Sharpness-Aware Minimization Alone can Improve Adversarial Robustness

Zeming Wei [*1] **Jingyu Zhu** [*1] **Yihao Zhang** [*1]

Abstract

Sharpness-Aware Minimization (SAM) is an effective method for improving generalization ability by regularizing loss sharpness. In this paper, we explore SAM in the context of adversarial robustness. We find that using only SAM can achieve superior adversarial robustness without sacrificing clean accuracy compared to standard training, which is an unexpected benefit. We also discuss the relation between SAM and adversarial training (AT), a popular method for improving the adversarial robustness of DNNs. In particular, we show that SAM and AT differ in terms of perturbation strength, leading to different accuracy and robustness trade-offs. We provide theoretical evidence for these claims in a simplified model. Finally, while AT suffers from decreased clean accuracy and computational overhead, we suggest that SAM can be regarded as a lightweight substitute for AT under certain requirements. Code is available at https://github.com/weizeming/SAM_AT.

1. Introduction

Sharpness-Aware Minimization (SAM) (Foret et al., 2020) is a novel training framework that improves model generalization by simultaneously minimizing loss value and loss sharpness. The objective of SAM is to minimize the *sharpness* around the parameters, which can be formulated as

$$\max_{\|\epsilon\| \leq \rho} L(w+\epsilon) + \lambda \|w\|_2^2, \tag{1}$$

where L is the loss function, w is the parameters of the model, $\|w\|_2^2$ is the regularization term and ρ controls the magnitude of weight perturbation. Intuitively, a larger ρ leads to stronger weight perturbation and pushes the model to find a flatter loss landscape. So far, SAM has become a powerful tool for enhancing the natural accuracy performance of machine learning models.

In this paper, we aim to explore SAM through the lens of adversarial robustness. Specifically, we study the robustness of SAM to defend against adversarial examples, which are natural examples with small perturbations that mislead the model into producing incorrect predictions (Szegedy et al., 2013; Goodfellow et al., 2014). The discovery of adversarial examples has raised serious concerns about the safety of critical domain applications (Ma et al., 2020), and has attracted a lot of research attention in terms of defending against them. Currently, Adversarial Training (AT) (Madry et al., 2017) has been demonstrated to be the most effective approach (Athalye et al., 2018) in improving the adversarial robustness of Deep Neural Networks (DNNs) among the various methods of defense. However, despite the success in improving adversarial robustness, there are still several defects remaining in adversarial training, such as decrease in natural accuracy (Tsipras et al., 2018), computational overhead (Shafahi et al., 2019) , class-wise fairness (Xu et al., 2021; Wei et al., 2023a) and the absence of formal guarantees (Wang et al., 2021; Zhang et al., 2023).

Surprisingly, we find that models trained with SAM exhibit significantly higher adversarial robustness than those trained using standard methods, which is an unexpected benefit. Also, when comparing SAM to AT, SAM has the advantage of lower computational cost and no decrease in natural accuracy. Based on the discussion above, we raise two research questions (**RQs**) in this paper:

- **RQ1**: Why does SAM improve adversarial robustness compared to standard training?
- **RQ2**: Can SAM be used as a lightweight substitution for adversarial training?

To answer the two questions above, we first provide a comprehensive understanding of SAM in terms of adversarial robustness. Specifically, we present the intrinsic relation between SAM and AT that they both apply adversarial data augmentations to eliminate non-robust features (Tsipras et al., 2018)) from natural examples during the training phase. As a result, both SAM and AT can effectively enhance the robustness of trained models, re-

[*]Equal contribution [1]Peking University. Correspondence to: Zeming Wei <weizeming@stu.pku.edu.cn>.

2^{nd} *AdvML Frontiers workshop at* 40^{th} *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

sulting in improved robust generalization ability. However, we also note that there are still several differences between SAM and AT. For instance, SAM adds adversarial perturbations *implicitly*, while AT applies perturbations *explicitly*. Additionally, the perturbation (attack) strength during training of SAM and AT differs, leading to different results in terms of natural and robust accuracy trade-offs.

Further, we verify the proposed empirical understanding with theoretical evidence in a simplified data model. Following the data distribution based on robust and non-robust features decomposition (Tsipras et al., 2018), we show that both SAM and AT can improve the robustness of the trained models by biasing more weight on robust features. In addition, we also show that SAM requires a larger perturbation budget to achieve comparable robustness to AT, which verifies our hypothesis that the perturbation strength of SAM is lower than AT.

Finally, we conduct experiments on benchmark datasets to verify our understanding. We find that models trained with SAM indeed outperform standard-trained models significantly in terms of adversarial robustness and also exhibit better natural accuracy. To sum up, our empirical and theoretical understanding can answer **RQ1**.

It is worth noting that, there still remains a large gap of robustness between SAM and AT. However, the natural accuracy of AT is consistently lower than standard training, not to mention SAM. Meanwhile, SAM also outperforms AT in terms of computational cost. Therefore, we finally answer **RQ2** with the conclusion that SAM can be considered a lightweight substitute for AT in improving adversarial robustness, under the following requirements: (1) no loss of natural accuracy and (2) no significant increase in computational cost.

To summarize, our main contributions in this paper are:

1. We point out that using SAM alone can notably enhance adversarial robustness without sacrificing clean accuracy compared to standard training, which is an unexpected benefit.

2. We provide both empirical and theoretical explanations to clarify how SAM can enhance adversarial robustness. In particular, we discuss the relation between SAM and AT and demonstrate that they improve adversarial robustness by eliminating non-robust features. However, they differ in perturbation strengths, which leads to different trade-offs between natural and robust accuracy.

3. We conducted experiments on benchmark datasets to verify our proposed insight. We also suggest that SAM can be considered a lightweight substitute for AT under certain requirements.

2. Background and related work

2.1. Sharpness awareness minimization (SAM)

In order to deal with the bad generalization problem in traditional machine learning algorithms, (Hochreiter & Schmidhuber, 1994; 1997) respectively attempt to search for flat minima and penalize sharpness in the loss landscape, which obtains good results in generalization (Keskar et al., 2016; Neyshabur et al., 2017; Dziugaite & Roy, 2017). Inspired by this, a series of works focus on using the concept of *flatness* or *sharpness* in loss landscape to ensure better generalization, *e.g.* Entropy-SGD (Chaudhari et al., 2019) and Stochastic Weight Averaging (SWA) (Izmailov et al., 2018). Sharpness-Aware minimization (SAM) (Foret et al., 2020) also falls into this category, which simultaneously minimizes loss value and loss sharpness as described in (1).

Theoretically, the good generalization ability of SAM is guaranteed by the fact that of the high probability, the following inequality holds:

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + h\left(\|\boldsymbol{w}\|_2^2/\rho^2\right), \quad (2)$$

where set $\mathcal{S}$ is generated from distribution $\mathscr{D}$, $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function.

There are also many applications of SAM in other fields of research like language models (Bahri et al., 2021) and fluid dynamics (Jetly et al., 2022), showing the scalability of SAM to various domains. In addition, many improvements of the algorithm SAM spring up, like Adaptive SAM (ASAM) (Kwon et al., 2021), Efficient SAM (ESAM) (Du et al., 2021), LookSAM (Liu et al., 2022), Sparse SAM (SSAM) (Mi et al., 2022), Fisher SAM (Kim et al., 2022) and FSAM (Zhong et al., 2022), which add some modifications on SAM and further improve the generalization ability of the model.

2.2. Adversarial robustness

The adversarial robustness and adversarial training has become popular research topic since the discovery of adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014), which uncovers that DNNs can be easily fooled to make wrong decisions by adversarial examples that are crafted by adding small perturbations to normal examples. The malicious adversaries can conduct adversarial attacks (Chen et al., 2023b; Wei et al., 2023b) by crafting adversarial examples, which cause serious safety concerns regarding the deployment of DNNs. So far, numerous defense approaches have been proposed (Papernot et al., 2016; Xie et al., 2019; Bai et al., 2019; Mo et al., 2022; Chen et al., 2023a), among which adversarial training

(AT) (Madry et al., 2017; Wang et al., 2019) has been considered as the most promising defending method against adversarial attacks, which can be formulated as

$$\min_{w} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\delta\|\leq\epsilon} L(w; x+\delta, y), \tag{3}$$

where $\mathcal{D}$ is the data distribution, ϵ is the margin of perturbation, w is the parameters of the model and L is the loss function (*e.g.* the cross-entropy loss). For the inner maximization process, Projected Gradient Descent (PGD) attack is commonly used to generate the adversarial example:

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}(x^t + \alpha \cdot \text{sign}(\nabla_x \ell(\theta; x^t, y))), \tag{4}$$

where Π projects the adversarial example onto the perturbation bound $\mathcal{B}(x,\epsilon) = \{x' : \|x'-x\|_p \leq \epsilon\}$ and α represents the step size of gradient ascent.

Though improves adversarial robustness effectively, adversarial training has exposed several defects such as computational overhead (Shafahi et al., 2019), class-wise fairness (Xu et al., 2021; Wei et al., 2023a), among which the decreased natural accuracy (Tsipras et al., 2018; Wang & Wang, 2023) has become the major concern. It is proved that there exists an intrinsic trade-off between robustness and natural accuracy (Tsipras et al., 2018), which can explain why AT reduces standard accuracy significantly.

In the context of adversarial robustness, there are also several works that attempt to introduce a flat loss landscape in adversarial training (Wu et al., 2020; Yu et al., 2022a;b). The most representative one is Adversarial Weight Perturbation (AWP) (Wu et al., 2020), which simultaneously adds perturbation on examples and feature space to apply sharpness-aware minimization on adversarial training. However, AWP also suffers from a decrease in natural accuracy. Also, the reason why a flat loss landscape can lead to better robustness has not been well explained.

To the best of our knowledge, we are the first to uncover the intrinsic relation between SAM and AT, and we reveal that SAM can improve adversarial robustness by implicitly biasing more weight on robust features.

3. Empirical understanding

In this section, we introduce our proposed empirical understanding on the relation between SAM and AT, which can explain how SAM improves adversarial robustness.

Recall that the goal of SAM is to minimize the generalization error and loss sharpness simultaneously. The sharpness term can be described as $\max_{||\boldsymbol{\epsilon}||<\rho} [L_S(\boldsymbol{w}+\boldsymbol{\epsilon}) - L_S(\boldsymbol{w})]$, and the loss term is $L_S(\boldsymbol{w})$. By combining the two terms, we get the objective of SAM is

$$\min_{\boldsymbol{w}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{||\boldsymbol{\epsilon}||<\rho} L_S(\boldsymbol{w}+\boldsymbol{\epsilon}; x, y). \tag{5}$$

Also, recall that the objective of AT is

$$\min_{\boldsymbol{w}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\delta\|\leq\epsilon} L_s(\boldsymbol{w}; x+\delta, y). \tag{6}$$

To illustrate their relation, we first emphasize that both techniques involve adding **perturbation** as a form of data augmentation for eliminating non-robust features (Ilyas et al., 2019). However, AT **explicitly** adds these perturbations to input examples, while SAM focuses on perturbing the parameters, which can be considered an **implicit** kind of data augmentation on the feature space. Therefore, both techniques involve perturbation on features, but in different spaces.

To be more specific and formal, we can derive our understanding with a middle linear layer in a model, which extracts feature z from input x: $z = Wx$. In AT, we add perturbations directly to the input space, resulting in $x \leftarrow x + \delta$. However, in SAM, the perturbation is not directly applied to the input space, but to the parameter space as $W \leftarrow W + \delta$. This leads to $Wx + W\delta$ for input perturbation and $Wx + \delta x$ for parameter perturbation. Both perturbations can be seen as a form of data augmentation, with the former being more explicit and the latter being more implicit.

In addition, we discuss the attack (perturbation) strength of AT and SAM. For SAM, the perturbation is relatively more moderate, as its perturbations are injected in the feature space. However, this small perturbation is still helpful in improving robustness, since it can eliminate the non-robust features implicitly. On the other hand, in order to achieve the best robustness by destroying the non-robust features, AT applies larger and more straightforward perturbations to the input space, leading to better robustness but a loss in natural accuracy.

Therefore, the difference and relation between SAM and AT can be considered as a trade-off between robustness and accuracy (Zhang et al., 2019). In summary, SAM applies small perturbations implicitly to the feature space to maintain good natural accuracy performance, while AT utilizes direct data augmentation magnitudes, which may result in a severe loss in natural accuracy. We provide more theoretical evidence for these claims in the next section.

4. Theoretical analysis

In this section, we provide a theoretical analysis of SAM and the relation between SAM and AT. Following the robust/non-robust feature decomposition (Tsipras et al.,

2018), we introduce a simple binary classification model, in which we show the implicit essential similarity and difference of SAM and AT. We first present the data distribution and hypothesis space, then present how SAM and AT work in this model respectively, and finally discuss their relations.

4.1. A binary classification model

Consider a binary classification task that the input-label pair $(\boldsymbol{x}, y)$ is sampled from $\boldsymbol{x} \in \{-1,+1\} \times \mathbb{R}^d$ and $y \in \{-1,+1\}$, and the distribution $\mathcal{D}$ is defined as follows.

$$\begin{aligned} y &\overset{\text{u.a.r}}{\sim} \{-1,+1\}, x_1 = \begin{cases} +y, & \text{w.p. } p, \\ -y, & \text{w.p. } 1-p, \end{cases} \\ x_2, \ldots, x_{d+1} &\overset{i.i.d}{\sim} \mathcal{N}(\eta y, 1), \end{aligned} \tag{7}$$

where $p \in (0.5, 1)$ is the accuracy of feature x_1, constant $\eta > 0$ is a small positive number. In this model, x_1 is called the *robust feature*, since any small perturbation can not change its sign. However, the robust feature is not perfect since $p < 1$. Correspondingly, the features $x_2, \cdots, x_{d+1}$ are useful for identifying y due to the consistency of sign, hence they can help classification in terms of natural accuracy. However, they can be easily perturbed to the contrary side (change their sign) since η is a small positive, which makes them called non-robust features (Ilyas et al., 2019).

Now consider a linear classifier model which predicts the label of a data point by computing $f_{\boldsymbol{w}}(\boldsymbol{x}) = \text{sgn}(\boldsymbol{w} \cdot \boldsymbol{x})$, and optimize the parameters $w_1, w_2, \cdots, w_n$ to maximize $\mathbb{E}_{x.y\sim\mathcal{D}} \mathbf{1}(f_{\boldsymbol{w}}(x) = y)$. In this model, given the equivalency of $x_i (i = 2, \cdots, n)$, we can set $w_2 = \cdots = w_n = 1$ by normalization without loss of generality. Therefore, the numerical value w_1 has a strong correlation with the robustness of the model. Specifically, larger w_1 indicates that the model bias more weight on the robust feature x_1 and less weight on the non-robust features $x_2, \cdots, x_{d+1}$, leading to better robustness.

In the following, we discuss the trained model under standard training (ST), AT, and SAM respectively. To make our description clear, we denote the loss function $\mathcal{L}(\boldsymbol{x}, y, w)$ as $1 - \Pr(f_w(\boldsymbol{x}) = y)$ and for a given $\epsilon > 0$, we define the loss function of SAM $\mathcal{L}^{SAM}$ as $\max_{|\delta|\leq\epsilon} \mathcal{L}(\boldsymbol{x}, y, w + \delta)$.

4.2. Standard training (ST)

We first show that there exists an optimal parameter w_1^* under standard training in this model by the following theorem:

Theorem 4.1 (Standard training)**.** *In the model above, under standard training, the optimal parameter value is*

$$w_1^* = \frac{\ln p - \ln(1-p)}{2\eta}. \tag{8}$$

Therefore, w_1^* can be regarded as the parameter w_1 returned by standard training with this model.

4.3. Adversarial training (AT)

Now let's consider when AT is applied. In this case, the loss function is no longer the standard one but the expected adversarial loss

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{||\boldsymbol{\delta}||_\infty \leq \epsilon} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, y; w) \right]. \tag{9}$$

Similar to standard training, there also exists an optimal parameter w_1^{AT} returned by adversarial training, which can be stated in the following theorem:

Theorem 4.2 (Adversarial training)**.** *In the classification problem above, under adversarial training with perturbation bound $\epsilon < \eta$, the adversarial optimal parameter value*

$$w_1^{AT} = \frac{\ln p - \ln(1-p)}{2(\eta - \epsilon)}. \tag{10}$$

We can see that w_1 has been multiplied by $\frac{\eta}{\eta-\epsilon}$, which has increased the dependence on the robust feature x_1 of the classifier. This shows the adversarially trained model pays more attention to robustness compared to the standard-trained one, which improves adversarial robustness.

4.4. Sharpness-Aware Minimization (SAM)

Now we consider the situation of SAM. Recall that the optimizing objective of SAM is

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{|\delta|\leq\epsilon} \mathcal{L}(x, y; w + \delta) \right]. \tag{11}$$

We first explain why SAM could improve the adversarial robustness by proving that the parameter w_1 trained with SAM is also larger than w_1^*, which is stated as follows:

Theorem 4.3 (Sharpness-aware minimization)**.** *In the classification problem above, the best parameter for SAM training w_1^{SAM} satisfies that*

$$w_1^{SAM} > w_1^*. \tag{12}$$

From theorem 4.2 and 4.3 we can see that both w_1^{AT} and w_1^{SAM} are greater than w_1^*, which indicates both SAM and AT improve robustness of the trained model. However, the qualitative relation is not sufficient to quantify how much robustness SAM achieves compared to adversarial training, and we attempt to step further by quantitatively estimating the w_1^{SAM} in the following theorem:

Theorem 4.4. *In the classification problem above, denote the best parameter for SAM training w_1^{SAM}. Suppose that ϵ is small, we have $w_1^{SAM} \approx w_1^* + \frac{2}{3} w_1^* \epsilon^2$.*

Table 1. Natural and Robust Accuracy evaluation on **CIFAR-100** dataset.

Method	Natural Accuracy	ℓ_∞-Robust Accuracy		ℓ_2-Robust Accuracy	
		$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
ST	76.9	13.6	1.7	44.5	21.2
SAM ($\rho = 0.1$)	78.0	19.6	3.0	51.5	27.2
SAM ($\rho = 0.2$)	78.5	23.1	4.2	54.2	31.3
SAM ($\rho = 0.4$)	**78.7**	**28.3**	**6.5**	**57.0**	**36.2**
AT (ℓ_∞-$\epsilon = 1/255$)	**73.1**	**60.4**	46.6	**67.4**	**61.5**
AT (ℓ_∞-$\epsilon = 2/255$)	70.1	60.3	50.6	65.7	60.6
AT (ℓ_∞-$\epsilon = 4/255$)	66.2	59.3	**52.0**	62.8	58.9
AT (ℓ_∞-$\epsilon = 8/255$)	60.4	55.1	50.4	57.0	54.3
AT (ℓ_2-$\epsilon = 16/255$)	**74.8**	52.8	31.4	66.3	57.7
AT (ℓ_2-$\epsilon = 32/255$)	73.2	57.4	40.6	**67.1**	**61.0**
AT (ℓ_2-$\epsilon = 64/255$)	70.7	58.1	45.9	66.1	60.7
AT (ℓ_2-$\epsilon = 128/255$)	67.4	**58.2**	**48.7**	63.9	60.4

4.5. Relation between SAM and AT

We further discuss the distinct attack (perturbation) strength between AT and SAM. Recall that in our empirical understanding in Section 3, the perturbation of SAM is more moderate than AT, which can be interpreted as SAM focusing on natural accuracy more and robustness less in the robustness-accuracy trade-off. Therefore, to reach the same robustness level (which is measured by the dependence on feature x_1, *i.e.* the magnitude of w_1), SAM requires a much larger perturbation range, while for AT, less perturbation over x is enough. Theoretically, the following theorem verifies our statement:

Theorem 4.5. *Denote the perturbation range ϵ of AT and SAM as ϵ_{AT} and ϵ_{SAM}, respectively. Then, when both methods return the same parameter w_1, we have the following relation between ϵ_{AT} and ϵ_{SAM}:*

$$2 + \frac{3}{\epsilon_{SAM}^2} \approx \frac{2\eta}{\epsilon_{AT}} \tag{13}$$

From theorem 4.5, we can identify the different perturbation strengths of AT and SAM. It can be easily derive from Theorem 4.5 that ϵ_{SAM} is larger than ϵ_{AT} when (13) holds, since we assume η is a small positive, ϵ is small in theorem 4.4 and $\epsilon_{AT} < \eta$ in theorem 4.2. Therefore, to gain the same weight w_1 on robust features x_1, ϵ_{AT} only need to be chosen much smaller than ϵ_{SAM}. On the other hand, under the same perturbation bound $\epsilon_{AT} = \epsilon_{SAM}$, the model trained under AT has larger parameter w_1 than SAM, hence it focuses on more robustness yet decreases more natural accuracy.

All proofs can be found in Appendix A. To sum up, we can conclude that AT utilizes explicit and strong perturbations for denoising non-robust features, while SAM leverages implicit and moderate perturbations. This is consistent with our empirical understanding in Section 3 and we also verify these claims with experiments in the following section.

5. Experiment

In this section, we present our experimental results to verify our proposed understanding.

5.1. Experiment set-up

To demonstrate the effectiveness of SAM in improving adversarial robustness, we compare models trained with the standard SGD optimizer to those trained with SAM. We also discuss adversarial training. However, we consider the robustness obtained by AT as an upper bound rather than a baseline for SAM.

In our experiment, we train the PreActResNet-18 (PRN-18) (He et al., 2016) model on the CIFAR-10 and CIAR-100 datasets (Krizhevsky et al., 2009) with Cross-Entropy loss for 100 epochs. The learning rate is initialized as 0.1 and is divided by 10 at the 75th and 90th epochs, respectively. For the optimizer, the weight decay is set to `5e-4`, and the momentum is set to 0.9.

For SAM, we select the perturbation hyper-parameter ρ from the range $\{0.1, 0.2, 0.4\}$. And for AT, we consider both ℓ_2 and ℓ_∞ robustness and train 4 models with different perturbation bounds for the two kinds of norms, respectively.

As for robustness evaluation, we consider robustness under ℓ_∞-norm perturbation bounds $\epsilon \in \{1/255, 2/255\}$ and ℓ_2-norm perturbation bounds $\epsilon \in \{16/255, 32/255\}$. The robustness is evaluated under a 10-step PGD attack.

Table 2. Natural and Robust Accuracy evaluation on **CIFAR-10** dataset.

Method	Natural Accuracy	ℓ_∞-Robust Accuracy		ℓ_2-Robust Accuracy	
		$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
ST	94.6	39.6	8.9	76.1	51.7
SAM ($\rho = 0.1$)	**95.6**	45.1	9.4	81.0	56.3
SAM ($\rho = 0.2$)	95.5	48.9	10.2	82.9	58.8
SAM ($\rho = 0.4$)	94.7	**56.1**	**15.6**	**84.0**	**64.4**
AT (ℓ_∞-$\epsilon = 1/255$)	**93.7**	86.4	75.5	**90.5**	86.4
AT (ℓ_∞-$\epsilon = 2/255$)	92.8	**87.4**	79.6	90.3	**86.9**
AT (ℓ_∞-$\epsilon = 4/255$)	90.9	86.4	**81.3**	88.3	85.7
AT (ℓ_∞-$\epsilon = 8/255$)	84.2	81.5	78.4	82.5	80.8
AT (ℓ_2-$\epsilon = 16/255$)	**94.5**	82.2	61.7	90.3	84.5
AT (ℓ_2-$\epsilon = 32/255$)	93.7	84.9	70.9	**91.0**	86.7
AT (ℓ_2-$\epsilon = 64/255$)	92.7	85.6	75.2	90.7	**87.5**
AT (ℓ_2-$\epsilon = 128/255$)	90.2	**85.7**	**78.4**	89.6	87.1

For all models, we run the experiment three times independently and report the average result. We omit the standard deviations since they are small (less than 0.5%) and do not affect our claims.

5.2. Accuracy and robustness evaluation

The results of the experiments conducted on the CIFAR-100 and CIFAR-10 datasets are presented in Table 1 and Table 2, respectively.

We first discuss the natural and robust accuracy performance of SAM. From the tables, we can see that all the models trained with SAM exhibit significantly better natural accuracy and robustness compared to those trained with standard training (ST). In particular, higher robustness is achieved by using larger values of ρ with SAM. For the CIFAR-100 dataset, the model trained with $\rho = 0.4$ demonstrates even multiple robust accuracy than ST, and its natural accuracy is still higher than that of ST. Compared to the improvement in natural accuracy (approximately 2%), the increase of robustness is more significant (**more than 10% in average**). Similarly, for the CIFAR-10 dataset, the model trained with SAM also outperforms ST in terms of clean accuracy and exhibits significant higher robustness than ST. Therefore, we can conclude that SAM with a relatively larger weight perturbation bound ρ is a promising technique for enhancing the performance of models in terms of adversarial robust accuracy without sacrificing natural accuracy.

Regarding adversarially trained models, although there remains a large gap between the robustness obtained by SAM and AT, all adversarially trained models exhibit lower natural accuracy than standard training, not to mention SAM. Particularly, for ℓ_∞-adversarial training, even training with perturbation bound $\epsilon = 1/255$ decreases natural accuracy at 3.8% for CIFAR-100 and 0.9% for CIFAR-10 datasets, respectively. And also note that the larger perturbation bound ϵ used in AT, the worse natural accuracy is obtained by the corresponding model. Therefore, a key benefit of using SAM instead of AT is that there is no decrease in clean accuracy. Additionally, note that solving the PGD process in AT results in significant computational overhead. Specifically, since we use 10-step PGD, all AT experiments require 10 times more computational cost compared to ST, while SAM only requires 1 time more.

Based on the discussion above, we reach the conclusion that SAM-trained models perform significantly better robustness without decreasing any natural accuracy compared to standard training methods. Furthermore, another benefit of SAM is that it does not require significant computational resources. Therefore, we point out that SAM can serve as a lightweight alternative to AT, which can improve robustness without a decrease in natural accuracy and significant training overhead.

6. Conclusion

In this paper, we show that using Sharpness-Aware Minimization (SAM) alone can improve adversarial robustness, and reveal the fundamental relation between SAM and Adversarial Training (AT). We empirically and theoretically demonstrate that both SAM and AT add perturbations to features to achieve better robust generalization ability. However, SAM adds moderate perturbations implicitly, while AT adds strong perturbations explicitly. Consequently, they lead to different accuracy and robustness trade-offs. We further conduct experiments on benchmark datasets to verify the validity of our proposed insight. Finally, we suggest that SAM can serve as a lightweight substitute for AT under certain requirements.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Bahri, D., Mobahi, H., and Tay, Y. Sharpness-aware minimization improves language model generalization. *arXiv preprint arXiv:2110.08529*, 2021.

Bai, Y., Feng, Y., Wang, Y., Dai, T., Xia, S.-T., and Jiang, Y. Hilbert-based generative defense for adversarial examples. In *ICCV*, 2019.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., and Zhu, J. Robust classification via a single diffusion model, 2023a.

Chen, H., Zhang, Y., Dong, Y., and Zhu, J. Rethinking model ensemble in transfer-based adversarial attacks, 2023b.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Neural Information Processing Systems*, 2019.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Jetly, V., Ibayashi, H., and Nakano, A. Splash in a flash: Sharpness-aware minimization for efficient liquid splash simulation. 2022.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kim, M., Li, D., Hu, S. X., and Hospedales, T. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *arXiv preprint arXiv:2210.05177*, 2022.

Mo, Y., Wu, D., Wang, Y., Guo, Y., and Wang, Y. When adversarial training meets vision transformers: Recipes from training to architecture. In *NeurIPS*, 2022.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *SP*, 2016.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Wang, H. and Wang, Y. Generalist: Decoupling natural and robust generalization. In *CVPR*, 2023.

Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.-J., and Kolter, J. Z. Beta-crown: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network robustness verification, 2021.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In *ICML*, 2019.

Wei, Z., Wang, Y., Guo, Y., and Wang, Y. Cfa: Class-wise calibrated fair adversarial training. In *CVPR*, 2023a.

Wei, Z., Zhang, X., Zhang, Y., and Sun, M. Weighted automata extraction and explanation of recurrent neural networks for natural language tasks, 2023b.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *CVPR*, 2019.

Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*, 2021.

Yu, C., Han, B., Gong, M., Shen, L., Ge, S., Du, B., and Liu, T. Robust weight perturbation for adversarial training. *arXiv preprint arXiv:2205.14826*, 2022a.

Yu, C., Han, B., Gong, M., Shen, L., Ge, S., Du, B., and Liu, T. Robust weight perturbation for adversarial training. *arXiv preprint arXiv:2205.14826*, 2022b.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, Y., Wei, Z., Zhang, X., and Sun, M. Using z3 for formal modeling and verification of fnn global robustness, 2023.

Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., and Tao, D. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint arXiv:2210.05497*, 2022.