary is to fool the model while restricting the perturbation within a threat model. The threat model is defined by:

$$||X^{'} - X||_p < \epsilon \text{ and } f_\theta : X \rightarrow R^N, \tag{1}$$

where $p$ represents the type of $\ell_p$ perturbation norm, $X^{'}$ represents the perturbed image and $\epsilon$ is the maximum allowed $\ell_p$ perturbation bound. In this work, we consider $\ell_\infty$ perturbation norm.

**Background.** Some of the past findings that motivate our proposed attack are discussed as follows:

- It is known that the gradients from a robust model are perceptually aligned [1, 22]. Recently, Ganz et al. [16] demonstrated that if the gradients from a model are perceptually aligned, then it implies that the model is adversarially robust.
- Through extensive human evaluation, Zhang et al. [43] demonstrated that LPIPS distance is a good perceptual model. Motivated by this, it has been further used in Laidlaw et al. [22] to define a perceptual threat model. Addepalli et al. [1] proposed to minimize the LPIPS distance between the clean and adversarial images to achieve robustness to larger perturbation bounds by ensuring that the images don't change their original class perceptually.

## 3. Motivation: Gradient Masking in VITs

Yu and Xu [40] demonstrated that the attacks involving softmax calculation may suffer from floating point underflow error leading to a suboptimal attack. Further, VITs use softmax to calculate the attention weights in every attention block. Therefore, the effect of floating point error should be more severe in the case of VITs as compared to CNNs. Based on this, we propose the following conjecture:

**Conjecture 1:** Presence of intermediate softmax in attention blocks of vision transformers makes them inherently vulnerable to the gradient masking effect.

**Justification:** To verify this, we plot the histogram of the difference between the largest and the second largest values before taking softmax ($\Delta$) for different attention blocks of the VIT-B16 [15] model in Figure 2 (first row). The histogram is plotted for the ImageNet-100 dataset, which is a 100-class random subset of ImageNet-1K [13] using a normally trained VIT-B16 model. We observe that for many images, the difference is significantly high and would lead to floating point underflow errors on taking exponential in the softmax calculations. As opposed to CNNs, since the floating point underflow error will occur in intermediate layers of VITs, the effect would simply magnify. Therefore, depending on the amount of floating point error, it can even lead to significant gradient masking. On using the proposed attack, as observed in Figure 2 (second row), the scale of the pre-softmax outputs gets significantly reduced, which helps

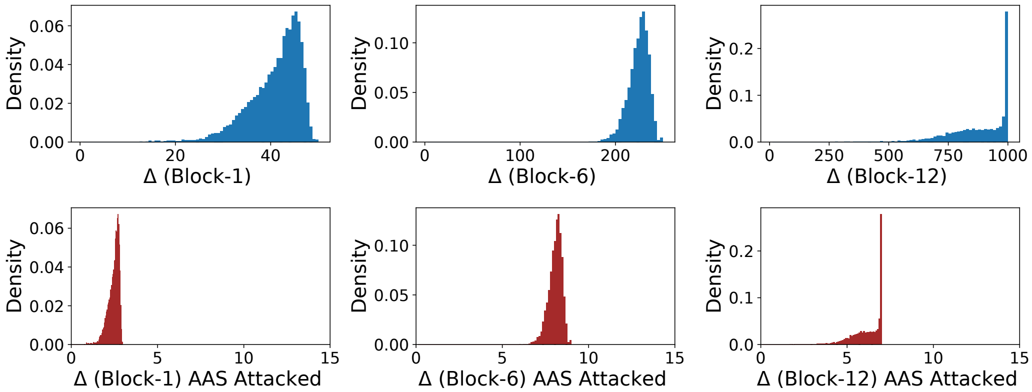


Figure 2. **Histogram of the difference between largest and second largest pre-softmax outputs** for different blocks of the normally trained VIT-B16 on ImageNet-100 dataset. As shown in the second row, the proposed AAS attack successfully downscales these values.

| Scaling | CIFAR10 | | CIFAR100 | |
|---|---|---|---|---|
| Factor | Clean | PGD-100 | Clean | PGD-100 |
| 1 | **87.43** | 61.10 | **62.47** | 30.01 |
| $10^{-4}$ | 86.23 | 60.23 | 61.13 | 28.45 |
| $10^{-3}$ | 86.79 | 60.14 | 61.42 | **27.76** |
| $10^{-2}$ | 87.01 | **59.43** | 61.79 | 28.01 |
| $10^{-1}$ | 87.22 | 60.87 | 62.04 | 28.12 |
| 10 | 85.71 | 66.78 | 61.03 | 31.78 |

Table 1. **Manually scaling down the pre-softmax outputs** leads to enhanced attack strength of PGD attacks on CIFAR10 and CIFAR100 datasets.

in overcoming the gradient masking effect. To understand this, we define the pre-softmax scaling factor as follows:
**Definition of Pre-Softmax Scaling Factor ($\lambda$):** Consider that the keys and queries are $R^d$. Let the set of queries be packed into a matrix denoted by $Q$, and the matrices for keys and values be $K$ and $V$. Let the pre-softmax scaling factor be $\lambda$; then we define the attention as follows:

$$\textbf{Attention}(Q, K, V) = \textbf{softmax}(\lambda QK^T/\sqrt(d))V. \quad (2)$$

We now analyze the effect of manually scaling down the features used for softmax computation (pre-softmax scaling factor $\lambda$) on CIFAR10 and CIFAR100 datasets in Table 1. We observe that even using the same scaling factor for all the attention blocks can give a boost of up to 1.67% in the PGD-100 and 2.7% in FGSM attack strengths on the CIFAR10 dataset. On CIFAR100, improved attack strength of up to 2% is observed on the PGD-100 attack.

## 4. Adaptive Attention Scaling (AAS) Attack

The process of finding the optimal combination of scaling factor is difficult because different attention blocks can have different scaling factors. We analyze if these scaling values can be found automatically using gradient-based optimization before generating the attack. Zhang et al. [43] performed a human study and demonstrated that LPIPS distance is a good perceptual metric. The LPIPS distance is a feature-level distance defined for a set of inputs to a given model. For a given set of clean ($x$) and adversarial ($x^{'}$) images, LPIPS distance is the sum of the normalized $\ell_2$ distances between the features of the two images taken after every attention block. A greater value of LPIPS distance indicates that the two images are perceptually dissimilar to each other [1, 22].
**Definition of LPIPS in terms of pre-softmax output:** In this work, instead of using a set of clean and adversarial images for calculating LPIPS distance, we perturb the pre-softmax output scaling factors denoted by $S = \{s_1, s_2, ..., s_m\}$ and calculate the LPIPS distance for a given set of normal and perturbed models. Here, $m$ is the number of attention blocks in the VIT. Thus, for an image $x$, the LPIPS distance, in our case, is defined by:

$$\textbf{LPIPS}(f_{\theta(S)}, f_{\theta(S')}) = \sum_{i=1}^{m} \frac{||f_{\theta(s'_i)}(x) - f_{\theta(s_i)}(x)||_2}{||f_{\theta(s'_i)}(x)||_2||f_{\theta(s_i)}(x)||_2}. \quad (3)$$

It is well know that the gradients calculated using an adversarially robust model are perceptually aligned [1, 22]. Through a human study, Zhang et al. [43] demonstrated that LPIPS is a good perceptual metric. Thus, maximizing LPIPS distance while perturbing the pre-softmax scaling factors should lead to finding the scaling factors which can produce gradients that are more perceptually aligned. As demonstrated by Ganz et al. [16], perceptually aligned gradients can imply adversarial robustness. Therefore, by making the gradients more perceptually aligned by maximizing the LPIPS distance between the original and perturbed models, we tend to overcome gradient masking and enhance the adversarial robustness of the model. As demonstrated in Addepalli et al. [1] (Figure S3), LPIPS distance between the clean images and the corresponding adversarial images perturbed using the adversarial attack generated from a standard model is less as compared to LPIPS distance between clean images and corresponding adversarial images generated by attacking a robust model. Therefore, maximizing LPIPS distance between the features of a normal and perturbed model should lead to the generation of perceptually aligned gradients, thereby overcoming the gradient masking effect.

Table 2. **Effect of perturbing the scaling factors using different loss functions**. A feature level distance like LPIPS [43] leads to better attack strength by overcoming gradient masking effectively.

| Loss Functions<br>(Perturb Scales) | CIFAR10 | | CIFAR100 | |
|---|---|---|---|---|
| | Clean | PGD-100 | Clean | PGD-100 |
| No Attack | **87.43** | 61.10 | **62.47** | 30.01 |
| Scaling (0.01) | 87.01 | 59.43 | 61.79 | 28.01 |
| Cross-Entropy | 86.84 | 60.13 | 61.48 | 29.11 |
| Max-Margin | 86.47 | 59.78 | 61.23 | 28.46 |
| GAMA | 86.74 | 59.85 | 61.41 | 28.31 |
| LPIPS (ASA) | 87.31 | **58.01** | 62.03 | **27.02** |

**Algorithm 1** Adaptive Attention Scaling (AAS) Attack

```
1: Input: Network $f_{\theta(S)}$ where $S = \{s_1, s_2, ..., s_m\}$ is the pre-softmax scaling factor and $m-1$ is the number of attention blocks in the model. Training Dataset $\mathcal{D} = \{(x_i, y_i)\}$, and $M$ training mini-batches of size $n$;
2: for iter $= 1$ to $M$ do
3:     $\delta = \mathcal{N}(0, 1)$;
4:     for steps $= 1$ to $10$ do
5:         $\delta = \delta + \nabla_S$ LPIPS$(f_{\theta(S)}(x_i), f_{\theta(S')}(x_i))$;
6:         $S^{'}$ = Clamp$(S+\delta, 10^{-r}, 1)$; % to prevent zero scaling factors, we considered $r = 7\%$
7:     $S = S^{'}$;
8: Generate the attack on the perturbed model;
```

• **Why LPIPS is better?** To understand whether maximizing a feature level distance, like LPIPS, is better than other standard attacks that use cross-entropy or max-margin loss, we analyze the effect of using different loss functions to perturb the scaling factors of pre-softmax outputs in Table 2. Since gradient masking is present at *feature level* in VITs, using a feature-level attack, like LPIPS, can give improved performance by up to $1-2\%$ over standard attacks. We observe that manually scaling the pre-softmax outputs outperforms the standard attacks, like GAMA, PGD, CW. This shows the importance of attacking at the feature level rather than the output space. Motivated by the above discussion, we propose to maximize LPIPS distance for perturbing the pre-softmax output scaling factors. The proposed AAS attack is presented in Algorithm 1. As common in practice, we initialize the attack with a standard normal distribution (L3). We perturb the pre-softmax output scaling factors in the attention blocks by maximizing the LPIPS distance before generating the actual attack (L4-L7). Later, the adversarial attack is generated using the model whose scaling factors is perturbed (L8).

## 5. Adaptive Attention Scaling Adversarial Training (AAS-AT)

Motivated by Conjecture-1, since VITs are inherently subjected to the gradient masking effect, adversarial training of VITs should be difficult. This is indeed observed in the prior works [12, 26], which demonstrate the need to use gradient clipping, larger weight decays and epsilon warmup to stabilize and achieve improved robustness on VITs by performing adversarial training. Though these tricks help in stabilizing the VITs, the reason for their effectiveness is not well known. Motivated by our previous discussion, we hypothesize that the reason for the sub-optimal adversarial robustness on training VITs is the gradient masking effect caused due to the large scale of pre-softmax outputs. To mitigate the associated gradient masking effect, we propose to use the proposed AAS attack in every few epochs of training while training on standard attacks for the remaining epochs. More specifically, we propose to perturb the pre-softmax scaling weights for every few epochs using LPIPS loss maximization to ensure that the scale of the pre-softmax values is within the suitable range. This will help in preventing floating point underflow errors resulting in a better estimate of the gradients and stronger attack generation during training. As shown in Algorithm-1 in Supplementary, the proposed AAS-AT is build on Trades AT [41] with an added scaling factors perturbation step (Lines 8-10) every $\lambda$ epochs. If the task was to perturb the pre-softmax output scaling factor, then the old scaling factors are reinitialized using the new perturbed ones (Line 17), otherwise, following Trades, the cross-entropy loss on the clean images and KL Divergence between the clean and the adversarial images is minimized (Lines 19-20).

## 6. Experimental Results

In this section, we present the results of the proposed Adaptive Attention Scaling (AAS) attack and Adaptive Attention Scaling Adversarial Training (AAS-AT). The performance is evaluated on CIFAR10, CIFAR100 and ImageNet-100 datasets, where ImageNet-100 is a random 100 class subset of ImageNet-1K. In all the experiments, we use an $\ell_\infty$ norm threat model of perturbation bound $8/255$. PGD-100, CW, DLR, GAMA use 100 iterations for generating the attack, whereas FGSM is a single-step attack. For training, we utilize a 10-step attack for all Adversarial Training methods. Training is done using VIT-B16 model (unless specified) using additional synthetic data generated from diffusion models (DDPM) [18] in case of CIFAR10 and CIFAR100. To evaluate the attack, we use VIT-B16 model trained on the standard PGD-AT [24] with AWP [38] for weight space smoothing. We use RTX-2080 and V100 GPUs for all experiments. Further training details of individual adversarial training methods and reproducibility evaluations (like reruns

Table 3. **Comparison of AAS attack** with different attack methods on CIFAR-10, CIFAR-100 and ImageNet-100 datasets.

| **Data** | **Attack** | **Clean Acc** | **Robust Acc** | **Clean Acc** | **Robust Acc** | **Clean Acc** | **Robust Acc** |
|---|---|---|---|---|---|---|---|
| | | | | **+ Scale** | **+ Scale** | **+ AAS** | **+ AAS** |
| CIFAR10 | FGSM | 87.43 | 66.48 | 87.01 | 63.78 | 87.31 | **61.04** |
| | PGD-20 | | 63.14 | | 60.64 | | **58.21** |
| | PGD-100 | | 61.10 | | 59.43 | | **58.01** |
| | CW | | 59.98 | | 58.13 | | **57.73** |
| | DLR | | 60.03 | | 58.31 | | **57.94** |
| | GAMA | | 59.78 | | 58.16 | | **57.61** |
| CIFAR100 | FGSM | 62.47 | 33.46 | 61.42 | 30.78 | 62.03 | **28.03** |
| | PGD-20 | | 30.78 | | 28.31 | | **27.21** |
| | PGD-100 | | 30.01 | | 27.76 | | **27.02** |
| | CW | | 29.03 | | 27.08 | | **26.31** |
| | DLR | | 29.21 | | 26.95 | | **26.42** |
| | GAMA | | 28.97 | | 26.64 | | **26.08** |
| ImageNet-100 | FGSM | 68.03 | 32.06 | 67.36 | 29.47 | 67.84 | **28.79** |
| | PGD-20 | | 30.02 | | 27.13 | | **26.41** |
| | PGD-100 | | 29.75 | | 27.03 | | **26.12** |
| | CW | | 28.69 | | 26.48 | | **25.81** |
| | DLR | | 28.71 | | 26.71 | | **25.90** |
| | GAMA | | 28.07 | | 26.15 | | **25.64** |

of AAS-AT) are presented in the Supplementary.

Table 4. **Incorporating AAS-AT with Trades** on ImageNet-100.

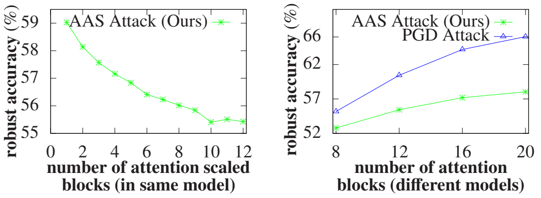
| **Method** | **Clean Accuracy** | **Auto-Attack** |
|---|---|---|
| PGD-AT | **68.32** | 25.78 |
| PGD-AT + Ours | 68.02 | **27.32** |
| Trades | **65.48** | 26.46 |
| Trades + Ours | 65.26 | **28.04** |

### 6.1. Evaluation of the proposed AAS attack

The results of the proposed Adaptive Attention Scaling (AAS) attack on CIFAR10, CIFAR100 and ImageNet-100 datasets are presented in Table 3. The bracket shows the scaling factor. The results are shown on incorporating AAS with different existing attack methods. In each of them w/o ASA/Scale represents the original accuracy achieved by the respective attack itself. Whereas + Scale represents the accuracy achieved on manually finding the best possible scaling factor for GAMA attack in the set $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The same scaling factor is used for all the attention blocks, and separate tuning for each attention block is not performed because it is $O(m^t)$ where $m$ is the attention blocks and $t$ is the size of the scaling factor set, which will be computationally very expensive. Finally, + AAS represents the performance of the respective attack on combining with the proposed Adaptive Attention Scaling attack. Since on performing scaling (+ Scale) or the proposed attack (+ AAS), the function mapping of the model will change; therefore, clean accuracy will also change.

As observed in Table 3, the clean accuracy drops by upto $1\%$ when finding the scaling factors manually. But on using AAS attack, this decrease is not more than $0.45\%$. On the other hand, on CIFAR10, the robust accuracy decreases by upto $5\%$ in the case of FGSM and PGD-20 attacks. Even for stronger attacks like CW, DLR and GAMA, an improved attack strength of up to $2\%$ is observed on CIFAR10. It is also observed that simply scaling the pre-softmax outputs can also help in improved attack strength by upto $1.6\%$ in case of the strongest GAMA attack, but since the scaling factors found in this fashion might not be optimal, therefore finding them using gradient-based optimization as used in AAS attack helps in a further boost of around $0.51\%$. Even on CIFAR100, on combining the proposed AAS attack with GAMA, it shows $2.89\%$ improved attack strength, whereas scaling shows $2.3\%$ improvements. We also compare the performance of the proposed AAS attack on ImageNet-100 where we observe around $3.5\%$ improvements over PGD-100 and $2.2\%$ improved results over the GAMA attack. As can be seen from this analysis, by overcoming the floating point underflow errors, our AAS attack gives consistent gains over the existing attacks.

- **Ablation experiments.** As shown in Figure 3 (a), on increasing the number of attention blocks in which the pre-softmax values are scaled using the proposed AAS attack, the robust accuracy on the CIFAR10 dataset falls continuously. Since floating point errors occur in each of the attention blocks, therefore when scaling is done for a larger number of attention blocks, the effect of gradient making is minimized, thus leading to stronger attacks. Further, as shown in Figure 3 (b), if the size of the model is increased by adding up more attention blocks, the drop in robust accuracy of the pro-



(a) (b)

Figure 3. **Ablation of AAS attack on CIFAR10. (a)** Effect of increasing the number of attention blocks used in AAS attack **(b)** Comparison between PGD and AAS attack on increasing the size of the model. Using a larger model with more attention blocks leads to larger gradient masking.

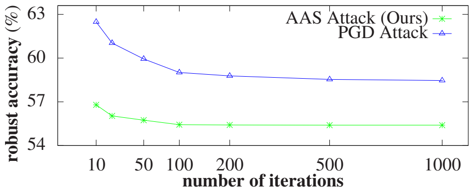


Figure 4. **Comparison between PGD and AAS attack** on increasing the number of attack iterations. Using a larger number of iterations also doesn't decrease the gap between the attack strengths.

posed AAS attack with respect to PGD-100 further increases. This demonstrates the effectiveness of overcoming gradient masking by using our AAS attack. Finally, we present the effect of increasing the number of iterations of attack for PGD and AAS in Figure 4. AAS attack saturates earlier than PGD, and PGD is not able to close up the gap between two attacks even on 1000 iterations. To confirm, this further, we observe that AAS attack indeed leads to less noisy gradients as compared to PGD in Figures-2, 3 of Supplementary. On ImageNet-100 in Table 4, we observe over 1.58% improved robust accuracy against Auto-Attack on using the proposed AAS-AT along with Trades. To emphasize that our claims can generalize to other VIT architectures, we present results of AAS attack on Swin Transformer (Swin-T) [23] and LeViT [19] in Table-3 of Supplementary. As shown in Tables-1 of the Supplementary, we observe that the proposed AAS attack shows improved attack strength for different perturbation radius. We present a discussion on computational efficiency of our method in Tables-5, 6 of Supplementary.

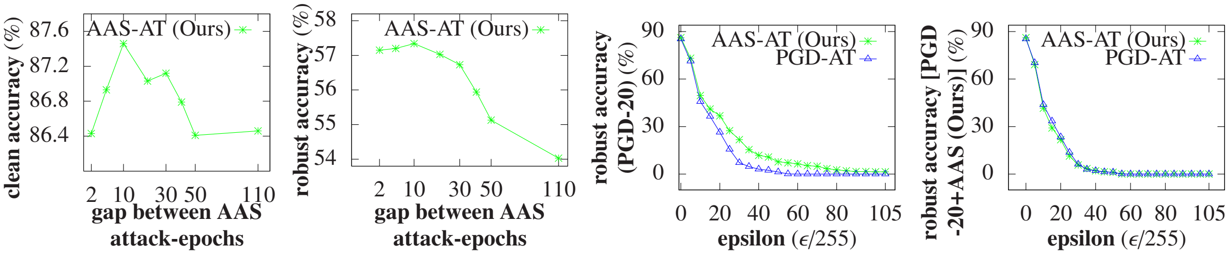
### 6.2. Evaluation of the proposed defense (AAS-AT)

As shown by Mo et al. [26], it is essential to use a pretrained initialization along with gradient clipping to enable stable and effective adversarial training of VITs. Therefore, we use ImageNet-1K initialization and gradient clipping in all our experiments. We utilize standard Pad-Crop along with Horizontal Flip as augmentations. Training is done for 110 epochs with a max learning rate of 0.1, and a cosine learning rate schedule is used for all experiments except XCIT-S12 [12]. For XCIT-S12 and XCIT-S12 + Ours, we train for 300 epochs instead. Further, SGD, along with a momentum of 0.9, is used as the optimizer in all the experiments. The simplicity of the proposed AAS-AT allows it to combine effectively with any existing adversarial training method. The results of combining the proposed AAS-AT with different adversarial training methods on CIFAR10 and CIFAR100 are shown in Table 5. We use PGD-100 [24], PGD-100 + AAS and AutoAttack [10] for evaluating the robustness of the defenses. As can be seen, AutoAttack remains the strongest white box attack. But the proposed PGD-100+AAS attack improves the attack strength by upto $4.5\%$ over PGD-100. In the case of PGD-AT + AAS-AT, the difference between PGD-100 and PGD-100+AAS is significantly reduced to only $0.34\%$. This shows that since the scale of the softmax is inherently lowered on training using AAS-AT, even PGD-100 remains effective. This demonstrates that large scaling of pre-softmax outputs indeed leads to the generation of weaker attacks. Though PGD-100 + AAS is weaker than AutoAttack [10], the difference between PGD-100 + AAS and AutoAttack is less than $1\%$ in all cases. Further, PGD-100 + AAS is significantly cheaper in terms of compute as compared to AA. This demonstrates the effectiveness of the proposed AAS attack. As shown in Table 5, it can be observed that on CIFAR10 incorporating AAS-AT with standard adversarial training methods like PGD [24] and Trades can improve the performance by upto $3.47\%$ on AA attack. Further on, incorporating AAS-AT with state-of-the-art adversarial training methods like Trades + AWP also gives an improved performance of upto $1.7\%$. On CIFAR100, we get even larger gains of upto $2.38\%$ on combining with Trades + AWP. We also demonstrate that AAS-AT can improve the performance of existing adversarial training methods, like [12, 26] that are crafted mainly for VITs. In the case of XCIT-S12 [12], AAS-AT improves the AA attack [10] performance about $1.38\%$ and $1.29\%$ for CIFAR10 and CIFAR100, respectively. On combining AAS-AT with [26], it shows gains of $1.76\%$ on CIFAR10 and $1.41\%$ on CIFAR100.

- **Ablation experiments.** As shown in Figure 5 (a,b), the proposed defense PGD-AT + AAS-AT is stable to using AAS attack every $5 - 40$ epochs. Using AAS attack too frequently or using it only once/twice in the entire training leads to suboptimal performance. The effect of varying $\epsilon$ and performing PGD-20 attack during evaluation is shown in Figure 5 (c). Since the proposed AAS-AT does not have a large scale of pre-softmax outputs, PGD-20 attack is stronger for PGD-AT + AAS-AT (this is also evident from Table 5) as compared to the baseline PGD-AT. Since PGD-AT suffers from gradient masking, its accuracy does not reach $0\%$ even on using an $\epsilon = 100/255$. But we get zero robustness on using $\epsilon = 65/255$. This shows AAS-AT does not suffer from gradient masking. Finally, in Figure 5 (d), we show that on using AAS attack along with PGD-20, the accuracy becomes zero for both PGD-AT as well as the proposed AAS-AT at $\epsilon$ close to $60/255$. Thus the proposed AAS attack is able to

Table 5. **Comparison of AAS-AT** with different adversarial training models on CIFAR10 and CIFAR100 datasets.

| **Model** | CIFAR10 | | | | CIFAR100 | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | PGD-100 | PGD-100 + AAS (Ours) | AA[10] | Clean | PGD-100 | PGD-100 + AAS (Ours) | AA[10] |
| PGD-AT [24] | **86.14** | 59.12 | 54.24 | 53.14 | 60.04 | 30.06 | 26.31 | 25.78 |
| PGD-AT + Ours | 85.32 | 58.12 | 57.78 | **56.61** | **62.06** | 29.03 | 28.34 | **27.71** |
| Trades [41] | 86.31 | 60.12 | 55.49 | 54.03 | 61.03 | 30.43 | 26.94 | 26.01 |
| Trades + Ours | **87.46** | 59.01 | 58.03 | **57.34** | **63.06** | 29.41 | 28.71 | **28.06** |
| Trades + AWP [38] | 86.21 | 60.48 | 56.84 | 56.03 | 62.78 | 31.86 | 27.94 | 27.03 |
| Trades + AWP + Ours | **87.10** | 59.78 | 58.41 | **57.73** | **63.14** | 31.76 | 30.43 | **29.41** |
| ART [36] | **86.19** | 59.13 | 54.76 | 54.12 | 62.41 | 32.61 | 27.12 | 26.47 |
| MART + Ours | 85.31 | 57.87 | 56.43 | **55.78** | **62.84** | 28.32 | 27.83 | **27.01** |
| XCIT-S12 [12] | 90.06 | 61.48 | 57.06 | 56.14 | 67.34 | 37.86 | 33.41 | 32.17 |
| XCIT-S12 + Ours | **90.78** | 59.94 | 57.84 | **57.42** | **67.12** | 35.35 | 33.97 | **33.46** |
| Mo *et al.* [26] | 86.43 | 60.03 | 56.12 | 55.03 | 61.76 | 31.30 | 27.84 | 27.01 |
| Mo *et al.* [26] + Ours | **86.71** | 58.46 | 57.44 | **56.79** | **61.43** | 30.86 | 29.16 | **28.42** |



(a) (b) (c) (d)

Figure 5. **Ablation of PGD-AT + AAS-AT on CIFAR10.** Variation of **(a)** clean and **(b)** PGD-100 (robust) accuracy with changing the gap between consecutive AAS attack epochs, respectively. **(c)** Lower robust accuracy (PGD-20) of AAS-AT and saturation to zero robustness of AAS-AT occurs at a much lower $\epsilon$ value as compared to PGD-AT. This indicates the absence of gradient masking in AAS-AT. **(d)** Using AAS attack on top of PGD-20 overcomes gradient masking, and the robust accuracy of PGD-AT decreases significantly as compared to **(c)**.

overcome the gradient masking effect observed in PGD-AT model. Additionally we present a comparison between the robustness of CNNs and VITs on incorporating the proposed AAS-AT and a discussion on computational efficiency of the proposed AAS-AT in Supplementary. We also observe that incorporating AAS-AT with existing AT methods leads to less than 1% increase in computation.

## 7. Conclusion

In this work, we demonstrate that the inherent design of attention blocks in VITs leads to floating point underflow errors, which causes weaker attack generation. To find the appropriate scaling factors for each attention block, we propose Adaptive Attention Scaling attack, which maximizes the LPIPS distance between the original and the perturbed model, where the perturbation is generated only on the pre-softmax output scaling factors. Since LPIPS distance is known as a good perceptual metric, maximizing it leads to perceptually aligned gradients, which is a characteristic of robust models [16]. We show that maximizing LPIPS distance indeed finds the appropriate scaling factors, thus overcoming gradient masking effect. We demonstrate that such an attack strategy can be integrated with any existing attack and leads to improved attack strength even on combining it with state-of-the-art single attacks, like GAMA attack. Further, we utilize this strategy in existing adversarial training methods and demonstrate improvements in robustness. We shown our empirical contributions against PGD, CW and GAMA attacks. Due to the simple design, the proposed method can be incorporated with any existing adversarial training method. Combining it with AT methods that are mainly designed for VITs also gives improved performance. We hope that by providing a fundamental understanding of gradient masking in VITs, this work will open new avenues of research in enhancing the robustness of VITs even further.

## 8. Acknowledements

We thank SERB MTR/2021/604 and ECR/2017/002419 for supporting this research.

## References

[1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and Venkatesh Babu Radhakrishnan. Scaling adversarial training to large perturbation bounds. In *The European Conference on Computer Vision (ECCV)*, pages 1–16, 2022. 3, 4

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 1–12, 2018. 1

[3] 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. 1

[4] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. 1

[5] 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*, 33:1877–1901, 2020. 1

[6] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, pages 1–22, 2018. 1

[7] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, pages 1–3, 2016. 1

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 1, 2

[9] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, pages 1–24, 2019. 1

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, page 2206–2216, 2020. 1, 7, 8

[11] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1

[12] Edoardo Debenedetti. A light recipe to train robust vision transformers. *arXiv preprint arXiv:2209.07399*, pages 1–29, 2022. 2, 3, 5, 7, 8

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1063–6919, 2009. 3

[14] 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*, pages 4171–4186, 2018. 1

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–22, 2021. 1, 2, 3

[16] Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply adversarial robustness? *arXiv preprint arXiv:2207.11378*, pages 1–21, 2022. 3, 4, 8

[17] Goodfellow. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, pages 1–11, 2015. 1

[18] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021. 5

[19] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herv'e J'egou, and Matthijs Douze. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. pages 12239–12249, 2021. 7

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, pages 1–12, 2017. 1

[21] Dorjan Hitaj, Giulio Pagnotta, Iacopo Masi, and Luigi V Mancini. Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2103.01914*, pages 1–6, 2021. 1

[22] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *International Conference on Learning Representations (ICLR)*, pages 1–25, 2021. 3, 4

[23] 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. pages 9992–10002, 2021. 7

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, pages 1–28, 2018. 1, 2, 3, 5, 7, 8

[25] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 1

[26] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *arXiv preprint arXiv:2210.07540*, pages 1–15, 2022. 2, 3, 5, 7, 8

[27] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, pages 1–24, 2021. 1

[28] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774: 1–100, 2023. 1

[29] 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*, pages 1–23, 2021. 1

[30] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, pages 1–20, 2017. 1

[31] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–12, 2020. 1, 2

[32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, pages 1–10, 2013. 1

[33] 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. 1

[34] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, pages 1–44, 2020. 1

[35] 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*, 30:1–11, 2017. 1

[36] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2020. 8

[37] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *AAAI Conference on Artificial Intelligence*, pages 2668–2676, 2022. 1

[38] Dongxian Wu, Shu-Tao Xia1, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–12, 2020. 2, 3, 5, 8

[39] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, pages 1–16, 2017. 1

[40] Yunrui Yu and Cheng-Zhong Xu. Efficient loss function by minimizing the detrimental effect of floating-point errors on gradient-based attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4056–4066, 2023. 1, 2, 3

[41] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 1–11, 2019. 2, 3, 5, 8

[42] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, pages 1–29, 2021. 1

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 3, 4, 5