Improving Adversarial Robustness of DNNs via Logits Constant Amplitude Training

Asynchronous submission [*1]

Abstract

Deep neural networks (DNNs) have been shown to be vulnerable to adversarial examples, which are carefully crafted by adding imperceptible perturbations to legitimate inputs. Adversarial training is the most widely used approach to enhance adversarial robustness. However, adversarial training requires the inclusion of adversarial examples during the training process, leading to significant computational overhead. Additionally, adversarial training provides limited insights into the fundamental causes of DNNs' susceptibility to adversarial attacks. In this paper, we explore robust training without the need for adversarial examples, instead utilizing examples perturbed with uniform noise. Specifically, we introduce a new logits norm regularization training paradigm, termed logits constant amplitude training (LCAT). LCAT consists of two strategies: logits amplitude alignment (LAA) and scaled bounded activation function (S-BAF). LAA aligns logits amplitudes between legitimate and uniform random noised examples, while S-BAF squeeze the logits values, increasing attack cost. Along with the proposed LAA loss scaled Tanh activation function, LCAT improve adversarial robustness of DNNs over 10% on CIFAR10 compared to the state-of-the-art methods.

1. Introduction

Deep neural networks (DNNs) have been widely applied in various domains, including face recognition and machine translation, owing to their exceptional performance(LeCun et al., 2015). However, their vulnerability to adversarial attacks has drawn increasing attention from researchers and practitioners alike(Bartoldson et al., 2024; Papernot et al., 2016). Among the numerous strategies proposed to address this issue, adversarial training has emerged as the most prominent and widely adopted method for enhancing the adversarial robustness of DNNs.

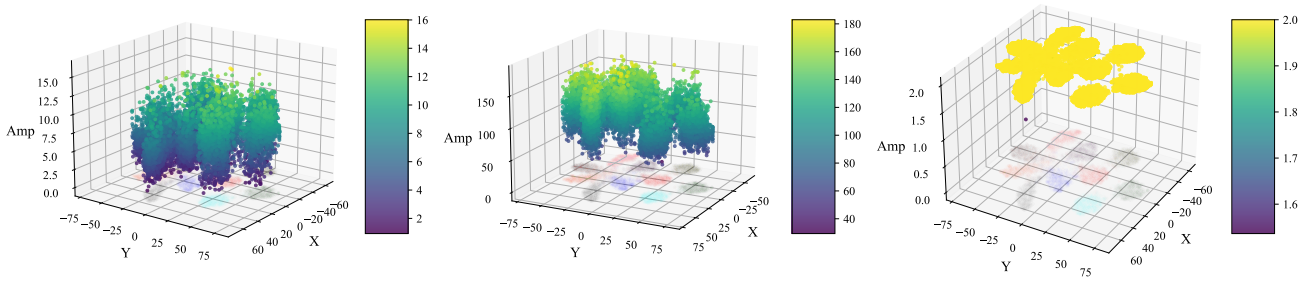
Adversarial training is a widely adopted method for enhancing the performance of deep neural networks (DNNs) against adversarial examples(Goodfellow et al., 2015). Specifically, standard adversarial training involves training DNNs using adversarial examples instead of legitimate examples. Variants of this approach include ensemble adversarial training(Tramèr et al., 2018), which leverages adversarial examples generated from multiple DNNs, and self-ensemble adversarial training, which updates model parameters using a momentum strategy(Wang & Wang, 2022). As outlined above, adversarial training relies heavily on the crafting of adversarial examples. However, adversarial examples generated by different methods often exhibit varying distributions. This raises an important open question: what types of examples are most effective for training DNNs to achieve robust adversarial performance?

Moreover, to the best of our knowledge, some theoretical studies have demonstrated that training DNNs with randomly perturbed examples can enhance adversarial robustness. Despite this promise, no existing method has achieved satisfactory performance against gradient-based attacks, such as FGSM (Goodfellow et al., 2015) and PGD (Kurakin et al., 2017) , and most fail entirely against stronger attacks like AutoAttack(Croce & Hein, 2020).

To address these challenges, this paper proposes a novel training paradigm aimed at improving the adversarial robustness of DNNs by introducing a logits norm interval constraint. The contributions of this paper are summarized as follows.

- We observe a "logits quasi-constant amplitude" phenomenon in adversarially robust DNNs, where the difference between the maximum and minimum logits remains constant.
- We propose a logits amplitude alignment (LAA) loss that enhances the adversarial robustness of DNNs while significantly reducing computational complexity.
- We design a scale-bounded activation function for DNNs. When combined with the LCAT loss, the proposed method outperforms existing approaches by more than 10% against AutoAttack on CIFAR-10.

As illustrated in Fig. 1, the logits amplitudes of adversarially



(a) MART (b) ORI (c) LCAT

Figure 1. Comparison of Logits Amplitude among adversarial robustness and vanilla models on CIFAR-10. In this figure, X-Y plane are tsne(Maaten & Geoffrey, 2008) manifold of test dataset, while Amp-axis represents logits amplitude. Here Amp is the logits amplitude.

robust models are noticeably more concentrated compared to those of a standard (vanilla) model. This observation highlights a key characteristic of adversarial robustness: models that exhibit condensed logits distributions tend to be less susceptible to adversarial perturbations. This insight serves as the primary motivation for developing the Logits Constant Amplitude Training (LCAT) method. By explicitly encouraging a more uniform and consistent logits amplitude distribution, LCAT aims to enhance DNNs's robustness against adversarial attacks while maintaining high performance on clean data.

2. Related work

Adversarial training (AT) is most popular method to improve adversarial robustness of DNNs. AT training DNNs with adversarial examples, which can be formulate as an mimmax optimization problem as follows.

$$\min_{\theta} \sum_{(x,y)\mathcal{D}} \max CE(f(x+\sigma;\theta),y) \tag{1}$$

TRADES (Zhang et al., 2019) and MART (Wang et al., 2020)are typical adversarial training methods. The two methods are similar, and their loss can be formulated as follows.

$$\begin{aligned}\mathcal{L}_{TRADES} = CE(f(x;\theta),y)+\\ \lambda KL(f(x;\theta),f(x_{adv};\theta))\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{MART} = CE(f(x_{adv};\theta),y)+\\ \lambda KL(f(x;\theta),f(x_{adv};\theta))(1-f(x;\theta)_y)\end{aligned} \tag{3}$$

here f is the deep model, and θ is parameters set of f. x is an example with the label y. CE is the cross-entropy, and KL is the KL-divergence.

Inverse cross entropy (ICE) (Chen et al., 2024)is also an method to improve adversarial robustness of DNNs.

$$\begin{aligned}\mathcal{L}_{ICE} = CE(f(x_{adv};\theta),y)-\\ \lambda\sum_{i=1}^{i=C} CE(f(x_{adv};\theta),y_i)\end{aligned} \tag{4}$$

Prior-Guided Knowledge (PGK) (Jia et al., 2024) adopts positive prior-guided adversarial initialization to prevent overfitting by improving adversarial example quality without extra training costs. The loss function of PGK can be formulated as follows.

$$\begin{aligned}\mathcal{L}_{PGK} = CE(f(x_{adv};\theta),y)-\\ \|f(x+\delta_{adv};\theta)-f(x+\delta_{pgi};\theta)\|_2^2\end{aligned} \tag{5}$$

Here δ_{adv} and δ_{pgi} is single-step adversarial perturbation and prior-guided initiation of adversarial perturbation.

3. Logits Constant Amplitude Training for DNNs

As shown in Fig. 2, our method is consist of bounded activation function (BAF) and logits constant amplitude loss. Here, DNNs are abbreviated as non-linear layer (NLL) and linear layer (LL). The logits constant amplitude loss aligns logits amplitude between legitimate and noised examples, i.e., logits amplitude alignment (LAA). While bounded activation function is introduced aiming accelerate the convergence of LCAT. Moreover, to explore more prior trade off between accuracy and adversarial robustness, the classification loss of noised examples is adopted in the training process. The details of LAA and BAF will be introduced in the following paragraphs.

Definition For given deep neural network classifier with specific activating function, $\forall x \in \mathbf{D}$, its corresponding output O, namely logit, $O_{max} - O_{min} \leq \gamma$. Here $\mathbf{D}$ is dataset, x is an example in $\mathbf{D}$. O^x_{max} and O^x_{min} is max value and min value of O, respectively. γ is a constant approximate to zero. x can be legitimate or noise example.

The logit constant-amplitude phenomenon demonstrates that stability of outputs of DNNs. It can be considered as an overfitting of DNNs. Logits constant amplitude loss is defined as follows.

$$\begin{aligned}\mathcal{L} = \sum_{\substack{B\in\mathcal{D}\\ B^*\in\mathcal{D}^*}} & focal(f(B^*)) + \lambda(\|\mathbf{l}(B^*) - \beta)\| + \\ & \|mean(\mathbf{l}(B^*)) - mean(\mathbf{l}(B))\|); \\ & where\ \mathbf{l}(B) = \mathbf{O}^x_{max} - \mathbf{O}^x_{min}\end{aligned} \tag{6}$$

Here B and B^* is a batch of examples-label pairs and its corresponding noised examples-label pairs. $\mathbf{O}_{max}$ and $\mathbf{O}_{min}$ is set of outputs w.r.t. B.

In this paper, we proposed the scaled Tanh activation, which can be formulated as follows.

$$S - Tanh(x) = \begin{cases} \dfrac{e^x - e^{-x}}{e^x + e^{-x}}; Traing = Ture \\ \dfrac{e^{s\cdot x} - e^{-s\cdot x}}{e^{s\cdot x} + e^{-s}}; Traing = False \end{cases} \tag{7}$$

Here s is a constant greater than 1.

3.1. Theoretical analysis

Here for the given DNNs $f(\cdot)$, we assume Lipschitz continuity, that is, $\forall x$, $\|f(x_y - f(x+\delta)_y\| \leq L\|\delta\|$. if $f(\cdot)$ is logits constant amplitude, Here let $z_x = f(x)_y, z_\delta = f(x+\delta)_y$. The probability distribution of $\|\delta\|_p$ is $p(\|\delta\|_p)$, where $\|\delta\|$. Then it can be concluded that $\mathbb{E}[\|z_\delta - z_x\|] \leq \int_0^{\frac{c}{L}} L\|\delta\|_p \cdot p(\|\delta\|_p)\, d\|\delta\|_p + \int_{\frac{c}{L}}^{\infty} c \cdot p(\|\delta\|_p)\, d\|\delta\|_p$.

Proof Since $f(\cdot)$ is logit constant amplitude, it can be concluded that $\|z_\delta - z_x\| \leq min(c, L\|\delta\|_p)$. Then we have

$$\begin{aligned}\mathbb{E}[\|z_\delta - z_x\|] \leq & \int_0^{\frac{c}{L}} L\|\delta\|_p \cdot p(\|\delta\|_p)\, d\|\delta\|_p + \\ & \int_{\frac{c}{L}}^{\infty} c \cdot p(\|\delta\|_p)\, d\|\delta\|_p\end{aligned} \tag{8}$$

Since, the distribution of δ is a uniform distribution or some form of known distribution, then we have

$$\int_0^{\frac{c}{L}} L\|\delta\|_p \cdot p(\|\delta\|_p)\, d\|\delta\|_p = L \cdot \mathbb{E}[\|\delta\|_p \big| \|\delta\|_p \leq \frac{c}{L}] \tag{9}$$

Moreover, s is constant. Then

$$\int_{\frac{c}{L}}^{\infty} c \cdot p(\|\delta\|_p)\, d\|\delta\|_p = c \cdot P(\|\delta\| > \frac{c}{L}) \tag{10}$$

As aforementioned, we have

$$\begin{aligned}\mathbb{E}[\|z_\delta - z_x\|] \leq L \cdot \mathbb{E}[\|\delta\|_p \big| \|\delta\|_p \leq \frac{c}{L}] + \\ c \cdot P(\|\delta\|_p > \frac{c}{L})\end{aligned} \tag{11}$$

This simplification result indicates that the upper bound of the expected loss increment depends on two components: one is the expected perturbation that is less than c/L, and the other is the probability of perturbations exceeding c/L. Together, these two factors jointly constrain the maximum change in loss under adversarial perturbations.

4. Experiment

4.1. Experiment Sets

Three commonly used datasets are employed to evaluate performance: MNIST (Lecun et al., 1998), CIFAR-10 (Krizhevsky, 2009), and Tiny-ImageNet (Le & Yang, 2015). Specifically, MNIST is a 10-class grayscale image dataset with 60,000 training samples and 10,000 test samples. The CIFAR-10 dataset consists of 50,000 training images and 10,000 test images, divided into 10 categories, with color images. Tiny-ImageNet contains 100,000 training images and 10,000 test images across 200 categories.

Moreover, EfficientNetB0 (Tan & Le, 2019) and ResNet18 (He et al., 2016) are trained on MNIST and CIFAR-10, while WideResNet34 and ResNet50 are used for Tiny-ImageNet. The data augmentation techniques for Tiny-ImageNet include RandomAffine and RandomGrayscale, implemented using PyTorch[1]. Additionally, the scale factor $s$in Eq. 7 is set to 10^2.

4.2. Adversarial Robustness Evaluation Against White-box Attacks

In this subsection, we evaluate the adversarial robustness of LCAT by comparing it with several state-of-the-art methods. AutoAttack is used to assess the adversarial robustness of the methods discussed in this paper. Specifically, the standard version of AutoAttack, which includes four adversarial attack methods—APGD-CE, APGD-T, FAB-T, and SQUARE—is employed. Among these methods, APGD-CE and APGD-T generate adversarial examples using gradients, while FAB-T formulates the adversarial attack as an optimization problem, minimizing the distance that causes misclassification in DNNs. In AutoAttack, the adversarial attack methods are applied in a cascading manner: the perturbed examples generated by one method are passed to the next, iteratively crafting increasingly effective adversarial examples.

The experimental results are presented in Table 1. APGD-

[1]https://pytorch.org/

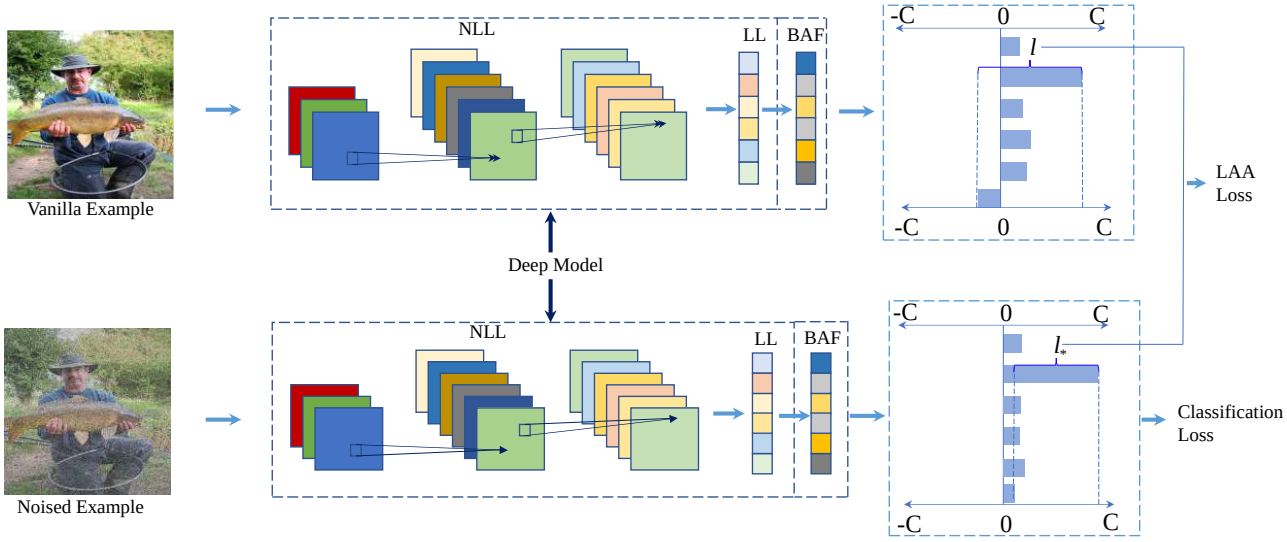


Figure 2. The workflow of logits constant amplitude training. Here we add bounded activation function to vanilla deep model. Vanilla deep model consist of non-linear layer and linear layer. LCAT consists of two strategies: logits amplitude alignment (LAA) and scaled bounded activation function (S-BAF). LAA aligns logits amplitudes between legitimate and uniform random noised examples, while S-BAF squeeze the logits values, increasing attack cost.

CE, APGD-T, FAB-T, and SQUARE represent the adversarial accuracy achieved after the corresponding attacks, while LEG denotes the standard accuracy, i.e., the accuracy on legitimate examples. EffNet, ResNet, and WRN correspond to EfficientNet, ResNet, and WideResNet, respectively. Additionally, $0.1|4/255$ indicates an attack intensity of 0.1 for MNIST and 4/255 for CIFAR-10 and Tiny-ImageNet; similarly, $0.2|8/255$ follows the same format. For LCAT, the noise intensity is set to 20/255 for CIFAR-10 and Tiny-ImageNet, and 0.8 for MNIST. The balance factor λ is set to 15 for MNIST and CIFAR-10, and 0.1 for Tiny-ImageNet.

According to Table 1, the proposed LCAT achieves the highest robustness accuracy in most scenarios. Moreover, we observe that ResNet18 outperforms EfficientNet, demonstrating that increasing model capacity benefits adversarial robustness. The effectiveness of LCAT stems from the logits amplitude alignment loss, which stabilizes the logits, effectively smoothing the decision boundary. This stabilization makes it harder for adversaries to extract useful information, such as gradients, to craft adversarial examples.

Moreover, the proposed scaled Tanh activation function further compresses the values of logits towards infinity or negative infinity. This means that the changes in logits caused by adversarial examples will be suppressed by the scaled Tanh activation function. With stable logits and the squeezing strategy, LCAT can effectively enhance the adversarial robustness of DNNs. Additionally, since PGK converges slowly on CIFAR-10, we do not report the corresponding experimental results. We observe that adversarial examples slow down convergence, whereas specific intensity random noised examples accelerate convergence. This suggests that the adversarial vulnerability of DNNs may be linked to the sparsity of the training set.

4.3. Black-box Adversarial Robustness Evaluation

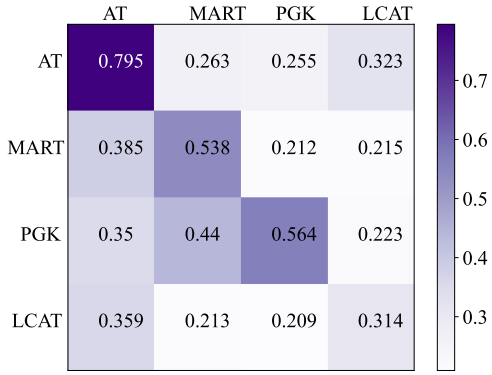
In this subsection, we conduct experiments to evaluate the adversarial robustness of LCAT in a black-box adversarial scenario. Specifically, we first investigate the adversarial transfer rate among the methods involved in this paper, including AT, MART, PGK, and LCAT. Additionally, since the scaled bounded Tanh activation is a key strategy in LCAT, we also examine the adversarial example transferability among DNNs trained with different scale factors s.

To investigate the adversarial transfer rate among different defense methods, we select APGD-CE and SQUARE as the adversarial attack methods. APGD-CE evaluates the gradient alignment between different methods, while SQUARE assesses the similarity of their decision boundaries. The experimental results are shown in Fig. 3.

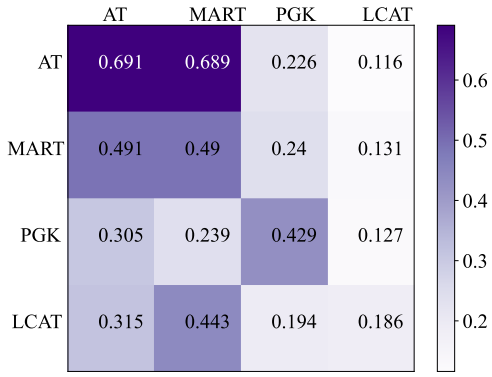
In Fig. 3, the models on the horizontal axis represent the victim models, while the models on the vertical axis represent the substitute models. Adversarial examples are crafted by the substitute model to attack the victim model. The values on the diagonal represent the success attack rate on the substitute models themselves. As shown in Fig. 3, adversarial examples crafted by PGK and LCAT are more likely to transfer to AT and MART, while adversarial examples crafted by

Table 1. Comparative analysis of the accuracy and robustness of TRADES, MART, PGK, and LCAT on the MNIST, CIFAR-10, and Tiny-ImageNet datasets.

Dataset	Method	Model	0.1|4/255				0.2|8/255				LEG
			APGD-CE	APGD-T	FAB-T	SQUARE	APGD-CE	APGD-T	FAB-T	SQUARE	
CIFAR10	VIN	EffNet	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.920
		ResNet	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.933
	AT	EffNet	0.493	0.492	0.492	0.492	0.364	0.364	0.364	0.950	0.776
		ResNet	0.952	0.601	0.599	0.559	0.559	0.353	0.353	0.353	0.780
	TRADES	EffNet	0.514	0.514	0.514	0.497	0.383	0.356	0.356	0.356	0.786
		ResNet	0.674	0.654	0.654	0.654	0.407	0.384	0.384	0.384	0.834
	MART	EffNet	0.473	0.423	0.423	0.423	0.374	0.307	0.307	0.307	0.793
		ResNet	0.665	0.636	0.636	0.636	0.454	0.403	0.403	0.403	0.833
	PGK	EffNet	-	-	-	-	-	-	-	-	-
		ResNet	0.837	0.700	0.675	0.952	0.557	0.497	0.487	0.497	0.817
	LCAT	EffNet	0.788	0.778	0.778	0.727	0.782	0.755	0.755	0.643	0.785
		ResNet	0.892	0.890	0.891	0.891	0.898	0.893	0.883	0.816	0.900
MNIST	VIN	EffNet	0.788	0.778	0.778	0.727	0.777	0.756	0.756	0.636	0.785
		ResNet	0.952	0.952	0.952	0.952	0.981	0.881	0.878	0.878	0.834
	AT	EffNet	0.942	0.942	0.942	0.942	0.938	0.938	0.938	0.938	0.945
		ResNet	0.952	0.952	0.952	0.952	0.951	0.944	0.944	0.944	0.969
	TRADES	EffNet	0.982	0.982	0.982	0.982	0.988	0.950	0.950	0.950	0.989
		ResNet	0.952	0.952	0.952	0.952	0.981	0.881	0.878	0.878	0.990
	MART	EffNet	0.983	0.983	0.983	0.983	0.952	0.952	0.952	0.951	0.988
		ResNet	0.952	0.960	0.960	0.952	0.960	0.930	0.930	0.930	0.987
	PGK	EffNet	0.092	0.090	0.090	0.090	0.076	0.076	0.076	0.076	0.191
		ResNet	0.711	0.437	0.435	0.435	0.516	0.070	0.050	0.050	0.975
	LCAT	EffNet	0.965	0.958	0.958	0.880	0.965	0.947	0.947	0.535	0.975
		ResNet	0.981	0.977	0.977	0.968	0.976	0.972	0.972	0.816	0.989
Tiny-ImageNet	VIN	WRN	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.512
		ResNet	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.490
	AT	WRN	0.245	0.232	0.232	0.232	0.197	0.189	0.189	0.431	0.451
		ResNet	0.207	0.198	0.165	0.165	0.164	0.147	0.116	0.116	0.389
	TRADES	WRN	0.236	0.225	0.164	0.160	0.201	0.215	0.213	0.213	0.495
		ResNet	0.005	0.005	0.005	0.005	0.001	0.001	0.001	0.001	0.481
	MART	WRN	0.291	0.231	0.230	0.230	0.153	0.104	0.104	0.103	0.306
		ResNet	0.232	0.183	0.183	0.180	0.127	0.006	0.005	0.05	0.287
	PGK	WRN	0.092	0.009	0.009	0.009	0.081	0.076	0.076	0.076	0.215
		ResNet	0.137	0.120	0.120	0.012	0.108	0.011	0.011	0.011	0.305
	LCAT	WRN	0.009	0.006	0.006	0.006	0.005	0.001	0.001	0.001	0.322
		ResNet	0.291	0.264	0.264	0.262	0.236	0.163	0.164	0.140	0.386



(a) APGD-CE



(b) SQUARE

Figure 3. Adversarial transfer rate between different adversarial defense methods, such as AT, MART, PGK, and LCAT.

AT and MART rarely transfer to PGK and LCAT. Notably, the adversarial example transfer rate is high between AT and MART. We believe this is due to the adversarial examples used during training, which may affect transferability. AT and MART use PGD adversarial examples, while PGK uses FGSM adversarial examples, and LCAT uses uniform random-noised examples. Additionally, we observe that the adversarial example transfer rate is low (less than 19%) against LCAT, suggesting that it is difficult for other methods to successfully attack LCAT. Furthermore, we conduct experiments to investigate the adversarial example transfer rate between models trained with different scale factors for LCAT. Similar to Fig. 3, the models on the horizontal axis represent the victim models, while those on the vertical axis represent the substitute models. APGD-CE is selected as the adversarial attack method to craft adversarial examples due to its superior performance. The experimental results are shown in Fig. 4.

According to Fig. 4, it can be observed that adversarial examples are more likely to transfer to models trained with a larger scaling factor s. However, when $s$is too small, the model itself tends to be more vulnerable. As mentioned earlier, we propose the scaled Tanh activation function to improve the adversarial robustness of DNNs. During training, a smaller s forces the logits to approximate infinity or negative infinity, while during testing, a larger s squeezes the outputs of the scaled bounded activation function to its upper and lower bounds.

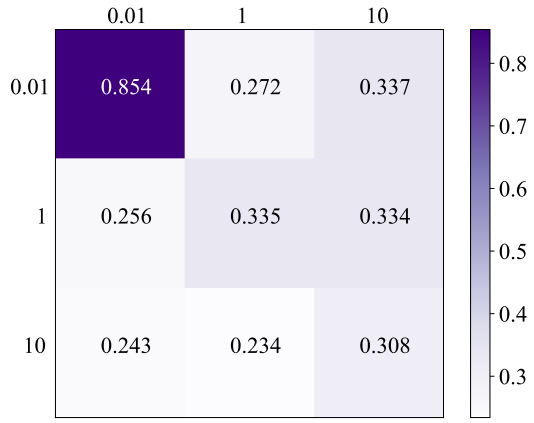


Figure 4. Adversarial transfer rate confusion matrix of variant scale factor s in Eq. 7 on CIFAR10. Here the models in horizontal axis are the victim model, while models in the vertical axis are the substitution model. 0.01, 1, and 10 represents different scale factor for LCAT.

4.4. Impact of λ on LCAT Performance

In this subsection, we conduct experiments to evaluate the impact of different balance factors λ from Eq. 6 on LCAT. As mentioned previously in Eq. 6, λ is a hyperparameter that balances the importance of the classification loss and the LAA loss. Specifically, we compare the performance of LCAT with various values of λ. The experimental results are shown in Fig. 5. In Fig. 5, APGD-CE, APGD-T, FAB, and SQUARE represent the adversarial accuracy after the corresponding attacks, while LEG denotes the standard accuracy. The model architecture used is ResNet18, and the noise intensity is set to 20/255.

As shown in Fig. 5, the standard accuracy decreases gradually, while the adversarial robustness accuracy of the DNN increases within a specific range of λ, then decreases as λ continues to increase. This behavior can be attributed to the LAA loss in LCAT, which improves the adversarial robustness of DNNs. As λ increases, the DNN becomes more robust; however, there is a trade-off between standard accuracy and robustness accuracy. Increasing λ harms the standard accuracy, leading to a decline in robustness accuracy. This is because when a legitimate example is misclassified by the DNN, its corresponding adversarial example is more likely to be misclassified as well. As shown in Fig. 5, when λ is approximately 15, the model achieves

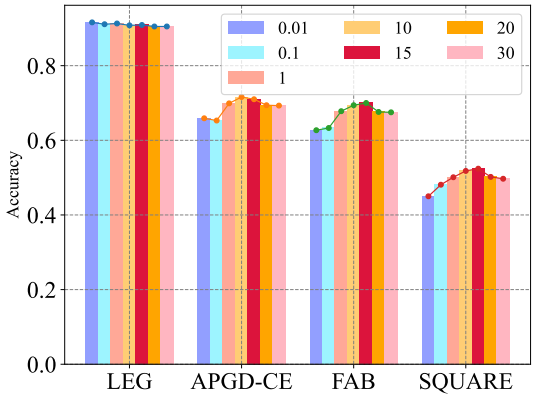


Figure 5. Comparison of LCAT performance with various λ in Eq. 6 on CIFAR10. APGD-CE, APGD-T, FAB, and SQUARE represents adversarial accuracy after corresponding attack. LEG represent standard accuracy.

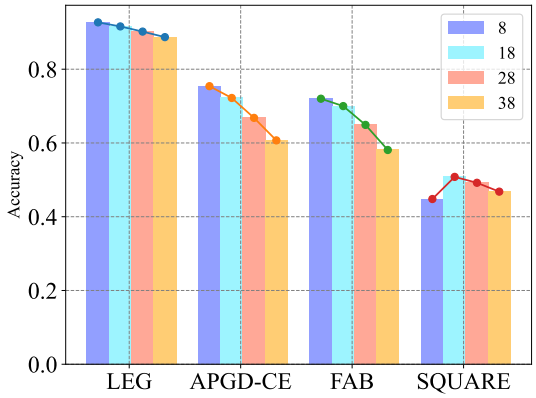


Figure 6. Comparison of LCAT performance with various noise intensity on CIFAR10. APGD-CE, APGD-T, FAB, and SQUARE represents adversarial accuracy after corresponding attack. LEG represent standard accuracy.

the highest robustness accuracy. Based on the experimental results, we believe that the optimal setting for λ depends on the dataset complexity. The more complex the dataset, the smaller λ should be.

4.5. Impact of Intensity of Noise on LCAT Performance

In this subsection, we conduct an experiment to investigate the impact of noise intensity on LCAT. We align the logits amplitudes between legitimate examples and their corresponding uniform random noise-perturbed examples to stabilize the logits. The experiments are conducted on CIFAR-10, with λ set to 15. The experimental results are shown in Fig. 6.

As shown in Fig. 6, it can be observed that standard accuracy decreases as noise intensity increases. However, adversarial robustness accuracy initially increases and then decreases with increasing noise intensity. This is because noise disrupts the semantic features of legitimate examples, and overwhelming noise intensity leads to a decline in standard accuracy. This effect is similar to the impact of varying the balance factor λ in Eq. 6. When a legitimate example is misclassified by DNNs, its corresponding adversarial example is more likely to be misclassified as well.

4.6. Impact of Bounded Activation Function

In this subsection, we investigate the impact of bounded activation functions on the performance of the proposed logits constant amplitude training (LCAT) framework. Specifically, we compare the performance of LCAT with various bounded activation functions, including Tanh and Sigmoid, which inherently constrain the output range of neural activations. To further highlight the role of bounded activation functions, we also include the unbounded activation function, Softplus, as a baseline.

The experiments are conducted on CIFAR-10 using the ResNet18 architecture. Additionally, $\lambda$is set to 15, and the uniform noise intensity is set to 20/255. APGD-CE and SQUARE adversarial attack methods are chosen to assess the adversarial robustness of models with different activations. The experimental results are shown in Fig. 7.

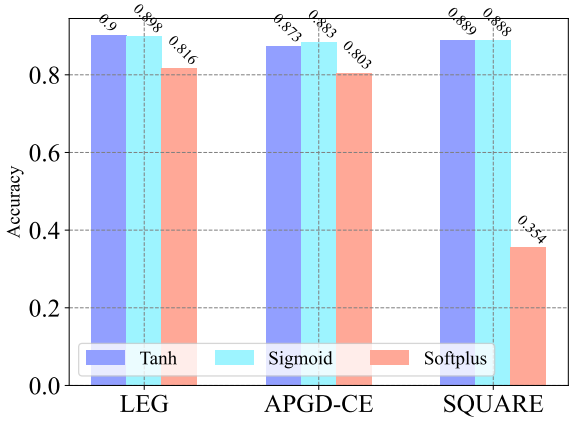


Figure 7. Comparison of LCAT performance with various bounded activation on CIFAR10. APGD-CE and SQUARE represents adversarial accuracy after corresponding attack. LEG represent standard accuracy.

According to Fig. 6, we can see that LCAT with various activation function can achieve adversarial robustness. However, unbounded activation function is inferior the bounded activation functions, such as Tanh and Sigmoid. The reason is that the unbounded activation function can not squeeze the logits values and converge slowly. Moreover, we note that Tanh prior Sigmoid. The reason is that

Table 2. Case study investigating different strategies to improve adversarial robustness of DNNs, such as ICE, ADV, NOI, Tanh, and LAA on CIFAR10. Here, ICE, ADV, NOI, Tanh, and LAA represents inverse cross entropy loss, adversarial training, training with uniform random noised example, scaled Tanh activation funciton, and logits amplitude aligning loss, respectively.

Strategies	LEG	APGD-CE	APGD-T	FAB-T	SQUARE
CE	0.934	0.000	0.000	0.000	0.000
ICE	0.921	0.013	0.000	0.000	0.000
CE+NOI	0.921	0.012	0.000	0.000	0.000
ICE+NOI	0.872	0.227	0.112	0.004	0.004
CE+Tanh	0.927	0.000	0.000	0.000	0.000
ICE+Tanh	0.884	0.026	0.001	0.001	0.000
CE+NOI+Tanh	0.892	0.334	0.317	0.315	0.197
ICE+NOI+Tanh	0.836	0.124	0.003	0.002	0.002
LAA+ADV+Tanh	0.793	0.480	0.423	0.423	0.423
LAA+NOI+Tanh	0.900	0.888	0.880	0.880	0.806

within infinity and negative infinity Tanh is more smoothing, $x \to \infty : Tanh^{'}(x) \to 0; Sigmoid^{'}(x) \to 1$. Tanh is more likely to squeeze logits values its bound. These results shed light on the potential benefits of leveraging bounded activation functions to enhance the adversarial robustness of DNNs.

4.7. Ablation Experiments

In this subsection, we evaluate the effectiveness of bounded activation functions and noisy examples. The training strategies discussed in this paper are summarized as follows: LAA loss, scaled Tanh activation function (Tanh), and high-intensity noisy examples (NOI). We then compare the performance of DNNs trained with different combinations of these strategies. The experiments are conducted on CIFAR-10 using the ResNet18 architecture.

The experimental results are shown in Table 2. CE and ICE represent deep models trained with cross-entropy and inverse cross-entropy loss, respectively. NOI indicates that the deep model is trained with uniformly random-noised examples, with the noise intensity set to 20/255. 'Tanh' refers to the addition of a scaled Tanh activation function to the last layer of ResNet18. ADV means the deep model is trained with PGD-5 adversarial examples, with the attack intensity set to 8/255.

As shown in Table 2, We find that ICE can enhance the adversarial robustness of deep models against the APGD-CE attack. This is because ICE induces gradient confusion, causing the gradient to approximate zero. Training deep models with NOI also improves adversarial robustness against APGD-CE, APGD-T, and FAB-T attacks. The reason for this is that NOI training helps smooth the decision boundary. Moreover, two counterintuitive phenomena are observed: ICE+NOI+Tanh performs worse than CE+NOI+Tanh, and LAA+ADV+Tanh performs worse than LAA+NOI+Tanh. The first phenomenon can be attributed to the fact that ICE introduces gradient confusion, meaning that the values of the logits become similar to each other. However, since $x \to 0 : Tanh^{'}(x) \to 0.5$, excessively large gradients hinder the convergence of deep models.. This observation is consistent with the large decline in standard accuracy for ICE+Tanh compared to ICE. The second counterintuitive phenomenon involves the inconsistency between model logits stability and overfitting to adversarial examples. While LAA aims to stabilize the logits, adversarial training (ADV) encourages the model to overfit to adversarial examples, which destabilizes the logits. As a result, the combination of LAA and ADV leads to unstable convergence. This explains why adversarial training often causes a decline in standard accuracy. Therefore, we conclude that Tanh, when combined only with the LAA loss (i.e., LCAT), is the most effective approach for improving adversarial robustness.

5. Conclusion

In this paper, we propose a new training paradigm called LCAT, which consists of two strategies: LAA and S-BAF. LAA aligns the logits amplitudes between legitimate examples and those with uniform random noise, while S-BAF reduces the logits values, thereby increasing the cost of adversarial attacks. LCAT outperforms existing methods by over 10% against AutoAttack on CIFAR-10. The proposed method demonstrates that adversarial robustness of DNNs can be improved with minimal additional computational overhead. LCAT is motivated by the observation of the quasi-constant amplitude phenomenon in logits, offering a novel perspective for addressing adversarial challenges in DNNs. Future work will explore more efficient bounded activation functions.

References

Bartoldson, B. R., Diffenderfer, J., Parasyris, K., and Kailkhura, B. Adversarial robustness limits via scaling-law and human-alignment studies. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=HQtTg1try7.

Chen, Y., Hu, P., Yuan, Z., Peng, D., and Wang, X. Integrating confidence calibration and adversarial robustness via adversarial calibration entropy. *Inf. Sci.*, 668:120532, 2024. doi: 10.1016/J.INS.2024.120532. URL https://doi.org/10.1016/j.ins.2024.120532.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Jia, X., Zhang, Y., Wei, X., Wu, B., Ma, K., Wang, J., and Cao, X. Improving fast adversarial training with prior-guided knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9):6367–6383, 2024. doi: 10.1109/TPAMI.2024.3381180. URL https://doi.org/10.1109/TPAMI.2024.3381180.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=BJm4T4Kgx.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Maaten, L. V. D. and Geoffrey, H. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605): 2579–2605, 2008.

Papernot, N., McDaniel, P. D., Sinha, A., and Wellman, M. P. Towards the science of security and privacy in machine learning. *CoRR*, abs/1611.03814, 2016.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., and McDaniel, P. D. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Wang, H. and Wang, Y. Self-ensemble adversarial training for improved robustness. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oU3aTsmeRQV.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rklOg6EFwS.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 7472–7482, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.