Explaining Adversarial Robustness of Neural Networks from Clustering Effect Perspective

Yulin Jin[1], Xiaoyu Zhang[1*], Jian Lou[2], Xu Ma[3], Zilong Wang[1] Xiaofeng Chen[1],
[1]State Key Laboratory of Integrated Service Networks (ISN), Xidian University
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[3]School of Cyber Science and Engineering, Qufu Normal University
jyl990903@163.com xiaoyuzhang@xidian.edu.cn jian.lou@zju.edu.cn
xma@qfnu.edu.cn zlwang@xidian.edu.cn xfchen@xidian.edu.cn

Abstract

Adversarial training (AT) is the most commonly used mechanism to improve the robustness of deep neural networks. Recently, a novel adversarial attack against intermediate layers exploits the extra fragility of adversarially trained networks to output incorrect predictions. The result implies the insufficiency in the searching space of the adversarial perturbation in adversarial training. To straighten out the reason for the effectiveness of the intermediate-layer attack, we interpret the forward propagation as the ***Clustering Effect****, characterizing that the intermediate-layer representations of neural networks for samples i.i.d. to the training set with the same label are similar, and we theoretically prove the existence of Clustering Effect by corresponding Information Bottleneck Theory. We afterward observe that the intermediate-layer attack disobeys the clustering effect of the AT-trained model. Inspired by these significant observations, we propose a regularization method to extend the perturbation searching space during training, named sufficient adversarial training (SAT). We give a proven robustness bound of neural networks through rigorous mathematical proof. The experimental evaluations manifest the superiority of SAT over other state-of-the-art AT mechanisms in defending against adversarial attacks against both output and intermediate layers. Our code and Appendix can be found at* `https://github.com/clustering-effect/SAT`.

1. Introduction

While the striking success of neural networks has been deployed into diverse real-world application scenarios [9, 28, 7, 15, 31, 32], recent studies have demonstrated that deep models are brittle to a series of crafted human-imperceptible perturbations which cause the target model to produce an incorrect output [8, 13, 3, 17, 30, 21]. This phenomenon leads to a significant controversy in the application of neural networks in safety-critical scenarios, e.g., automatic driving systems [4], brain-computer interface systems, etc. Thus, resistance to adversarial perturbations on the inputs [17], *i.e.*, adversarial samples, is becoming a crucial design goal that pushes researchers to dive into proposing a sizable number of defense mechanisms for the adversarial robustness settings.

The most prosperous methodology among those defense mechanisms, *i.e.*, Adversarial Training (AT) [8, 17, 29, 22], attempts to solve a *min-max* optimization problem of the loss function. AT firstly searches the constraint perturbation added to the sample as the maximum of the loss function in the input space. Thereafter, AT updates the parameters of the model utilizing the stochastic gradient descent (SGD) algorithm to approach the minimum of loss function in the parameter space. [8, 17] choose the ordinary Cross-Entropy (CE) loss where the inner maximization is equivalent to an adversarial attack to acquire incorrect outputs. [29, 22, 25] defined rectified CE loss to characterize the distance from the sample to the decision boundary. These mechanisms obtain pleasurable achievements against output-layer attacks (OLA), *i.e.*, FGSM [8], PGD [17], AutoAttack [5].

However, a recent research[27] manifests the extra vulnerability of the above defense methods to the *intermediate-layer attack* (ILA) [27] which disturbs features extracted by intermediate layers. This novel fragility implies AT [8, 17, 29, 22, 25] can merely regularize the model to defend against OLA [8, 17, 5] rather than ILA [27]. AT merely adopts Cross-Entropy loss which entirely relies on the output of the model, so perturbations involved in training are located in the maximum of Cross-Entropy loss in input space. These perturbations are constrained in a narrow searching space during AT and can not represent the

*Corresponding Author

maximum of other loss functions. Besides, ILA constructs a novel loss function different from Cross-Entropy loss by exploiting the information of both output and intermediate layers, resulting in perturbations far from the searching space of OLA. Therefore, AT may be powerless against ILA since the searching space of ILA and OLA are dissimilar. In a nutshell, ***AT is insufficient in defense against ILA***. A host of other defense methods [1, 26, 12, 18] have been on the scene, which exploit the abnormal behaviors of the extracted features induced by adversarial perturbations. Unfortunately, this research yet hardly states a sufficient AT defending both output-layer and intermediate-layer attacks.

In this paper, we go deeply into the *insufficiency* of AT to derive a unified AT framework protecting the entire network. We set things moving by delving into the diverse impact of ILA and OLA at intermediate layers on "robust" AT-trained models. We highlight the effect of the adversarial perturbation through forward propagation which is universally acknowledged as the process of feature extraction. To fundamentally elucidate the distortion at intermediate layers induced by adversarial attacks, we raise and answer the following question as the preliminary: *"How to interpret and materialize the feature extraction process of the neural network?"*

As the answer to the question, it turns out that model extracts features progressively in a *clustering-type manner* as the features are passed forward as we called ***Clustering Effect*** of the model. To put it simply, given examples from the same label, which are i.i.d. to the training set, the outputs of the intermediate layer will close to a fixed vector in the sense of L_p norm. Specifically, the vector is the centroid of all outputs of the intermediate layer. The result is understandable which fits the intuition that the trained model similarly encodes the samples with the same label. We define the logical definition of the metric of the performance at the intermediate layer as the clustering accuracy (Clu.Acc). We find Clu.Acc converges to the classification accuracy in deep layers, indicating the model indeed extracts essential features from samples. Finally, we theoretically explain the existence of clustering accuracy by corresponding with *Information Bottleneck Theory*[19, 20].

Then, we observe the explicit distinction between ILA and OLA where the Clu.Acc does not converge to the classification accuracy under ILA on AT-trained 'robust' models. Therefore, we pinpoint that the AT-trained model does not ever train sufficiently on perturbations generated by ILA. To thoroughly eliminate the vulnerability of AT, we propose the *sufficient adversarial training (SAT)* with a regularization item characterizing the deviation from extracted features to the corresponding clustering centroid, which is then incorporated into the cross-entropy loss as our devised loss function. We adversarially train models utilizing the proposed loss function as a novel AT framework Sufficient Adversarial Training (SAT). SAT is a generalized form of previous AT since we additionally train robust intermediate layers. Mathematically, we strictly prove that minimizing the proposed regularization item is equivalent to minimizing the *Information Bottleneck* loss function, and give a proved robustness lower bound relevant to the proposed regularization item. To summarize, we make the following contributions.

- We demonstrate the Clustering Effect of the intermediate layer in extensive experiments, which characterizes the extracted features for samples i.i.d. to the training set with the same label are similar. Thereafter, we theoretically prove the existence of the Clustering Effect by corresponding *Information Bottleneck Theory*.

- We observe perturbations generated by ILA deviating from the Clustering Effect, demonstrating the AT-trained model has not ever been trained sufficiently on ILA perturbations. We further visualize the distinction between OLA and ILA at intermediate presentations clearly, which is set as our motivation.

- We propose a sufficient adversarial training framework to defend against both ILA and OLA by incorporating a regularization loss characterized by the intermediate layers' clustering effect into Cross-Entropy loss. Mathematically, we rigidly prove proved robustness lower bound relevant to the proposed regularization item.

- We demonstrate the capacity and efficiency of SAT to enhance the robustness faced against ILA. We evaluate SAT on CIFAR10, SVHN, and CIFAR100 against six state-of-the-art adversarial attacks (including 5 of OLA and 1 of ILA), manifesting its remarkable performance in improving the adversarial robustness of the neural network.

2. Related Work

Adversarial Attacks. A host of literature exposes image classification models' serious vulnerability to manufactured tiny perturbations called *adversarial perturbations*. Therefore, the attackers may pursue bringing out adversarial perturbations to impel the target model predicts incorrect outputs. Most of the adversarial attacks are realized based on the commonly used Fast Gradient Sign Attack (FGSM) [8]. BIM attack [13] is the direct variant of FGSM which iteratively conducts FGSM with a small perturbation step; PGD attack [17] further selects several initial start points in a neighbor of the input and implements BIM attack [13] in parallel; CW attack [3] transfers the constraint feasible region of the adversarial perturbation to an unconstrained

domain and maximizes a substitute loss function; AutoAttack [5] ensembles several surrogate loss functions and updates the adversarial perturbation with an adaptive step size. Recently, a novel intermediate-layer attack LAFEAT [27] searches the perturbation that mostly distorts the extract features by exploiting the vulnerability of intermediate layers in neural networks. Besides, LAFEAT [27] reports extra vulnerability of intermediate layers of "robust" models trained by AT. The findings in LAFEAT [27] promote us to devise a unified adversarial training framework, strengthening both output and intermediate layers.

Adversarial Defense. Research on the robustness of image classification models is emerging in endlessly [8, 17, 2, 6, 10]. The field of adversarial training is concentrating on solving a min-max problem of involving adversarial perturbations to training set. PGD-AT [17] implements the inner maximization by PGD attack [17] during training and involves them into the training set; TRADES [29] propose a surrogate loss function quantifying the gap from examples to the decision boundary to substitute CE loss in PGD-AT [17]. Hein et al. [10] first proposed a proved robustness bound for a two-layer fully-connected ReLu network with a cross-Lipschitz regularization loss function; Weng et al. [24] expands the bounds in [10] applicable to any network by the Extreme Value Theory; Lin et al. [14] prove a proved robustness for quantized DNNs. [6] trains model to possess more interpretable saliency maps of adversarial samples to improve the robustness of models; Bai et al. [1] attribute the poor robustness of models to the channel-wise activation of adversarial samples which is at opposite poles to that of clean samples. Yan et al. [26] further propose a suppressing method to rectify the channel-wise activation of adversarial samples. In summary, most of the defense mechanisms [8, 17, 10, 24] target to obtain adversarial robust models utilizing the output from the last layer. The other [1, 26] exploit the distortion of the features extracted by intermediate layers reduced by the output-layer attacks. However, these researches not yet take the robustness of intermediate layers into account.

3. Background

To study the exceeding effect of ILA on the AT-trained model compared to OLA, as preliminaries, we provide a brief introduction to adversarial training and its intrinsic property about feature extraction named ***Clustering Effect***. The Clustering Effect property characterizes that the intermediate-layer representations are similar for samples from the same label, which are i.i.d. to the training set. The property is significant in explaining the peculiar vulnerability of AT-trained models to ILA. Further, we theoretically correspond the Clustering Effect with *Information Bottleneck* [19, 20] theory to explain its existence.

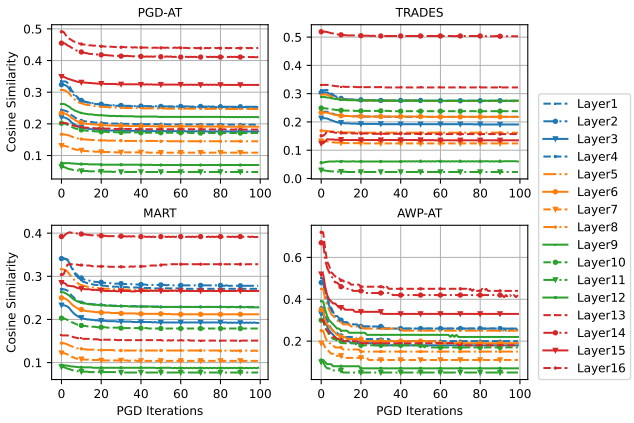

Figure 1. The intermediate-layer gradient versus the shift caused by the perturbation at different PGD_{100} iterations. The model ResNet-18 is trained on CIFAR10 dataset by present SOTA AT methods including PGD-AT, TRADES, MART, and AWP-AT.

3.1. Adversarial Training

Let $f_{\boldsymbol{\theta}}^{L+1}$ be an $L+1$ layer neural network with parameter $\boldsymbol{\theta}$ taking input $\boldsymbol{x} \in R^n$ from the input space $\mathbb{X}$ and output a probability vector $f_{\boldsymbol{\theta}}^{L+1}(\boldsymbol{x}) \in R^{\|\mathbb{Y}\|}$, where $\mathbb{Y}$ denotes the label space. A loss function $\mathcal{L} : f \times \mathbb{X} \times \mathbb{Y} \mapsto R$ measures the performance of model f given a dataset. The field of adversarial training is concentrating on solving a min-max problem in the following equation, aiming to minimize the upper bound of the loss function $\mathcal{L}$ of f in the neighbor of samples bounded by p-norm:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \max_{\|\boldsymbol{r_i}\|_p \leq \epsilon} \mathcal{L}_{CE}(f_{\boldsymbol{\theta}}^{L+1}, \boldsymbol{x_i} + \boldsymbol{r_i}, y_i), \tag{1}$$

where $\mathcal{N}$ is the size of the training set. Intuitively, adversarial training generates imperceptible adversarial perturbations automatically and involves them in training set. The inner maximization problem can be solved by a multi-step gradient ascent algorithm with a random initialization mechanism. The outer minimization problem selects a parameter $\boldsymbol{\theta}$ fitting perturbations.

So far the existing adversarial training generates perturbations based on OLA, which updates perturbations by the backward propagation. Therefore, the adversarial perturbations generated by OLA drive the l-th intermediate-layer representation on perturbed examples to the direction similar to the intermediate-layer gradient $\nabla_{f_{\boldsymbol{\theta}}^l(\boldsymbol{x})} \mathcal{L}_{CE}(f_{\boldsymbol{\theta}}^{L+1}, \boldsymbol{x}, y)$. As shown in Fig. 1, for all layers of AT-trained ResNet-18 and every PGD_{100} iteration t on CIFAR-10 testset, the shift of the intermediate-layer representation $f_{\boldsymbol{\theta}}^l(\boldsymbol{x} + \boldsymbol{r_t}) - f_{\boldsymbol{\theta}}^l(\boldsymbol{x})$ caused by the perturbation $\boldsymbol{r_t}$ is close to the intermediate-layer gradient since the average cosine similarity is always positive. The results imply that the shift in intermediate layers and the perturbation in in-

put space in AT all overfit the Cross-Entropy loss function, since the direction of the shift of the intermediate layer and perturbation are fixed by it. ***In other words, the perturbation searching space of AT can be insufficient,*** which is not compatible with the perturbation space of ILA. Therefore, attacks with other loss functions than Cross-Entropy loss may exploit extra fragility in AT-trained models. Although extensive AT frameworks spare no efforts to eliminate the generalization gap between the training set and the whole data population, they neglect the notion of the generalization in the space of loss function.

3.2. Clustering Effect

As the preliminary to reveal the intrinsic difference between the impact of ILA and OLA on the AT-trained model, we first interpret the property of the feature extraction process of AT-trained neural networks. We empirically demonstrate that the forward propagation of AT-trained neural networks extracts features progressively in a *clustering-type manner*. We assemble all samples of any label i in the training set $\mathcal{X}_{train}$ into a subset $\mathcal{X}_{train,i} = \{(\boldsymbol{x_i^j}, i)|1 \leq j \leq \mathcal{N}_{train,i}\}$, $1 \leq i \leq |\mathbb{Y}|$, and $\mathcal{X}_{train} = \bigcup_{i=1}^{|\mathbb{Y}|} \mathcal{X}_{train,i}$, where $\mathcal{N}_{train,i}$ is the cardinal of $\mathcal{X}_{train,i}$. Given a sample $\boldsymbol{x}$, we denote $f_{\boldsymbol{\theta}}^l(\boldsymbol{x})$ as the l-th intermediate-layer output of the neural network f and $\boldsymbol{\mu_i^l} = E_{\boldsymbol{x}\in\mathcal{X}_{train,i}}[f_{\boldsymbol{\theta}}^l(\boldsymbol{x})]$ as the corresponding clustering centroid, $1 \leq l \leq L+1$. Furthermore, we define the set of mean vectors of the lth intermediate-layer output $f_{\boldsymbol{\theta}}^l(\boldsymbol{x})$ as $\mu^l = \{\boldsymbol{\mu_i^l}|1 \leq i \leq |\mathbb{Y}|\}$, $f_{\boldsymbol{\theta}}^l(\boldsymbol{x}), \boldsymbol{\mu_i^l} \in R^{d_l}, d_l > 0$.

For any sample $\boldsymbol{x}$ from the test set $\mathcal{X}_{test}$, we search $\arg\max_{1\leq k\leq|\mathbb{Y}|} \|f_{\boldsymbol{\theta}}^l(\boldsymbol{x_i^j}) - \boldsymbol{\mu_k^l}\|_p$ as the clustered label. Below, we introduce the *clustering accuracy* of the model $f_{\boldsymbol{\theta}}$'s lth layer on the test set $\mathcal{X}_{test}$.

Definition 1 *Given μ^l and test set $\mathcal{X}_{test}$ split into $\mathcal{X}_{test,i} = \{(\boldsymbol{x_i^j}, i)|1 \leq j \leq \mathcal{N}_{test,i}\}$, where $\mathcal{X}_{test} = \bigcup_{i=1}^{|\mathbb{Y}|} \mathcal{X}_{test,i}$, and $\mathcal{N}_{test,i}$ is the cardinality of $\mathcal{X}_{test,i}$. The clustering accuracy of the l-th layer of the model f is defined as,*

$$\text{Clu.Acc} = \frac{\sum_{i=1}^{|\mathbb{Y}|} \sum_{j=1}^{\mathcal{N}_{test,i}} \mathcal{I}(\arg\max_{1\leq k\leq|\mathbb{Y}|} \|f_{\boldsymbol{\theta}}^l(\boldsymbol{x_i^j}) - \boldsymbol{\mu_k^l}\|_p = i)}{\sum_{i=1}^{|\mathbb{Y}|} \mathcal{N}_{test,i}}. \quad (2)$$

Fig. 2 shows the comparison results of Clu.Acc of convolution layers in different convolution layers of models trained by PGD-AT, TRADES, and MART. The results of other deeper networks on larger datasets are shown in Appendix. We find that the deeper the layer is, the higher Clu.Acc of intermediate-layer representations, until be comparable to the classification accuracy. That is, the capacity of extracted features gradually increases with the deepening of network layers for a given trained model. In addition, this phenomenon indicates that, in the sense of

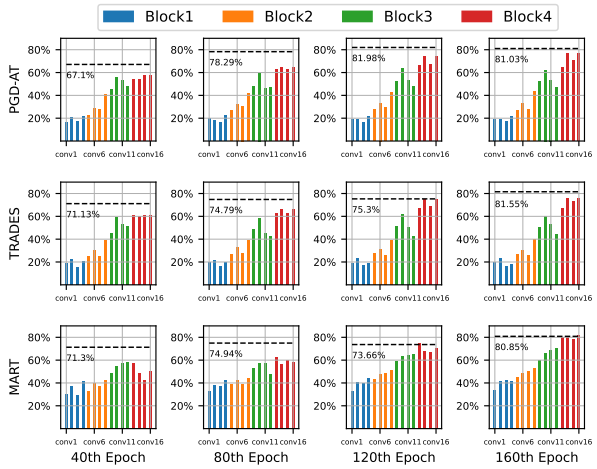

Figure 2. The figure shows the Clu.Acc for convolution layers of ResNet-18 trained by PGD-AT, TRADES, and MART at different epochs on CIFAR10 testset. The color of each column represents the residual block located. The black dashed line indicates the classification accuracy of the model at the present epoch.

Euclidean distance, the trained model performs similar encoding for examples sampled from the trained distribution with the same label. We term this phenomenon the ***Clustering Effect*** of intermediate layers. In the next section, by utilizing the property Clustering Effect, we manifest the distinction between OLA and ILA to explain the extra vulnerability of AT-trained models to ILA.

3.3. Connection to Information Bottleneck

Definition 2 *Given the input variable X and its corresponding ground-true label variable Y, the loss function $\mathcal{L}_{IB}$ of the* Information Bottleneck *principle on the neural network f is defined as:*

$$\mathcal{L}_{IB} = I(X;T) - \gamma I(T;Y), \gamma > 0, \quad (3)$$

where T is the output of any intermediate-layer representation of X, and the function $I(;)$ represents the mutual information between the two input variables.

Minimizing Equation (3) aims to squeeze the mutual information between the sample X and the intermediate-layer representation T during the forward propagation, and conversely retain the critical information regarding the ground truth label Y. That is, training models to throw features of little significance away. We reformulate the *Information Bottleneck* loss function [19, 20] referring to the correlation between mutual information and entropy, *i.e.*, $I(X;T) = H(T) - H(T|X)$ and $I(T;Y) = H(T) - H(T|Y)$, as follows,

$$\mathcal{L}_{IB} = (1-\gamma)H(T) + \gamma H(T|Y), \gamma > 0. \quad (4)$$

Note that T is fully determined by X when the parameters of model are fixed, that is, $H(T|X) = 0$. We discover two potential minimal solutions of the *information bottleneck* loss function determined by the value of γ, including,
Minimum 1 P_T is a Dirac distribution.
Minimum 2 $P_{T|Y=i}$ for any $1 \leq i \leq |\mathbb{Y}|$ is a Dirac distribution and P_T is a discrete probability distribution.

Apparently, for any $\gamma > 0$, $H(T|Y)$ would verge to 0 but $H(T)$ displays different states in the light of the range of γ. Therefore, we divide the value of γ into three regions, *i.e.*, $0 < \gamma < 1$, $\gamma = 1$, $1 < \gamma$, to further discuss when the *Information Bottleneck* loss function would converge to which minimum solution.
Case 1. For $0 < \gamma < 1$, the minimization of $\mathcal{L}_{IB}$ aims to minimize both $H(T)$ and $H(T|Y)$. The lower of $H(T|Y)$ represents the lower uncertainty of the conditional probability distribution $P_{T|Y=i}$, $1 \leq i \leq |\mathbb{Y}|$, *i.e.*, $P_{T|Y=i}$ is more like a Dirac distribution than uniform distribution, where we analyze $H(T)$ the same. Hence, the minimum of $\mathcal{L}_{IB}$ is only **Minimum 1**.
Case 2. The minimization of $\mathcal{L}_{IB}$ is equivalent to minimizing $H(T|Y)$, thus, the minimums of $\mathcal{L}_{IB}$ are both **Minimum 1** and **Minimum 2**.
Case 3. For $\gamma > 1$, in contrast to Case 1, Minimizing $\mathcal{L}_{IB}$ would maximize $H(T|Y)$. The higher $H(T)$ represents the probability distribution P_T is more uniform. Therefore, the minimum of $\mathcal{L}_{IB}$ is only **Minimum 2**.

When $\mathcal{L}_{IB}$ converges to the **Minimum 2**, the intermediate-layer representations of samples from different labels will be located in different tiny regions, and samples with the same label will be mapped in the same area. This suggests that the model can distinguish samples from different labels clearly. Therefore, this minimum fits the ground-true experimental results shown in Fig. 2. Conversely, the undesired **Minimum 1** implies that the model will map samples similarly at the intermediate layer whatever their labels, which means the model can not distinguish samples with different labels, is conflicted with experimental results shown in Fig. 2.

Proposition 1 *If T is a K-dimension random variable with finite mean vector μ and covariance matrix Σ, then the maximum entropy distribution of T is $\mathcal{N}(\mu, \Sigma)$.*

Proposition 2 *If T is a K-dimension Gaussian random variable with mean vector μ and finite covariance matrix Σ, then $Klog2\pi + \frac{1}{2}\sum_{i=1}^{K}\Sigma_{ii}$ is an upper bound of $H(T)$.*

Theorem 1 *Given a series of continuous K-dimension probability density distributions $\{p_i(t)|1 \leq i \leq N\}$ with their corresponding finite covariance matrices $\{\Sigma^i|1 \leq i \leq N\}$ and mean vectors $\{\mu^i|1 \leq i \leq N\}$, with $\lambda > 0$, the following two minimization problems (5) and (6) have the same optimal solution:*

$$\min_{p_i(t),1\leq i\leq N} -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\|\mu^i - \mu^j\|_2 + \lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_k^i, \quad (5)$$

$$\min_{p_i(t),1\leq i\leq N} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\int p_i(t)p_j(t)dt + \lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_k^i, \quad (6)$$

Theorem 1 proves that the formation of the Clustering Effect (Eq. 5) we investigate is equivalent to an upper bound of $\mathcal{L}_{IB}$ (Eq. 6, obtained by Proposition 1 and 2). Researches [19] point out that the training of complicated neural network doesn't obey the minimization of $\mathcal{L}_{IB}$ strictly. However, if we relax the constraint that the training obeys the minimization of an upper bound of $\mathcal{L}_{IB}$, the existence of the Clustering Effect and *Information Bottleneck* can be mutually verified. The detailed proof of the theorem is delegated in Appendix.

4. Problem Formulation

To uncover the mystery that the vulnerability of AT-trained "robust" models to ILA. We are concerned about the different influences of OLA and ILA acting on intermediate layers of AT-trained models. We connect our cognition of the Clustering Effect with robustness by finding the markedly different behavior between OLA and ILA, where ILA damages the Clu.Acc of layers of AT-trained models less severe than ILA but results in lower classification accuracy. The results show that the Clustering Effect we investigate is suitable for OLA, but ILA enhances our presumption that perturbations generated by ILA locate in some blind zone of AT and trained insufficiently by the model. Inspired by the results above, we compare the shift of intermediate layers caused by OLA and ILA to constitute our motivation for expanding the searching space during adversarial training. We observe ILA causes smaller distortions at intermediate layers than OLA in the sense of 2-norm, which is regarded as our motivation.
OLA v.s. ILA on Clustering Effect. We evaluate Clu.Acc of different intermediate layers of AT-trained models under OLA and ILA. We choose commonly used PGD_{100} and LAFEAT_{100} to represent OLA and ILA, respectively. As Fig. 3 shows, ILA performs less violent damage on clustering accuracy than OLA through all residual blocks but results in lower classification accuracy. Besides, opposite to OLA, Clu.Acc under ILA does not close to the classification accuracy, indicating the Clustering Effect we investigate is not suitable for ILA. This phenomenon confirms our hypothesis that the model does not be trained sufficiently on perturbations generated by ILA.

From a more intuitive perspective to explain the results in Fig. 3, ILA resembles sampling different potential parameters of trained models and generates adversarial perturbations to undermine the performance of all of them. There-

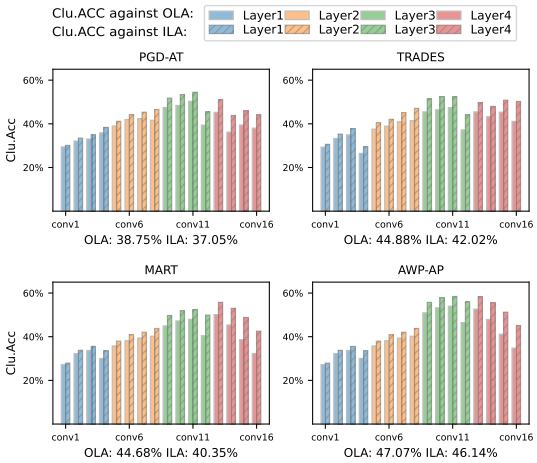

Figure 3. Clu.Acc at intermediate layers caused by OLA and ILA. The result is derived from AT-trained ResNet-18 on CIFAR10 testset. The color of each column represents the residual block the layer locates.

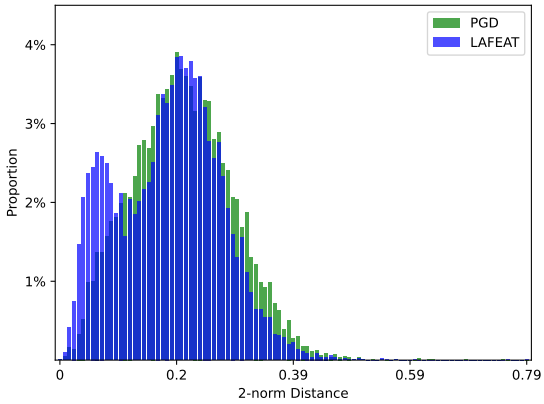

Figure 4. The distribution of the size of the shift at intermediate layers caused by OLA and ILA. The result is derived from the last convolution layer of ResNet-18 trained by PGD-AT on CIFAR10 testset.

fore, it's hard to generate ILA perturbations during training since the model developer can only hold the present parameters. As the matter of fact, the results are understandable that ILA perturbations probably locate in some blind zones of the inner maximization of AT. These blind zones deeply hide the imperceptible weakness of AT-trained "robust" models.

OLA v.s. ILA on the shift of intermediate layers. A reasonable way to exploit the space of ILA perturbations is by substituting a more general loss function for the Cross-Entropy loss in the inner maximization of AT. Inspired by this point of view, we make a profound study of the concrete difference between OLA and ILA, the size of the shift at the intermediate layers. We choose PGD_{100} and LAFEAT_{100} to represent OLA and ILA as before. As Fig. 4 shows, compared to OLA, ILA performs smaller distortion to the intermediate layer on average. The results of other layers are shown in Appendix. Therefore, we can construct the perturbation space to cover both the space of OLA and ILA by controlling the shift at intermediate layers. The perturbation will converge to OLA without the constraint of the shift at intermediate layers, conversely, the perturbation will close to ILA if paid large concentration to decrease the size of the shift. We set this notion as our insight to construct a large enough space covering that of ILA and OLA.

5. Proposed Method

5.1. Generalized Inner Maximization

The insights gained above prompt us to make an attempt to produce a more robust neural network which is also robust to ILA. To do this, we design a generalized loss function $\mathcal{L}_{Gen}$ as Equation (7) substituting Cross-Entropy loss $\mathcal{L}_{CE}$ in Equation (1) to quantify the shifting capacity of generated perturbations to the intermediate layers. For the layer l of the neural network $f_{\boldsymbol{\theta}}^{L+1}$ and any $x \in \mathcal{X}_{train,i}$, we utilize $2-$norm to measure the size of the shift between $f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x}+\boldsymbol{r})$ and $f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})$.

$$\mathcal{L}_{Gen}(f_{\boldsymbol{\theta}}^{l}, \boldsymbol{x}, \boldsymbol{r}) = \mathcal{L}_{CE} - \frac{\lambda}{n}\sum_{l=1}^{n} \|f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x}+\boldsymbol{r}) - f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})\|_2 \tag{7}$$

Apparently, $\mathcal{L}_{Gen} = \mathcal{L}_{CE}$ when $\lambda = 0$. The second item comes into effect with $\lambda > 0$ and enforces the perturbation to impact the output by intermediate layers as small as possible. Therefore, the perturbation we compute approximates ILA perturbation to some extent. Moreover, the perturbation generated by maximizing the Equation (7) may deviate the searching space of OLA, we set a random variable Λ subject to uniform distribution $\mathcal{U}(0, \lambda)$ substitute λ in Equation (7). Under the random variable Λ, our searching space can cover both OLA and ILA during training, enhancing the sufficiency of AT.

5.2. Outer Minimization

Standard AT only minimizes the Cross-Entropy loss $\mathcal{L}_{ce}$ but ignores the impact of the size of the searching space, which can increase the difficulty of searching the optimal. The generalized loss $\mathcal{L}_{Gen}$ provides extended searching space that amplifies the difficulty of searching the optimal. To reduce the difficulty of sufficiently generating perturbations from $\mathcal{L}_{Gen}$, we strive for squeezing the range of the whole searching space including ILA and OLA during the outer minimization by selecting proper parameter $\boldsymbol{\theta}$. We add a regularization term $\mathcal{L}_{Cluster}$ as Equation (8) to

squeeze the size of the shift at intermediate layers induced by perturbations.

$$\mathcal{L}_{Cluster,l}\left(f_{\boldsymbol{\theta}}^{l}, \boldsymbol{x}, \mu^{l}\right)=\frac{1}{\|\mathbb{Y}\|} \sum_{i=1}^{|\mathbb{Y}|} \sum_{k=1}^{d_{l}} \sqrt{\Sigma_{l}}_{kk}^{i}, \tag{8}$$

where $\sqrt{\Sigma_{l}}^{i}=E_{\mathcal{X}_{train,i}}[\sqrt{(f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})-\boldsymbol{\mu_{l}^{i}})(f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})-\boldsymbol{\mu_{l}^{i}})^{T}}]$, $\sqrt{\Sigma_{l}}_{kk}^{i}$ is the kth diagonal element of $\sqrt{\Sigma_{l}}^{i}$, the mean vectors $\boldsymbol{\mu_{i}^{l}}=E_{\mathcal{X}_{train,i}}[f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})]$. Then $\mathcal{L}_{Cluster,l}$ measures the ability of clustering of the $l-$th intermediate layer. We approximate the ground-true centroid of a label at the l-th layer by averaging the intermediate representation of samples in the batch from that label. The proposed $\mathcal{L}_{Cluster,l}$ contributes to pulling the intermediate representation of samples under perturbations to the centroid, decreasing the influence of perturbations at intermediate layers. Therefore, ILA and OLA will behave similarly at intermediate layers, the parameter of the model will promote the searching space of ILA and OLA to merge into a single, which decreases the range of the searching space.

5.3. Sufficient Adversarial Training

Based on the proposed generalized loss function $\mathcal{L}_{Gen}$ and clustering loss function $\mathcal{L}_{Cluster,l}$, we derive ***sufficient adversarial training (SAT)*** as a general form of conventional AT. SAT can be iteratively represented by the following two optimization problems Equation (9) and (10),

$$\boldsymbol{r_{*}}=\arg\max_{\boldsymbol{r}} \mathcal{L}_{CE}-\frac{\Lambda}{n} \sum_{l=1}^{n}\left\|f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x}+\boldsymbol{r})-f_{\boldsymbol{\theta}}^{l}(\boldsymbol{x})\right\|_{2} \tag{9}$$

$$\boldsymbol{\theta}^{*}=\arg\min_{\boldsymbol{\theta}} \mathcal{L}_{CE}+\frac{\beta}{n\|\mathbb{Y}\|} \sum_{i=1}^{|\mathbb{Y}|} \sum_{l=1}^{n} \sum_{k=1}^{d_{l}} \sqrt{\Sigma_{l}}_{kk}^{i}, \tag{10}$$

Theorem 2 *Given an $(L+1)$-layer neural network $f_{\boldsymbol{\theta}}^{L+1}$ with activation function Relu, and an input $\boldsymbol{x}$ with the p-norm constrained perturbation r, $\|r\|_{p} \leq \varepsilon$. Assume that the network classifies x as label y, $1 \leq y \leq|\mathbb{Y}|$, then if the inequality below holds,*

$$\Delta \leq \min_{\boldsymbol{r}}\left\{\left\|\left(\operatorname*{argmin}_{\boldsymbol{r}} \operatorname{ReLu}\left(\min_{j \neq y} \frac{f_{\boldsymbol{\theta}}^{L+1,(y)}(\boldsymbol{x})-f_{\boldsymbol{\theta}}^{L+1,(j)}(\boldsymbol{x})}{\left\|W_{L+1}^{y}-W_{L+1}^{j}\right\|_{u}}\right.\right.\right.$$
$$\left.\left.\left.-\left\|f_{\boldsymbol{\theta}}^{L+1}(\boldsymbol{x}+\boldsymbol{r})-f_{\boldsymbol{\theta}}^{L+1}(\boldsymbol{x})\right\|_{v}\right)\right)\right\|_{p}, \varepsilon\Bigg\}, \tag{11}$$

where $1=\frac{1}{u}+\frac{1}{v}$, and the classification labels of the set $\left\{\boldsymbol{x}+\boldsymbol{r} \middle\| \|\boldsymbol{r}\|_{p} \leq \Delta\right\}$ will be the same.

Similar to the bounds in [10], we further prove a proved bound in Theorem 2. The detailed proof is delegated to Appendix due to space limitations. Specifically, given training sample x with the ground truth label y, according to the Triangle Inequality, we have the inequality $\|f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{r})-f_{\boldsymbol{\theta}}(\boldsymbol{x})\|_{p} \leq\|f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{r})-\boldsymbol{\mu_{y}^{L}}\|_{p}+\|f_{\boldsymbol{\theta}}(\boldsymbol{x})-\boldsymbol{\mu_{y}^{L}}\|_{p}$. We simulate the perturbation r in both two augmentation strategies in SAT. Therefore, the proved bound Δ is related to $\mathcal{L}_{Cluster,l}$, where $\|f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{r})-f_{\boldsymbol{\theta}}(\boldsymbol{x})\|_{p}$ diminishes along with minimizing $\mathcal{L}_{Cluster,l}$, *i.e.*, $\sum_{i=1}^{\|\mathbb{Y}\|} \sum_{k=1}^{d_{l}} \sqrt{\Sigma_{l}}_{kk}^{i}$. The results of empirical evaluation of the proved bound Δ are shown in Appendix.

6. Experiments

6.1. Experiment Setup

Models and Datasets. For a fair comparison against existing defense techniques across different attacks, we evaluate our proposed SAT on three baseline datasets, *i.e.*, CIFAR10, SVHN, and CIFAR100. Worthy noting that we do not use any data augmentation method through experiments. We choose commonly used ResNet-18 and WideResNet28×10 for these three-channel datasets.

Baseline Attack & Defense Methods. For any model and dataset, we select FGSM, BIM_{100}, PGD_{100}, CW_{∞}, AutoAttack_{100}, and LAFEAT_{100} as baseline attacks. For all datasets, we set PGD as $\varepsilon=8/255$, $\alpha=2/255$ with 7 iterations during the inner maximization of PGD-AT. Besides, we compare SAT with the most valuable defense methods PGD-AT [17], TRADES [29], MART [22], and AWP-AT[25]. Further, we combine SAT with TRADES and AWP by rectifying $\mathcal{L}_{CE}$ in SAT and adding perturbations on weight, named SAT-TRADES and SAT-AWP-AT. We implement these AT methods following the hyperparameters settings of original literature.

Hyperparameter Settings. We involve layers in the last residual block of models in Equation (7) and (8). $\Lambda \sim \mathcal{U}(0, \lambda)$ is the most critical hyperparameter in SAT. We first choose λ=0 and update λ greedily with stepsize 0.01. We derive distributions in Fig. 4 induced by OLA, ILA, and proposed Equation (7) from the model trained by PGD-AT. Then vectorize the distributions to compute the inner product among them. If the inner product of distribution induced by OLA and Equation (7) is larger than that of ILA and Equation (7), we reduce λ with the fixed stepsize. Otherwise, we increase λ with the stepsize.

For all datasets, we train ResNet-18 and WRN28×10 for 200 epochs in SAT by SGD with the initial learning rate of 0.1. For each model and dataset setting, we add SGD with momentum 0.9, weight decay 5×10^{-4} and the decayed epochs are 100 and 150. The iteration of the inner maximization in SAT is 10.

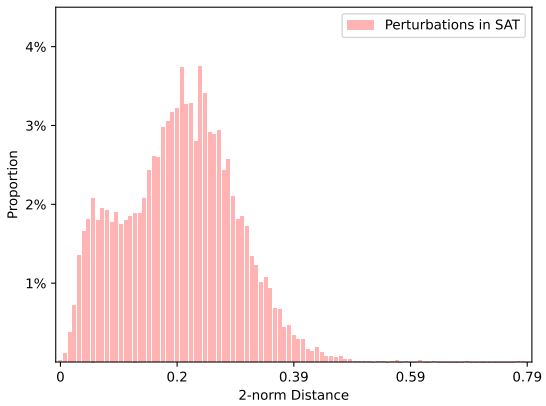

Figure 5. The distribution of the size of the shift at intermediate layers caused by Equation (7). We evaluate the size of the distortion at the last intermediate layer over the whole testset.

6.2. Experiments Results

Perturbations in SAT. After the searching of λ, we set $\Lambda \sim \mathcal{U}(0, \lambda)$ and generate perturbations firstly. As shown in Fig. 5, the size of the shift at intermediate layers caused by perturbations in SAT resembles the average of that of ILA and OLA in Fig. 4. Therefore, the result demonstrates that the searching space of the perturbation generated by maximizing Equation (7) has the ability to cover both OLA and ILA. The result of WRN28×10 and other datasets are shown in Appendix.

Performance Evaluation. Table 1 describes the comparison of the adversarial robustness of neural networks trained by baseline robust training methods [17, 29, 22, 25] and trained by the proposed SAT and its variants. The results show that SAT framework achieves better adversarial robustness especially against ILA than those state-of-the-art defense methods with comparable clean accuracy. The results of SVHN and CIFAR100 datasets with other ILAs [11, 16] are delegated to Appendix. The results indicate that extending the searching space of perturbations during training is helpful to improve the adversarial robustness of neural networks against both OLA and ILA, which experimentally demonstrates our earlier conjecture in the article.

Defense Against Converged Attack. Here we present the results of SAT against baseline attacks in very large iterations where attacks are converged. We set PGD_{1000}, AutoAttack_{1000}, and LAFEAT_{1000} as baseline converged attacks. As shown in Table 2, the robustness accuracy of SAT varies a little compared to the results in Table 1. The results demonstrate that the perturbation generated by Equation (7) can represent both ILA and OLA perturbations sufficiently.

Enhanced Clustering Effect. Similar to Fig. 4, for ResNet-18 trained by SAT, we evaluate the distribution of the distortion at the last intermediate layer induced by ILA and OLA as shown in Fig. 6. We observe that ILA

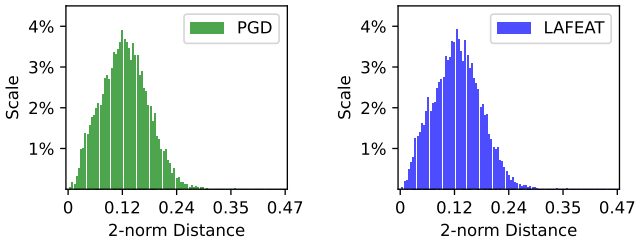

Figure 6. The size of the distortion at the last intermediate layer induced by ILA and OLA of ResNet-18 trianed by SAT on CIFAR10. We select PGD_{100} and LAFEAT_{100} to represent OLA and ILA as before.

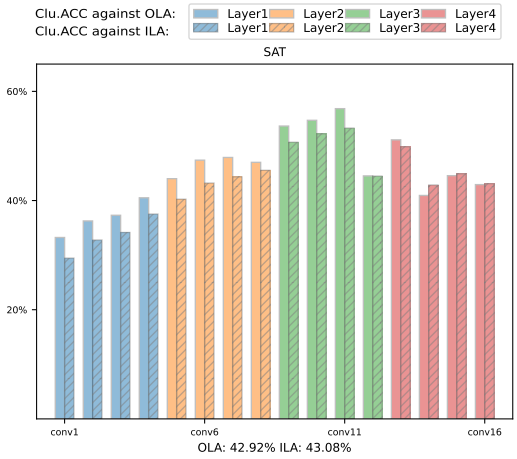

Figure 7. Clu.Acc at intermediate layers caused by OLA and ILA of ResNet-18 trained by SAT on CIFAR10 testset.

and OLA behave more similarly for SAT-trained ResNet-18 compared to Fig. 4 of the AT-trained model. The results demonstrate our proposed merging the searching space of ILA and OLA as a single, explaining the remarkable robustness against ILA of SAT and its variants as shown in Table 1. Besides, we additionally verify Clu.Acc of ResNet-18 trained by SAT on CIFAR10 dataset under PGD_{100} and LAFEAT_{100} as shown in Fig. 7. Compared to Fig. 3, we observe the model trained by SAT manifests favourable Clustering Effect on the last intermediate layer, demonstrating the model trains perturbations from Equation (7) sufficiently.

Performance on Generated Samples. Recent work [23] reports remarkable performance of TRADES on large amounts of generated data, showing that an "exhausted" manner is effective for improving robustness. Intuitively, more data can supply summits of loss function on points that are not sampled by original training set. Table 3 reports the comparison between SAT and trades, both are given 100K to 1M augmented data generated by the denoising diffusion probabilistic model. Both defenses have better performance with more augmented data, while SAT offers better robustness than TRADES, especially against

Dataset	Model	Method	Clean	FGSM	BIM_{100}	PGD_{100}	CW_∞	$AutoAttack_{100}$	$LAFEAT_{100}$
CIFAR10	ResNet-18	PGD-AT	81.32	60.83	40.19	38.75	38.68	38.02	37.05
		TRADES	82.45	62.94	45.33	44.88	45.46	43.38	42.02
		MART	81.07	63.36	45.72	44.68	46.67	44.36	40.14
		AWP-AT	**82.97**	63.15	48.94	47.07	47.57	47.36	46.14
		SAT	81.77	61.03	42.83	42.92	43.46	42.52	43.08
		SAT-TRADES	82.25	**63.40**	45.83	45.80	46.25	43.69	45.35
		SAT-AWP-AT	82.68	63.34	**49.08**	**48.82**	**48.07**	**47.64**	**48.95**
	WRN28×10	PGD-AT	83.52	65.21	44.81	44.72	44.12	44.24	41.71
		TRADES	84.25	66.88	49.74	49.92	49.15	47.57	45.31
		MART	**85.17**	66.52	52.41	52.03	50.95	49.34	45.90
		AWP-AT	84.16	67.54	54.38	54.36	52.47	50.14	48.61
		SAT	84.22	65.52	44.18	44.95	44.43	44.10	43.75
		SAT-TRADES	84.90	67.56	50.56	50.40	50.48	48.28	49.35
		SAT-AWP-AT	84.50	**68.49**	**54.79**	**54.60**	**53.55**	**51.44**	**54.70**

Table 1. Comparison of clean accuracy and robust accuracy against baseline attacks of neural networks across different defense mechanism, *i.e.*, PGD-AT [17], TRADES [29], MART [22], AWP-AT[25], and SAT. The bold indicates the best accuracy of the model under different attacks (%).

Dataset	Model	PGD_{1000}	$AutoAttack_{1000}$	$LAFEAT_{1000}$
CIFAR10	ResNet-18	42.28	42.06	42.49
	WRN28×10	44.80	44.05	44.55
CIFAR100	ResNet-18	31.35	30.04	31.69

Table 2. The robust accuracy of SAT against converged baseline attacks (%).

Generated	TRADES		SAT-TRADES	
	AutoAttack(OLA)	LAFEAT(ILA)	AutoAttack(OLA)	LAFEAT(ILA)
100K	53.27	50.85	53.36	54.60
500K	62.82	61.39	62.51	62.64
1M	63.46	62.88	63.56	64.17

Table 3. Comparison of SAT-TRADES and TRADES on generated samples(%).

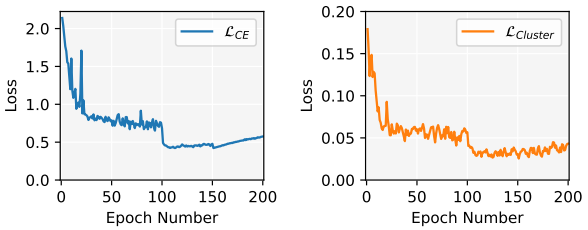

Figure 8. The evolution of $\mathcal{L}_{Cluster}$ and $\mathcal{L}_{CE}$ of ResNet-18 on SVHN.

ILA. Worthy noting that our proposed SAT is orthogonal to this exhausted manner.

The Convergence of SAT. To evaluate the convergence of the added regularization term $\mathcal{L}_{Cluster}$ in the minimization of SAT, we record the variation of $\mathcal{L}_{Cluster}$ during training. We present the transmission of $\mathcal{L}_{Cluster}$ in Fig. 8, which converges rapidly during training, and does not affect the convergence of $\mathcal{L}_{CE}$.

7. Conclusion

We observe and term the Clustering Effect in the forward propagation process and put down the weak robustness of model to the poor clustering robustness of intermediate layers attacks. Further, we theoretically connect the *Information Bottleneck* theory to prove the existence of the Clustering Effect. The result indicates the perturbation searching space of AT does not overlap with that of ILA. Besides, we propose SAT to explicitly extend the searching space of AT to further enhance the adversarial robustness. In addition, we strictly prove a robustness bound. The experiments show the superiority of SAT in improving the adversarial robustness of the output layer as well as the intermediate layers of the neural network.

Acknowledgment This work is supported by the National Natural Science Foundation of China (Nos. 62102300, 62206207, 61960206014, and 62121001).

References

[1] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. *arXiv preprint arXiv:2103.08307*, 2021.

[2] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pages 1014–1023. PMLR, 2020.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[4] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015.

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

[6] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

[7] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30, 2017.

[11] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019.

[12] Ahmadreza Jeddi, Mohammad Javad Shafiee, Michelle Karg, Christian Scharfenberger, and Alexander Wong. Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1241–1250, 2020.

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[14] Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. Integer-arithmetic-only certified robustness for quantized neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7828–7837, 2021.

[15] Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, and Willy Susilo. Erm-ktp: Knowledge-level machine unlearning via knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20147–20155, 2023.

[16] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2022.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[18] Geon Yeong Park and Sang Wan Lee. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7758–7767, 2021.

[19] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

[20] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[21] Pengfei Tang, Wenjie Wang, Jian Lou, and Li Xiong. Generating adversarial examples with distance constrained adversarial imitation networks. *IEEE Transactions on Dependable and Secure Computing*, 19(6):4145–4155, 2021.

[22] 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*, 2020.

[23] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023.

[24] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[25] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[26] Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *International Conference on Machine Learning*, pages 11693–11703. PMLR, 2021.

[27] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. Lafeat: Piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5735–5745, 2021.

[28] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[29] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[30] Xiaoyu Zhang, Chao Chen, Yi Xie, Xiaofeng Chen, Jun Zhang, and Yang Xiang. A survey on privacy inference attacks and defenses in cloud-based deep neural network. *Computer Standards & Interfaces*, 83:103672, 2023.

[31] Xiaoyu Zhang, Xiaofeng Chen, Joseph K Liu, and Yang Xiang. Deeppar and deepdpa: privacy preserving and asynchronous deep learning for industrial iot. *IEEE Transactions on Industrial Informatics*, 16(3):2081–2090, 2019.

[32] Xiaoyu Zhang, Yulin Jin, Tao Wang, Jian Lou, and Xiaofeng Chen. Purifier: Plug-and-play backdoor mitigation for pre-trained models via anomaly activation suppression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4291–4299, 2022.