Supplementary: Towards Understanding and Improving Adversarial Robustness of Vision Transformers

Samyak Jain
IIT (BHU) Varanasi
samyakjain.cse18@iitbhu.ac.in

Tanima Dutta
IIT (BHU) Varanasi
tanima.cse@iitbhu.ac.in

1. Additional details on the setup

We first train a robust vision transformer model using some adversarial training method and then perturb the pre-softmax scaling factors using the same training data so that the gradient masking is now prevented. Now, this model is given to the attacker, who can craft any attack within the threat model. The model's weights and pre-softmax scaling factors are fixed and cannot be changed/perturbed by the attacker. The perturbation of pre-softmax scaling factors is thus not adjusted as per the attacked data, but rather it is adjusted according to the training data.

2. Related Works

Adversarial Attacks. Goodfellow [15], Szegedy et al. [27] showed that DNNs are vulnerable to adversarial attacks.PGD [22] maximizes the cross-entropy loss to generate an attack. It is one of the most commonly used attacks to analyze the robustness of a defense. Carlini and Wagner [5] showed that maximizing max-margin loss instead of the standard cross-entropy leads to stronger attacks. But there have been several instances where these attacks fail to give a true estimate of the robustness. Croce and Hein [8] proposed AutoAttack, which is an ensemble of four attacks including three white box (i.e., Adaptive PGD with cross-entropy loss, Adaptive PGD with difference of logits ratio loss, Fast adaptive boundary attack [7]) and one black box (square attack [2]). AutoAttack is stronger than existing attacks. However, AutoAttack is computationally expensive because it is an ensemble of four attacks. Therefore, stronger single attacks like GAMA attack [26] have also been proposed, which are weaker than AutoAttack, but give a close estimate of the robustness. It uses a ℓ_2 norm regularizer between the outputs of clean and perturbed images along with a max-margin objective in the first few iterations of the attack. Later, only max-margin loss is maximized and the regularizer is shown to help in improved optimization of the attack.

Gradient Masking. Athalye et al. [3], Carlini et al. [6], Tramer et al. [28] demonstrated that many defenses claiming to achieve enhanced robustness can actually be broken down by using adaptive attacks. Due to gradient masking, the attacker ends up calculating a false estimate of actual gradients, resulting in a false sense of security. Alike Logit Scaling Attack [19], Yu and Xu [32] demonstrated that larger scale of logits leads to floating point underflow error. In this work, we hypothesize that the gradient masking effect due to floating point underflow errors is more intense in VITs.

Robustness of ViT models. Paul and Chen [25] focus on understanding the robustness comparison between the VITs and CNNs on common corruptions (like CIFAR-10 C and ImageNet-C) and natural adversarial image datasets (like ImageNet-A). Through extensive experiments, the authors conclude that VITs are more robust than CNNs. But the authors do not consider white-box adversarial attacks, which are known to be stronger. Our work aims to get the worst-case robustness against the stronger white-box attacks by overcoming the gradient masking effect.

The authors in [25] compare the adversarial robustness between VIT and CNN models against white box and black box attacks, but they utilize standard trained models for this comparison and do not use stronger white box attacks like Auto-Attack. They conclude that VITs are more robust than CNNs. But as mentioned the introduction of the main paper, this hypothesis has been challenged by later works [4, 24, 30].

Adversarial Training (AT). PGD-AT [22] showed that maximizing the cross-entropy loss helps to generate a multi-step attack and minimising the same for training the model helps in achieving robustness. MART [29] uses a different minimization loss for the misclassified and correctly classified examples. Trades [33] maximizes the Kullback-Leibler (KL) loss between the outputs of clean and adversarial images while minimizing the same with the cross-entropy loss on clean samples. Trades [34] demonstrated the existence of the fundamental tradeoff between clean and adversarial accuracy. Adversarial Weight Perturbations (AWP) [31] showed that perturbing the weights within a fixed ℓ_2 norm perturbation bound leads to convergence to a flatter minima. This helps to enhance robustness.

Mo et al. [23] showed the importance of using pre-trained

initializations for training VITs adversarially. The authors demonstrated the importance of using gradient clipping for stabilizing the adversarial training of VITs. They also claim to randomly remove the gradient flow through some multi-head attention modules and randomly mask the input perturbation during forward propagation.Debenedetti [10] showed that using a larger value of weight decay and a few initial epochs of epsilon warmup can help in improved adversarial robustness. Debenedetti [10] demonstrated that these tricks helps in enhancing the robustness of VITs significantly. On CIFAR100, the authors achieve significant improvement leading to a second entry on the robustness leaderboard [9]. This is the first successful demonstration that VITs can indeed achieve good adversarial robustness.

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. Apart of adversarial attacks, we also tested our models for semantic attacks and patch attacks.

3. Gradients of an Adversarially Robust Model

As shown by Addepalli et al. [1], Laidlaw et al. [20] the gradients calculated from an adversarially robust model are perceptual in nature [14]. Through a human study, Zhang et al. [34] 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. [14], perceptually aligned gradients 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. In order to make this claim stronger, as shown in OAAT [1], we analyze if calculating LPIPS distance using a robust model can indeed differentiate between the perturbations generated from standard verses adversarially robust models.

We train a robust vision transformer model using some adversarial training method and then perturb the pre-softmax scaling factors using the same training data. Now, this model is given to the attacker, who can craft any attack within the threat model. The model's weights and pre-softmax scaling factors are fixed and cannot be changed/perturbed by the attacker. The perturbation of pre-softmax scaling factors is thus not adjusted as per the attacked image.

3.1. Performance of AAS for different perturbation bounds

As shown in Algorithm-1 of main paper, AAS attack maximizes LPIPS distance using clean images rather than adversarial images. Thus, there is no relationship between the scaling factor's perturbation and input image's perturbation. Therefore, the generated perturbations of scaling factor are expected to generalize for different perturbation values of input images. As shown in Table 1, incorporating AAS attack with PGD-100 and GAMA-PGD indeed gives improved performance over these attacks for different perturbation radii. Further, the boost in the attack strength obtained by incorporating AAS attack increases with an increase in perturbation radius. This is because the gradient masking effect is expected to increase with an increase in perturbation radius. On CIFAR10, the gains by incorporating AAS attack with ϵ=16 is over 7.55%, while 3.09% with ϵ=8. For GAMA attack, we observe improved attack strength of 5.08% for ϵ =16, while 2.17% for ϵ =8. Similar observations are seen on CIFAR-100.

4. Why to choose LPIPS?

• **Motivation of using LPIPS distance.** In this section, we describe the motivation for using the LPIPS distance to perturb the pre-softmax scaling factors. It is well known that perturbations from a robust model are perceptually aligned, whereas from a standard model or model suffering from gradient masking are not perceptually aligned. Further, Ganz et al. [14] has recently demonstrated that a model giving perceptual aligned gradients is in fact robust. Therefore, if we can ensure that gradients become perceptually aligned after perturbing the pre-softmax scaling factors (motivated by [14]), the model can achieve improved robustness by overcoming gradient masking.

• **Our Observations.** To demonstrate this, we further present the comparison between CIFAR-10 perturbed images generated with a perturbation radius of 12/255 using 0,2,4,6,8,10 steps of AAS attack (Algorithm-1 of main paper) followed by 100 steps PGD attack in Figure 1 and Figure 2. As can be seen, the gradients generated without using AAS attack are more noisy and less perceptually aligned. Also, it is seen that maximizing LPIPS distance (Figure 1) generates more perceptually aligned gradients as compared to maximizing the cross-entropy loss (Figure 2). This observation is also supported by [20, 34], which shows that LPIPS is a good perceptual metric (this means two images having a lower LPIPS distance would be perceptually similar) using a human evaluation.

5. Details of the proposed Adaptive Attention Scaling (AAS) Attack

5.1. Baselines

The performance of different defences is evaluated on CIFAR10, CIFAR100 and ImageNet-100 [11] datasets, comprising of 10, 100 and 100 classes, respectively. The resolution of images in the CIFAR10 and CIFAR100 datasets is 32x32, while it is 256X256 in the ImageNet-100 dataset. For all the experiments, VIT-B16 architecture is used. CI-

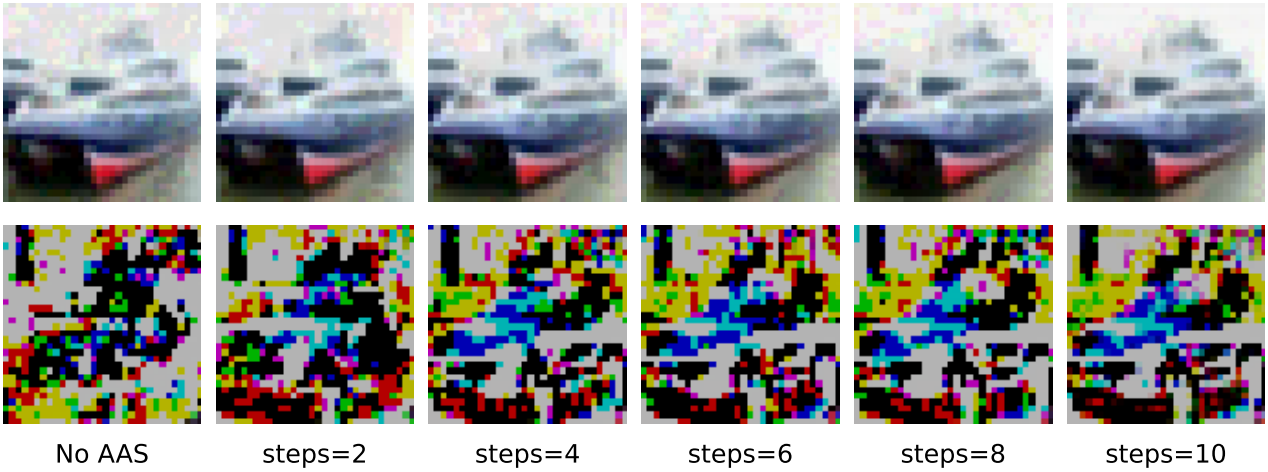

Figure 1. Effect of the number of AAS attack steps when the proposed **LPIPS distance** is maximized followed by 100 step PGD attack. On not using AAS (No AAS) attack, the perturbation generated by PGD-100 is more noisy.

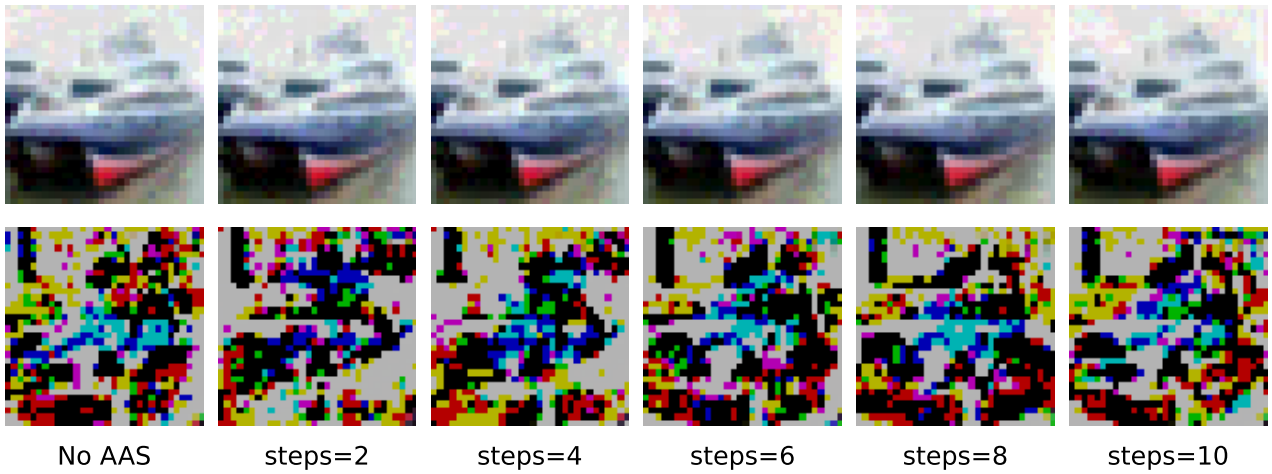

Figure 2. Effect of the number of AAS attack steps when **cross-entropy (CE) loss** is maximized instead of LPIPS distance, followed by 100-step PGD attack. Comparing the perturbation with Fig-1 where LPIPS distance is maximized, it can be seen that maximizing CE loss generates less perceptual/more noisy perturbations, especially for smaller number of attack steps.

FAR100 is a more challenging dataset as compared to CIFAR10 because it has one-tenth the number of images in each class along with a larger number of classes. We use the ImageNet-100 dataset to show that the proposed attack is generalizable to larger-resolution images as well. For all the attacks and training, we consider an ℓ_∞ threat model with $\epsilon = 8/255$. For all our attack and adversarial training experiments, we use the codebase of Mo et al. [23] [1].

5.2. Evaluation Details

We train the VIT-B16 model using PGD-AT [22] along with AWP [31] to evaluate the existing and proposed AAS attacks. VIT-B16 is trained for 110 epochs using a cosine learning rate schedule with a maximum learning rate of 0.1. SGD with a momentum of 0.9 is used for training the model. We utilize additional data generated from DDPM [16] for all the experiments on CIFAR10 and CIFAR100 datasets.

We evaluate PGD-AT + AWP trained VIT-B16 model against several attacks like PGD [22], CW [5], AutoAttack [8] and GAMA [26] attack. While AutoAttack [8] is the strongest attack, it is computationally expensive. Amongst the single-attack methods, which are relatively cheaper in terms of compute, GAMA attack is the strongest one. Amongst the multistep attacks except for AutoAttack, we use 100 iterations for generating the attack. The details of individual attacks is given by:

- **Fast Gradient Sign Method (FGSM)** [15]: FGSM is a single-step attack which maximizes cross-entropy loss (defined as L_{CE}). The attack objective of FGSM is described as follows:

$$\textbf{argmax}_{x'} L_{CE}(f_\theta(x'), y) \quad \textbf{s.t.} ||x' - x||_\infty < \epsilon. \quad (1)$$

[1] https://github.com/mo666666/When-Adversarial-Training-Meets-Vision-Transformers

- **Projected Gradient Descent (PGD)** [22]: PGD is a multi-step version of FGSM where the perturbation is initialized using random noise sampled from a uniform distribution.
- **Carlini and Wagner (in short CW)** [5]: CW attack maximizes max-margin loss (defined as L_{MM} instead of the standard cross-entropy loss used in PGD attack. The attack formulation of the CW attack considered in this work is shown below:

$$\textbf{argmax}_{x'} L_{MM}(f_\theta(x'), y) \quad \textbf{s.t.} ||x' - x||_\infty < \epsilon. \quad (2)$$

- **Guided Margin Aware Attack (GAMA)** [26]: GAMA attack proposed to aid the initial optimization path by maximizing the ℓ_2 norm between the output logits of the adversarial and the clean images along with maximizing the standard max-margin loss. The objective function of GAMA attack is shown as follows:

$$\textbf{argmax}_{x'} L_{MM}(f_\theta(x'), y) + \lambda ||f_\theta(x^{'}) - f_\theta(x)|| \quad \textbf{s.t.} ||x' - x||_\infty < \epsilon. \quad (3)$$

  Over the training iterations, the value of λ is decayed.
- **AutoAttack (AA)** [8]: AutoAttack is an ensemble of four attacks including three white box (Adaptive PGD with cross-entropy loss, Adaptive PGD with difference of logits ratio loss, Fast adaptive boundary attack [7]) and one black box (square attack [2]). The details of these attacks are described below:
  - **Adaptive PGD**: APGD is the same as the standard PGD attack but as opposed to PGD it adjusts the step size of the attack automatically. An untargeted APGD is used in the AutoAttack framework.
  - **Adaptive PGD with difference of logits ratio loss**: This attack uses the DLR loss instead of the standard cross-entropy loss. Further, the target attack is used in the AutoAttack framework. Therefore this attack is expensive because its frequency depends on the number of classes in the dataset.
  - **Fast Adaptive Boundary Attack (FAB)**: FAB attack aims to find the minimum perturbation required to change the true class predicted by the model. In the AutoAttack framework, a targeted FAB attack is used and it is the most expensive attack amongst all others in the AutoAttack framework.
  - **Square Attack**: Square attack is the only black box attack present in AutoAttack. It is a search-based attack, where randomly coloured squared and rectangles are added to the input image and then they are retained if there is an increase in loss value. Since the square attack is a gradient-free attack, it helps to circumvent and identify the models that are suffering from gradient masking issues.

6. Details of the proposed Adaptive Attention Scaling Adversarial Training (AAS-AT)

6.1. Baselines

We use the VIT-B16 model for all the experiments. VIT-B16 is trained for 110 epochs using a cosine learning rate schedule with a maximum learning rate of 0.1. SGD with a momentum of 0.9 is used for training the model. We utilize additional data generated from DDPM [16] for all the experiments on CIFAR10 and CIFAR100 datasets except for Debenedetti [10] we train for 300 epochs. We use a pretrained ImageNet-1K initialization and also use gradient clipping in all the experiments. We utilize simple Pad-Crop-Horizontal Flip as the augmentations for training. We compare the performance of the proposed AAS-AT with the following adversarial training methods:

- **PGD-AT** [22]: PGD-AT performs the standard ten-step PGD attack to generate the adversarial images and later minimized the cross-entropy loss on the generated adversarial images to train the model. The objective function of PGD-AT is shown below:

$$\textbf{argmax}_{x'} L_{CE}(f_\theta(x'), y) \quad \textbf{s.t.} ||x' - x||_\infty < \epsilon, \quad (4)$$

$$\textbf{min}\ L_{CE}(f_\theta(x'), y). \quad (5)$$

- **Trades** [33]: Trades maximizes the KL divergence (L_{KL}) loss between the adversarial and the clean image to generate the perturbations and later minimizes the combination cross-entropy loss on clean image and KL divergence loss between the clean and the generated adversarial image. The objective function of PGD-AT is shown below:

$$\textbf{argmax}_{x'} L_{KL}(f_\theta(x'), f_\theta(x)) \quad \textbf{s.t.}\ ||x' - x||_\infty < \epsilon, \quad (6)$$

$$\textbf{min}\ L_{CE}(f_\theta(x), y) + \lambda L_{KL}(f_\theta(x'), f_\theta(x)). \quad (7)$$

- **Trades-AWP** [31]: Trades-AWP proposes to first generate the attack by maximizing the KL divergence between the clean and the perturbed image and then perturb the weights of the model within an ℓ_2 norm perturbation bound (ρ). Later the Trades adversarial training is performed on the perturbed model. The objective function of AWP is shown below:

$$\theta^{'} = \theta + \delta,\ \delta = \textbf{argmax}_{\theta'}(L_{CE}(f_{\theta'}(x), y) + \lambda L_{KL}(f_{\theta'}(x'), f_{\theta'}(x))) - \theta, \quad \textbf{s.t.}\ ||\theta' - \theta|| < \rho, \quad (8)$$

$$\textbf{argmax}_{x'} L_{KL}(f_{\theta'}(x'), f_{\theta'}(x)) \quad \textbf{s.t.}\ ||x' - x||_\infty < \epsilon, \quad (9)$$

$$\textbf{min}\ L_{CE}(f_{\theta'}(x), y) + \lambda L_{KL}(f_{\theta'}(x'), f_{\theta'}(x)), \quad (10)$$

$$\theta = \theta' - \delta. \quad (11)$$

Algorithm 1 Adaptive Attention Scaling Adversarial Training (AAS-AT)

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 $\mathcal{D} = \{(x_i, y_i)\}$, Adversarial Threat model: ℓ_∞ bound of radius ε, coefficient of KL divergence term β, Cross-entropy loss ℓ_{CE}, number of epochs E, M training mini-batches of size n, Maximum Learning Rate LR_{max}, Frequency of AAS attack λ;
2: **for epoch** $= 1$ **to** E **do**
3: $\text{LR} = 0.5 \cdot \text{LR}_{max} \cdot (1 + \textbf{cosine}((\textbf{epoch} - 1)/E \cdot \pi))$;
4: **for iter** $= 1$ **to** M **do**
5: **if epoch**$\%\lambda == 0$ **then**
6: $\delta = \mathcal{N}(0, 1)$;
7: **for steps** $= 1$ **to** 10 **do**
8: **if epoch**$\%\lambda == 0$ **then**
9: $\delta = \delta + \nabla_S \textbf{LPIPS}(f_{\theta(S)}(x_i), f_{\theta(S')}(x_i))$;
10: $S^{'} = \textbf{Clamp}(S + \delta, 10^{-r}, 1)$; *% to prevent zero scaling factors,we used* $r = 7\%$
11: **else**
12: $\delta = 0.001 \cdot \mathcal{N}(0, 1)$;
13: $\delta = \delta + \varepsilon_{\text{asc}} \cdot \textbf{sign}\left(\nabla_\delta \textbf{KL}(f_\theta(x) || f_\theta(x + \delta))\right)$;
14: $\delta = \textbf{Clamp}\left(\delta, -\varepsilon_{\text{asc}}, \varepsilon_{\text{asc}}\right)$;
15: $\widetilde{x} = \textbf{Clamp}\left(x + \delta, 0, 1\right)$;
16: **if epoch**$\%\lambda == 0$ **then**
17: $S = S^{'}$;
18: **else**
19: $\mathcal{L}_{\text{TR}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{CE}}(f_\theta(x_i), y_i) + \beta \cdot \textbf{KL}(f_\theta(x_i) || f_\theta(\tilde{x}_i))$;
20: $\theta = \theta - \text{LR} \cdot \nabla_\theta(\mathcal{L}_{\text{TR}}(\theta))$;

6.2. AAS-AT Training Algorithm

In this section, we explain the proposed Adaptive Attention Scale Adversarial Training (AAS-AT). The detailed description of AAS-AT is shown in Algorithm 1. We train the VIT-B16 model for 110 epochs (L2) using a cosine learning rate schedule (L3). The proposed AAS attack is performed every λ epoch (L5). Firstly, the AAS attack is initialized using Gaussian noise (L6). Every λ epochs (L8) AAS attack is performed where the LPIPS distance is maximized to perturb the pre-softmax scaling factors (L9) and later clamped to lie between $(10^{-r}, 1)$ (L10). For the remaining epochs, we perform the standard Trades [33] adversarial training where the KL-Divergence between the clean and the adversarial images is maximized to perturb the images (L12-L15). Finally, if the task was to perturb the pre-softmax output scaling factor, then the old scaling factors are reinitialized using the new perturbed ones (L17). Otherwise, standard Trades adversarial training occurs on the perturbed image (L19-20) where the cross-entropy loss on the clean images and KL Divergence between the clean and the adversarial images is minimized.

6.3. Error Analysis

We perform multiple reruns of the proposed AAS-AT, and we observe small variations in the robust and standard accuracy across the reruns. We performed three reruns of Trades+AAS-AT and observed 87.23% as the mean accuracy and a standard deviation of 0.21% in clean accuracy. Against AutoAttack, we observed 57.61% as the mean adversarial accuracy and a standard deviation of 0.16%.

7. Ablation Experiments on proposed AAS attack

As shown in Figure-3 (a) of main paper, 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) of main paper, if the size of the model is increased by adding up more attention blocks, the drop in robust accuracy of the proposed AAS attack with respect to PGD-100 further increases. This demonstrates the effectiveness of overcoming gradient masking by using the proposed AAS attack. Finally, we present the effect of increasing the number of iterations of attack for PGD and AAS attacks in Figure-4 of main paper. As can be seen, AAS attack saturates earlier than PGD, and PGD is not able to close up the gap between the two attacks even on using 1000 iterations. As shown in Table 1, incorporating AAS attack with PGD-100 and GAMA-PGD indeed gives improved performance over these attacks for different perturbation radii. Further, the boost in the attack strength obtained by incorporating AAS attack increases with an increase in perturbation radius. This is because the gradient masking effect is expected to increase with an increase in perturbation radius. On CIFAR10, the gains by incorporating AAS attack with ϵ=16 is over 7.55%, while 3.09% with ϵ=8. For GAMA attack, we observe improved attack strength of 5.08% for ϵ=16, while 2.17% for ϵ=8. Similar observations are seen on CIFAR-100.

As observed, on incorporating AAS-AT with AdvCam-AT, we obtain gains of over 1.16%. This shows that including AAS-AT helps in improved robustness. Further, if the model is trained using AdvCam-AT, then there is a drop of 2.08% on including AAS attack in the evaluation along with the AdvCam-1000 attack. Thus, AdvCam-1000 is not able to generate an optimal attack.

Table 1. **Effect of increase in the perturbation radius.**

	CIFAR10				CIFAR100			
$\epsilon/255$	PGD-100	PGD-100 + AAS	GAMA	GAMA + AAS	PGD-100	PGD-100 + AAS	GAMA	GAMA + AAS
0.00	87.43	87.31	87.43	87.31	62.47	62.03	62.47	62.03
4.00	72.13	69.76	70.87	**69.16**	41.87	38.46	39.47	**38.03**
8.00	61.10	58.01	59.78	**57.61**	30.01	27.02	28.97	**26.08**
12.00	48.87	43.48	46.79	**42.86**	27.46	22.13	24.63	**21.46**
16.00	42.96	35.41	40.03	**34.95**	24.97	19.30	22.46	**18.76**

Table 2. **Results of Semantic Attack (AdvCam)** on ImageNet-100.

Method	Clean Accuracy	AdvCam-1000	AdvCam-1000 + AAS
AdvCam-AT	76.84	65.18	63.10
AdvCam-AT + Ours	76.96	64.3	**64.26**

8. Performance on Swin Transformer and LeViT

We train Swin Transformer [21] Swin-T) and LeViT [17] without distillation head using PGD-AT and AWP and present the evaluation on different attacks with and without incorporating our AAS attack in Table 3. On incorporating the AAS attack, an improvement in attack strength across all the architectures on both CIFAR10 and CIFAR100 datasets is observed.

9. Adversarial robustness of CNNs verses VITs

We would like to highlight Table-7 (b,d) of Debenedetti et al. [10]. We present an analysis by considering WideResNet-28-10 and ReNet50 as CNN models and XCIT-S12 for VIT (as used in [10]) in Table 4. Note that robust and standard accuracy of CNNs are taken from Table-7 (b,d) of [10]. As can be seen, XCIT-S12+Ours outperforms around 1.4 times the parameter count WideResNet-28-10 model by over 5.04% on CIFAR100 and 2.5% on CIFAR10. This shows that VITs exhibits superior adversarial robustness than CNNs.

• As observed, the computational complexity (measured in FLOPS) on incorporating AAS attack increases by about 0.19% and attack time increases by about 5.11%. We also present the training time comparison of PGD-AT, Trades, Ensemble of PGD-AT+Trades, PGD-AT+Ours and Trades+Ours and Ensemble of PGD-AT+Ours and Trades+Ours for VIT-B16 model on CIFAR10 dataset in Table 7.

10. Ablation Experiment on our Defense (AAS-AT)

As shown in Figure-5 (a,b) of main paper, 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 ϵ and performing PGD-20 attack during evaluation is shown in Figure-5 (c) of main paper. 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 of main paper) as compared to the baseline PGD-AT. Since PGD-AT suffers from gradient masking, thus 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) of main paper, 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 ϵ close to $60/255$. Thus, the proposed AAS attack is able to overcome the gradient masking effect in PGD-AT model.

11. Performance on Patch Attacks

We observe improved performance on combining AAS attack with existing Patch attacks. We incorporate our AAS attack with Patch-Fool [13] and Gu et al. [18]. We perform this analysis on the ImageNet-100 dataset with a normally trained VIT-B16 model giving 83.64% clean accuracy in Table-8. We consider that the attacker can perturb a single patch of 8×8 dimensions in the image.

Due to the simplicity of the proposed AAS attack, incorporating it along with existing patch attacks helps in improving their attack strength. On incorporating AAS with Sparse Patch-Fool [13], we observe improved attack strength by over 2.24% and with Gu et al. [18], gains of over 4.46% are observed.

12. Performance on Semantic Attack (AdvCam)

We utilize the AdvCam attack proposed in [12] to generate adversarial samples for training. For training VIT-B16 model on ImageNet-100, we use a 10-step attack with a coefficient of the adversarial loss linearly varying from 1000 to 10000 over these 10 steps. For all these experiments, we specify the size of the attack region as 40×40 and use a random image

Table 3. **Effectiveness of AAS attack on Swin Transformer and LeViT.**

Data	Model	Attack	Clean Acc. w/o AAS	Robust Acc. w/o AAS	Clean Acc.+ AAS	Robust Acc. + AAS
CIFAR10	VIT-B16	PGD-100	87.43	61.10	87.31	**58.01**
		GAMA	87.43	59.78	87.31	**57.61**
	Swin-T	PGD-100	76.03	50.13	76.16	**48.48**
		GAMA	76.03	49.03	76.16	**47.32**
	LeVIT-256	PGD-100	82.42	56.13	82.16	**53.87**
		GAMA	82.42	55.34	82.16	**53.39**
CIFAR100	VIT-B16	PGD-100	62.47	30.01	62.03	**27.02**
		GAMA	62.47	28.97	62.03	**26.07**
	Swin-T	PGD-100	56.79	26.49	56.81	**24.16**
		GAMA	56.79	25.79	56.81	**23.78**
	LeVIT-256	PGD-100	60.10	27.06	60.03	**25.79**
		GAMA	60.10	26.74	60.03	**25.46**

Table 4. **Adversarial robustness of CNNs and VITs.**

		CIFAR10		CIFAR100	
Model	Parameter Count	Clean Accuracy	AA accuracy	Clean Accuracy	AA accuracy
WideResNet-28-10 [10]	36.5M	87.11	54.92	59.23	28.42
ResNet-50 [10]	25M	84.8	41.56	61.28	22.01
XCIT-S12 [10]	26M	90.06	56.14	**67.34**	32.17
XCIT-S12 [10] + Ours	26M	**90.78**	**57.42**	67.12	**33.46**

Table 5. **Number of samples required for AAS attack.**

Data	Number of samples	Clean Accuracy	GAMA Accuracy
CIFAR10	0	87.43	59.78
	500	87.26	58.31
	1000	87.18	**57.73**
	1500	87.24	57.66
	2000	87.19	57.67
	50000 (reported)	87.31	57.61
CIFAR100	0	62.47	28.97
	500	61.84	28.03
	1000	62.12	27.16
	1500	62.01	26.42
	2000	61.94	**26.13**
	50000 (reported)	62.03	26.08

from the same class for the target style. For evaluation, we use the same AdvCam attack but with 1000 steps. We present our observations in Table 2.

13. Performance on large datasets: ImageNet-100, ImageNet-200

To analyze the scalability of our AAS-AT, we train VIT-B16 using PGD-AT and Trades on 100 class - randomly chosen subset of the ImageNet dataset. We also trained separate models by incorporating AAS-AT with PGD-AT and Trades. Due to computational limitations, we are currently not able to perform adversarial training on the ImageNet dataset. We will certainly consider extending our method to ImageNet for the final version.

We have compared the performance of the proposed AAS attack on ImageNet-100 in Table-4 of main paper, where we observe around 3.5% improvements over PGD-100 and 2.2% improved results over the GAMA attack. This highlights that we can overcome the floating point underflow errors and the proposed AAS attack gives consistent gains over the existing attacks. The results against Auto-Attack on the 5000 images validation set are presented in Table-5 of main paper. On Auto-Attack, we observe gains of over 1.54% on incorporating AAS-AT with PGD-AT and 1.58% with Trades.

14. Discussion on Computational Efficiency

We would like to highlight that in all our experiments, we first train a robust vision transformer model using some adversarial training method and then perturb the pre-softmax scaling factors using the same training data. This model is then given to the attacker, who can craft any attack. Thus, the weights and pre-softmax scaling factors are fixed and cannot be changed by the attacker. Therefore, AAS attack can be considered like an additional training epoch (on clean images). If the scaling factors are perturbed right after training itself (when the training data is available), then this limitation can be overcome. But even if this is not done, we observe that securing the entire training set is not important during inference; rather, less than 5% of the training set on CIFAR10 and CIFAR100 can give a reasonable estimate of the scaling factors.

• We present an analysis of the amount of training data

Table 6. **Computational time in Flops on CIFAR10.**

Attack	# Forward passes	# Backward passes	Flops (attack)	Attack time (sec)	Clean Accuracy	Robust Accuracy
PGD-100	100	100	2.2380E + 13	1153	87.43	61.1
PGD-100 + AAS	100 (PGD) + 1 (AAS)	100 (PGD) + 1 (AAS)	2.2454E+13	1212	87.31	58.01
GAMA	200	100	2.9840E+13	1411	87.43	59.78
GAMA + AAS	200 (GAMA) + 1 (AAS)	100 (GAMA) + 1 (AAS)	2.9914E+13	1470	87.31	57.61

Table 7. **Training Time Comparison on CIFAR10.**

Method	Training time (sec)	Clean Accuracy	AA Accuracy
PGD-AT	32315	86.14	53.14
PGD-AT + Ours	32613	85.32	56.61
Trades	33213	86.31	54.03
Trades + Ours	33546	87.46	57.34
Ensemble (PGD-AT, Trades)	65528	88.03	56.36
Ensemble (PGD-AT + Ours, Trades + Ours)	66159	88.06	58.01

Table 8. **Performance of AAS on Patch Attack** on ImageNet-100.

Method	**Robust Accuracy**
Sparse Patch-Fool [13]	71.36
Sparse Patch-Fool [13] + AAS	**69.12**
Gu et al. [18]	74.98
Gu et al. [18] + AAS	**70.52**

needed to optimize the pre-softmax scaling factors on CIFAR10 and CIFAR100 datasets in the Table 5. To test the robustness of the model, we use the GAMA attack. As can be seen, on CIFAR100, using 2000 training images (around 4% of total training images) is enough to get a close estimate of the pre-softmax scaling factors. On CIFAR10, around 1000 training images can give a close estimate.

• We present the cost of computational time in terms of flops for PGD-100, GAMA, PGD-100 + AAS and GAMA + AAS attacks on VIT-B16 in Table 6 on CIFAR10 dataset for both clean and robust accuracy. We consider same number of forward and backward passes. We have also highlighted the attack time in seconds.

• On incorporating AAS-AT with PGD-AT in Table 7, we observe an increase of about 0.92% in the training time and about 1% with Trades. We also include the results of an ensemble of PGD-AT and Trades and an ensemble of PGD-AT+Ours and Trades+Ours. We take the average of the output softmax distribution of two models and report the highest confidence class as the predicted class. In the case of an ensemble, we observe an increase in training time by around 0.96%.

15. Limitations

In this work, though we propose a gradient-based optimization to get the scaling factors automatically, it is bound to give an approximate value. It is difficult to analyze how close this is with respect to the optimal scaling factors one can find by trying out all possible combinations of scaling factors. Further, on perturbing the scaling factors, we observe a drop in the clean accuracy. This is bound to happen since the model is not finetuned on these perturbed scaling factors. Though there is a drop in the clean accuracy, it is not significant.

16. Social Impact

By highlighting the reason for gradient masking in VITs, this work aims to improve the robustness of VITs and prevent the development of future defences, which might give a false sense of security because existing attacks are weak on VITs. We hope that this work will help in the development of more robust defences on VITs in the future.

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. 2

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, pages 1–34, 2020. 1, 4

[3] 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

[4] 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

[5] 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, 3, 4

[6] 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

[7] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, pages 1–24, 2020. 1, 4

[8] 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, 3, 4

[9] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, pages 1–29, 2021. 2

[10] Edoardo Debenedetti. A light recipe to train robust vision transformers. *arXiv preprint arXiv:2209.07399*, pages 1–29, 2022. 2, 4, 6, 7

[11] 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. 2

[12] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, and Yun Yang. Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 997–1005, 2020. 6

[13] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations? In *International Conference on Learning Representations (ICLR)*, pages 1–18, 2022. 6, 8

[14] Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply adversarial robustness? *arXiv preprint arXiv:2207.11378*, pages 1–21, 2022. 2

[15] Goodfellow. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, pages 1–11, 2015. 1, 3

[16] 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. 3, 4

[17] 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. 6

[18] Jindong Gu, Volker Tresp, and Yao Qin. Are Vision Transformers Robust to Patch Perturbations? In *European Conference on Computer Vision (ECCV)*, page 404–421, 2022. 6, 8

[19] 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

[20] 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. 2

[21] 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. 6

[22] 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, 3, 4

[23] 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. 1, 3

[24] 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

[25] Sayak Paul and Pin-Yu Chen. Vision Transformers are Robust Learners. In *AAAI Conference on Artificial Intelligence*, page 2071–2081, 2022. 1

[26] 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, 3, 4

[27] 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

[28] 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

[29] 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. 1

[30] 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

[31] 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. 1, 3, 4

[32] 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

[33] 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. 1, 4, 5

[34] 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. 1, 2