InfoAT: Improving Adversarial Training Using the Information Bottleneck Principle

Mengting Xu, Tao Zhang, Zhongnian Li, and Daoqiang Zhang, Senior Member, IEEE

Abstract—Adversarial training (AT) has shown excellent high performance in defending against adversarial examples. Recent studies demonstrate that examples are not equally important to the final robustness of models during AT, that is, the so-called hard examples that can be attacked easily exhibit more influence than robust examples on the final robustness. Therefore, guaranteeing the robustness of hard examples is crucial for improving the final robustness of the model. However, defining effective heuristics to search for hard examples is still difficult. In this article, inspired by the information bottleneck (IB) principle, we uncover that an example with high mutual information of the input and its associated latent representation is more likely to be attacked. Based on this observation, we propose a novel and effective adversarial training method (InfoAT). InfoAT is encouraged to find examples with high mutual information and exploit them efficiently to improve the final robustness of models. Experimental results show that InfoAT achieves the best robustness among different datasets and models in comparison with several state-of-the-art methods.

Index Terms—Adversarial training (AT), information bottleneck (IB), mutual information, robustness.

I. INTRODUCTION

Despite the great success of deep neural networks on diverse tasks such as image classification [1], speech recognition, and natural language processing [2], recent studies have shown that their adversarial robustness is very poor [3], [4]. When attacked by adversarial examples which added visually imperceptible perturbations, deep neural networks can be easily misled by high confidence. The vulnerability of deep neural networks to adversarial examples may pose a huge threat to applications and services that are highly dependent on security performance, such as autonomous driving systems [5], [6], computer-aided disease diagnosis based on medical images [7], and malware detection protocols [8].

Considering the significance of adversarial robustness in neural networks, a range of defense methods have been proposed [7], [9]–[12]. Sun et al. [13] achieved a high level of robustness by projecting both clean and adversarially attacked input images into low-dimensional quasi-natural image space. Zhang et al. [14] proposed to progressively minimize the divergence of different distributions, such that the gap between the clean and adversarial domains in the intermediate feature space and decision space is minimum. Among them, adversarial training (AT), which continuously adds newly generated adversarial examples in the neural network training process, has been proved to be one of the most effective defense strategies against adversarial attacks. Many AT strategies employ the “worst-case” adversarial example to maximize the loss for updating the current model such as standard AT [16], TRADES [15], and robust self-training (RST) [17].
Zhang and Wang [18] generated adversarial images for AT through feature scattering in the latent space, which is unsupervised in nature and avoids label leaking. Liu et al. [19] and [20] designed a GanDef, which utilizes a competition game to regulate the feature selection during the AT. And [21] showed that adversarially training a network with an attack method helps defend against that particular attack method, but has limited effect on other attack methods. Furthermore, recent studies [22], [23] have shown examples along with their adversarial variants are not equally important to the robustness of models during AT. Specifically, the so-called hard examples (which can be attacked easily) exhibit more influence than robust examples on the final robustness. Therefore, it is crucial to define effective heuristics to search for hard examples for improving the robustness of the model.

In this article, we propose a novel and effective adversarial training strategy (InfoAT) from an information-theoretic perspective. It uses the information bottleneck (IB) principle [24] to find the hard examples. Recently, the studies of the IB principle to deep neural network training have developed rapidly [25], [26], but its effect on the effectiveness of AT still remains unclear. The IB principle shows that when the valid information of the input cannot be well compressed by the model, the neural network cannot generalize the out-of-distribution data well, that is, the mutual information between the input data and its related latent representation is high. As shown in Fig. 1(a), hard (i.e., nonrobust, can successfully be attacked by project gradient descent (PGD) [16] attack) examples are more likely to have higher mutual information than robust examples. What is more, Fig. 1(b) shows the larger the mutual information for example, the smaller the perturbation required for a successful attack. These phenomena show that an example with high mutual information is more vulnerable than other examples. This is because when the mutual information is high, the latent representation of the network has learned a large number of features of the input image. And these features often contain some redundant information and nonrobust features, so when there are adversarial perturbations in the image, the model will be easily fooled. Motivated by these conceptual and experimental observations, we adopt the mutual information of the original examples between the input and latent distribution with the current model to select the hard examples and guarantee them to be robust against attackers to improve the robustness of the model.

Our main contributions are:

1) We show that the example with high mutual information is more likely to be attacked in Fig. 1. High mutual information means that the latent representation of the network contains a lot of information about the input image, including the redundant information and adversarial perturbations that are useless for the prediction results, which makes the network more vulnerable in these examples;

2) We propose a novel and effective adversarial training (InfoAT) using the IB principle and address the problem of selecting a hard example with high mutual information. To the best of our knowledge, it has not been studied yet;

3) We show the effectiveness of InfoAT by different networks and datasets. Specifically, under the most common comparison setting (PGD attack with 20 steps on CIFAR-10, with \( L_\infty \) perturbation \( \epsilon = 8/255 \)), InfoAT achieves the state-of-the-art robustness of 60.46% with WideResNet, which improves ~7% over standard AT [16], ~6% over GanDef [19], ~4% over TRADES [15], and ~2% even over MART [22].

II. BACKGROUND AND RELATED WORK

A. Base Classifier

For a \( K \)-class \((K \geq 2)\) classification problem, denote a dataset \( \{(x_i, y_i)\}_{i=1}^n \) with distribution \( x_i \in \mathbb{R}^d \) as natural input and \( y_i \in \{1, \ldots, K\} \) represents its corresponding true label, a classifier \( h_\theta \) with parameter \( \theta \) predicts the class label of an input example \( x_i \)

\[
    h_\theta(x_i) = \arg \max_{k=1,\ldots,K} p_k(x_i, \theta)
\]

where

\[
    p_k(x_i, \theta) = \frac{\exp(z_k(x_i, \theta))}{\sum_{k=1}^K \exp(z_k(x_i, \theta))}
\]

where \( z_k(x_i, \theta) \) is the network logits output with respect to class \( k \), and \( p_k(x_i, \theta) \) is the probability (softmax on logits) that \( x_i \) belonging to class \( k \).

B. Adversarial Training

One of the most successful empirical defenses to date is AT. It was first proposed in [3] and [4], where they showed that adding adversarial examples to the training set can improve the robustness against attacks. More recently, [16] formulated AT as a min-max optimization problem and demonstrated that AT with PGD attack leads to empirical robust model. Specifically, the objective function of standard AT is

\[
    \mathcal{L}_{\text{AT}} = \min \frac{1}{n} \sum_{i=1}^n \ell(h_\theta(\tilde{x}_i^\prime), y_i)
\]

where

\[
    \tilde{x}_i^\prime = \arg \max_{x_i^\prime \in B_\mathcal{L}(x_i)} \ell(h_\theta(x_i^\prime), y_i).
\]

\( \ell() \) represents the loss function such as commonly used cross-entropy loss, and \( B_\mathcal{L}(x_i) = \{x_i^\prime : \|x_i^\prime - x_i\|_{\mathcal{L}} \leq \epsilon\} \) denotes the \( L_\mathcal{L} \)-norm ball centered at \( x_i \) with radius \( \epsilon \). In this work, we only focus on the \( L_\infty \)-norm.

Based on the primary AT framework, recent research has proposed many improvements from different perspectives and demonstrated strong performance. TRADES [15] trades adversarial robustness off against accuracy. The objective function of TRADES is a linear combination of natural loss and regularization terms. Its regularization term uses Kullback–Leibler (KL) divergence to constrain the output probability between the original example and its adversarial example. The objective function of TRADES can be described as follows:

\[
    \mathcal{L}_{\text{TRADES}} = \min \frac{1}{n} \sum_{i=1}^n \{\ell(h_\theta(x_i), y_i)
\]

\[
    + \lambda \cdot \text{KL}(p(x_i, \theta) || p(\tilde{x}_i^\prime, \theta))\}
\]
where
\[
\hat{x}_i^\prime = \arg \max_{x_i^\prime \in B(x_i)} \text{KL}(p(x_i, \theta)||p(x_i^\prime, \theta))
\]
measures the difference of two distributions. Furthermore, MART [22] gives the misclassified examples and correctly classified examples different weights during AT and adopts a regularized adversarial loss involving both adversarial and natural examples. The objective function of MART is as follows:

\[
\mathcal{L}_{\text{MART}} = \min 1 \cdot \sum_{i=1}^{n} \left\{ \text{BCE}(h_\theta(\hat{x}_i^\prime), y_i) + \lambda \cdot \text{KL}(p(x_i, \theta)||p(\hat{x}_i^\prime, \theta)) \cdot (1 - p_{y_i}(x_i, \theta)) \right\}
\]

C. Information Bottleneck (IB) Principle

Different from these methods, our method selects hard examples to efficiently improve the model robustness using the IB principle. The IB [24] is a principled method of finding a latent representation Z, which represents the information contained in the input variable X about the output Y. The mutual information of X and Z can be written as I(X; Z). I(X; Z) = KL(P_{XZ}/P_XP_Z), where P_{XZ} is the joint distribution and P_X, P_Z are the marginal distributions of input variables X, Z, respectively. Intuitively, I(X; Z) measures the uncertainty of X given Z. It not only reflects how much Z compresses X through I(X; Z), but also reflects how well Z predicts Y through I(Y; Z). In practice, this IB principle is achieved by minimizing the following IB Lagrangian:

\[
\min[I(X; Z) - \lambda \cdot I(Y; Z)]
\]

where \(\lambda\) represents a positive parameter, used to controls the tradeoff between compression and prediction. Through the compression item I(X; Z) to control the amount of compression in Z, we can adjust the required features of the training model, such as data augmentation [27], generalization error [26], and detection of out-of-distribution data [28].

In this article, we use the IB principle to select hard examples with high mutual information during AT. Specifically, we use I(X = x_i; Z = z_i) to denote the point-wise mutual information [29] of the input x_i and its latent representation z_i, which means the relevant information that z_i provides about x_i [24]. Hereafter, to avoid notational clutter, we will shorten “X = x_i” to x_i, “Z = z_i” to z_i, and “Y = y_i” to y_i in all conditionals and point-wise mutual information theoretic functions.

III. METHODS

A. Regularizations Using IB Principle

Note that the adversarial risk in (3) and (4) treat all examples equally, regardless of the different effects of these examples on the robustness of the model. Therefore, the main idea of our article is to apply the IB principle to AT to exploit hard examples efficiently so as to improve the robustness of models.

According to (10), when the natural input information cannot be well compressed by the model, the neural network cannot well generalize the out-of-distribution data, that is, the mutual information of the natural input and its related latent representation is very high, as confirmed in Fig. 1. Based on these observations, we consider adding the IB-weighted regularization of natural input to the standard cross-entropy loss. The IB-weighted regularization on natural input is defined as follows:

\[
\mathcal{L}_{\text{IB}}(x, \theta) = 1 \cdot \sum_{i=1}^{n} I(x_i; z_i) \cdot \mathbb{1}(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i))
\]

where \(\hat{x}_i^\prime\) is the “worst-case” adversarial examples with the \(\epsilon\)-ball of natural input \(x_i\), and I(\(x_i; z_i\)) is the point-wise mutual information of the input \(x_i\) and its latent representation \(z_i\), which was used to select the hard examples (this will be large for hard examples and small for robust examples). I(\(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i)\)) aims to encourage the output of neural network to be stable against example which has high mutual information. I(\(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i)\)) = 1 when \(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i)\), and I(\(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i)\)) = 0 otherwise.

Moreover, [30] disentangles adversarial examples as a natural consequence of nonrobust features. A common belief is that adversarial robustness comes from feature representations learned through adversarial examples [30], [31]. From this perspective, the network is expected to maintain more information on adversarial examples (robust features), that is, the mutual information of the adversarial input and its associated latent adversarial representation should be high. Therefore, the newly added adversarial regularizer on adversarial examples is defined as follows:

\[
\mathcal{L}_{\text{Reg}}(\hat{x}_i^\prime, \theta) = 1 \cdot \sum_{i=1}^{n} -I(\hat{x}_i^\prime; \hat{z}_i^\prime)
\]

where I(\(\hat{x}_i^\prime; \hat{z}_i^\prime\)) is the point-wise mutual information of the adversarial example \(\hat{x}_i^\prime\) and its latent representation \(\hat{z}_i^\prime\).

Finally, by combining the proposed IB-weighted regularization and adversarial regularization in the AT framework, we can train a network by proposed InfoAT that minimizes the following risk:

\[
\mathcal{L}_{\text{min}}(x, \theta) = \min_{\theta} 1 \cdot \sum_{i=1}^{n} \left\{ \ell(h_\theta(\hat{x}_i^\prime), y_i) + \lambda \cdot I(x_i; z_i) \cdot \mathbb{1}(h_\theta(\hat{x}_i^\prime) \neq h_\theta(x_i)) - \beta \cdot I(\hat{x}_i^\prime; \hat{z}_i^\prime) \right\}
\]

where \(\lambda, \beta\) are tunable scaling parameters that are fixed for all training examples, \(\ell(\cdot)\) is the standard cross-entropy loss. Next, we will describe how to generate the “worst-case” adversarial example \(\hat{x}_i^\prime\) in our InfoAT.
Algorithm 1 Adversarial Training Using the Information Bottleneck Principle (InfoAT)

**Input:** \( \{x_i, y_i\}_{i=1-n} \): training data; \( T_0 \): outer iteration number; 
\( T_I \): inner iteration step; \( \epsilon \): maximum perturbation; \( a_l \): step size for inner optimization; \( a_o \): step size for outer optimization; 
\( m \): the number of each batch size; \( \lambda \) and \( \beta \): tunable scaling parameters.

**Initialization:** Standard random initialization of \( h_\theta \)

1: for \( t = 1, \ldots, T_O \)  
2: Uniformly sample a minibatch of training data \( B(t) \)  
3: for \( x_i \in B(t) \)  
4: \( x_i' = x_i + \epsilon \cdot \zeta \), with \( \zeta \sim U(-1, 1) \)  
5: \( \# B(t) \) is a uniform distribution  
6: for \( s = 1, \ldots, T_I \)  
7: \( x_i' \leftarrow x_i + \alpha_I \cdot \text{sign}(\nabla_x \mathbb{P}(\theta(x_i'), y_i)) + \lambda \cdot \mathbb{P}(\theta(x_i')) \cdot \mathbb{I}(\mathbb{P}(\theta(x_i')) - \mathbb{P}(x_i))^2 \)  
8: \( \# B(t) \) is the projection operator  
9: endfor  
10: \( \hat{x}_i' \leftarrow x_i' \)  
11: endfor  
12: \( \theta \leftarrow \theta - a_o \sum_{x_i \in B(t)} \nabla_\theta [-\beta \cdot \mathbb{P}(\theta(x_i')) + \mathbb{P}(\theta(x_i')) \cdot \mathbb{I}(\mathbb{P}(\theta(x_i')) - \mathbb{P}(x_i))^2] \)  
13: endfor

Standard AT essentially minimizes a lower bound on the inner maximum “worst-case” loss, which will mislead an optimizer to search for a local optimal value [32]. To alleviate this pessimism, we also consider incorporating IB regularizer into the maximization phase. In maximization phase, the model parameters \( \theta \) are fixed, and \( z_i \) is a deterministic representation of input \( x_i \), that is, any given input is mapped to a fixed representation. Consequently, it holds that \( I(x_i; z_i) \) is a different constant for different input \( x_i \). Therefore, the IB regularization can be used to search for more out-of-distribution adversarial data (i.e., the mutual information of its natural example is high) in the maximization phase. From this perspective, we extend the proposed IB regularization to the maximization phase to search for the “worst-case” (i.e., more out-of-distribution) data to improve model robustness. We have verified its effectiveness in experiments. The generation of \( \hat{x}_i' \) can be defined as

\[
\hat{x}_i' = \arg \max_{x_i \in B(x_i)} \left\{ \ell(h_\theta(x_i'), y_i) + \lambda \cdot I(x_i; z_i) \cdot \mathbb{I}(h_\theta(x_i') \neq h_\theta(x_i)) \right\}.
\]

**B. Optimization for InfoAT**

As analyzed in the previous section, we derived the adversarial risk of InfoAT based on point-wise mutual information [i.e., \( I(x_i; z_i), I(\hat{x}_i'; z_i') \)] and 0-1 loss [i.e., \( \mathbb{I}(h_\theta(x_i') \neq h_\theta(x_i)) \)]. However, mutual information in example level is shown to be intractable in the literature [33], [34] since \( z_i \) is high-dimensional and optimization over 0-1 loss is also difficult, and therefore directly optimizing this objective is challenging. We then propose the optimization strategies, by surrogating mutual information with entropy and replacing 0-1 loss with proper distance function, which are both physically meaningful and computationally tractable.

1) **Surrogating Mutual Information With Entropy:** A variety of methods have been proposed for estimating mutual information [35]–[37]. One of the most effective estimators is the mutual information neural estimator (MINE) [33]. It proves that, for all \( \eta \geq 0 \), there exists \( M \in \mathbb{Z} \) that satisfies certain constraints, such that for any \( m \geq M \), \( |I_m(X; Z) - I(X; Z)| \leq \eta \) almost always, where \( m \) is the number of sampling examples.

However, our work exploits mutual information of each example (instead of on \( m \geq M \) examples) to select the hard example. From this perspective, MINE may not be applicable in our case. On example level, the mutual information \( I(x_i; z_i) \), for example, \( x_i \) is still intractable. Our approach in this article is that when in the task of classification scenarios, \( I(x_i; z_i) \) can be effectively approximated during training. As we will show, this process can be effectively realized by the entropy \( H(\hat{y}_i) \) predicted by the network, which is a tractable lower bound of \( I(x_i; z_i) \), \( H(\cdot) \) is the Shannon entropy, and \( \hat{y}_i \) is the output prediction of current model \( h_\theta(\cdot) \).

As described in [26], a deep neural network can be viewed as a Markov chain of continuous representation of the input, where the information flow follows the following structure: \( x_i \mapsto z_i \mapsto \hat{y}_i \), where \( \hat{y}_i \) is network prediction. According to the data processing inequality [38], \( I(x_i; z_i) \geq I(x_i; \hat{y}_i) \) can be got, where \( I(x_i; \hat{y}_i) = H(\hat{y}_i) - H(\hat{y}_i|x_i) \). In other words, when calculating mutual information during the AT process, the model parameter \( \theta \) is fixed, and any given input \( x_i \) is mapped to a single class. That is, \( H(\hat{y}_i|x_i) = 0 \). After putting all these together, we have the following proposition.

**Proposition 1:** Consider a deterministic neural network, the parameter \( \theta \) of which is fixed. Given an input \( x_i \), a network prediction \( \hat{y}_i \), and a latent representation \( z_i \). Then, the entropy \( H(\hat{y}_i) \) is the lower bound of the mutual information \( I(x_i; z_i) \) that is,

\[
I(x_i; z_i) \geq I(x_i; \hat{y}_i) = H(\hat{y}_i) - H(\hat{y}_i|x_i) = H(\hat{y}_i). \tag{15}
\]

There are two important benefits of surrogating mutual information with entropy to be discussed. First, \( H(\hat{y}_i) \) is the lower bound of \( I(x_i; z_i) \), so it is the most representative term for choosing a hard example that is nonrobust to adversarial attack (i.e., a high entropy example indicates a high mutual information). Besides, \( H(\hat{y}_i) \) can be easily computed from the softmax output of a classification network. Therefore, for point-wise mutual information, we surrogate \( I(x_i; z_i) \) with \( H(p(x_i)), I(\hat{x}_i'; z_i') \) with \( H(p(x_i')) \).

2) **Replacing 0-1 Loss With Distance Function:** For the 0-1 indicator function \( \mathbb{I}(h_\theta(x_i') \neq h_\theta(x_i)) \), we use mean square error (MSE) divergence as the surrogate loss function [39], since \( h_\theta(x_i') \neq h_\theta(x_i) \) represents that the output distributions of the adversarial example and the natural example are different. Thus, we have

\[
\text{MSE}(x_i', x_i) = \| p(x_i') - p(x_i) \|_2^2. \tag{16}
\]

This will be large when the difference of distribution between adversarial examples and natural examples is high, and small otherwise.
Fig. 2. Sensitivities to parameters $\lambda$ and $\beta$. The units of “Test robustness” is (%). The dataset is CIFAR-10. The model is ResNet-18. The attack method is PGD with 20 steps. (a) Choice of $\lambda$. (b) Choice of $\beta$.

| TABLE I | ROBUSTNESS (%) AND ACCURACY (%) BASED ON MINE ESTIMATION MUTUAL INFORMATION ON DIFFERENT PARAMETER $\lambda$, WHERE $\beta = 0.2$ |
|---------|-------------------------------------------------|
|         | $\lambda$ | Natural | PGD$^{20}$ |
| layer #1| 0.5       | 58.42   | 56.71      |
|         | 1.0       | 58.53   | 56.79      |
|         | 1.5       | 58.69   | 57.21      |
|         | 2.0       | 58.52   | 56.09      |
| layer #2| 1.0       | 58.43   | 57.02      |
|         | 1.5       | 58.74   | 56.88      |
|         | 2.0       | 58.79   | 57.29      |
|         | 2.5       | 58.36   | 57.73      |
|         | 3.0       | 58.30   | 57.46      |
| layer #3| 1.0       | 58.90   | 57.07      |
|         | 1.5       | 58.23   | 57.55      |
|         | 2.0       | 58.31   | 57.54      |
|         | 2.5       | 58.20   | 57.55      |
|         | 3.0       | 58.36   | 57.74      |
| layer #4| 1.0       | 58.15   | 57.39      |
|         | 1.5       | 58.46   | 57.76      |
|         | 2.0       | 58.92   | 57.67      |
|         | 2.5       | 58.62   | 57.46      |
|         | 3.0       | 58.90   | 57.40      |

C. Overall Objective Function

Based on the above surrogate loss functions, we can illustrate the final objective function of InfoAT we proposed

$$L_{\text{InfoAT}}(x; \theta) = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left( CE(p(x_i^t), y_i) + \lambda \cdot H(p(x_i)) \cdot ||p(x_i^t) - p(x_i)||^2_2 - \beta \cdot H(p(x_i^t)) \right)$$

where

$$x_i^t = \arg \max_{x_i \in B_i(x_i)} \left( CE(p(x_i^t), y_i) + \lambda \cdot H(p(x_i)) \cdot ||p(x_i^t) - p(x_i)||^2_2 \right)$$

are fixed for all training examples. Algorithm 1 describes the complete training procedure of InfoAT.

IV. EXPERIMENTS

In this section, first, we conduct a large number of ablation studies to fully understand our proposed InfoAT and then we evaluate its robustness on the benchmark datasets under the black-box and white-box settings. Finally, we use the WideResNet to benchmark the state-of-the-art robustness and use unlabeled data for further robustness evaluation of our method.

1) Experimental Setup: Three commonly used datasets (i.e., CIFAR-10, CIFAR-100, and SVHN) and two networks (i.e., ResNet18 and WideResNet-34-10) are utilized in our experiments. All the models with three datasets are trained using SGD, the momentum is 0.9, weight decay is $3 \times 10^{-3}$, and the initial learning rate is 0.01, which is divided by 10 at the 75th, 90th, and 100th epoch. The attack used for training is PGD$^{10}$ (PGD attack with ten steps) with a random start. The maximum perturbation $\epsilon = 8/255$ and step size $\epsilon/4$, while the test attack is PGD$^{10}$ with random start unless otherwise stated.

We perform all experiments on a desktop PC using a single GeForce RTX 2080 Ti GPU and 12-core Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz. We implement training and certification in PyTorch.

A. Ablation Study

Here, we first investigate the parameter $\lambda \in [0, 8]$ and $\beta \in [0.1, 2.0]$ in InfoAT objective function. We present the results of $\lambda$ and $\beta$ in Fig. 2(a) and (b), respectively. InfoAT achieves good robustness across different choices of $\lambda$ and $\beta$. We choose $\lambda = 2.5, \beta = 0.2$ for further experiments.

Then, we investigate InfoAT from four aspects: 1) evaluating the effectiveness of surrogating mutual information with entropy; 2) employing different weighting methods; 3) replacing divergence of InfoAT; and 4) exploring different regularizations in the outer minimization.

1) Effectiveness of Surrogating Mutual Information With Entropy: In this part, we evaluate the effectiveness of our method (i.e., surrogating mutual information with entropy) by comparing it with MINE. MINE is a method estimates the mutual information of $m \geq M$ examples. Therefore, we regularize each batch of $m$ examples with the same mutual information estimated by MINE, as follows:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{b} I(x_i; z_i) I(h_\theta(x_{ij}^{t}) \neq h_\theta(x_{ij}))$$

where $b$ is the number of batch, $m$ is the number of examples in each batch, and $I(x_i; z_i)$ represents the mutual information on batch $i$. Here, we show the results of model robustness and accuracy based on MINE estimation mutual information on different parameter $\lambda$ in Table I, where $\#1, \#2, \#3, \#4$ represent the mutual information estimated by MINE on different layer latent representations $z(l); l \in \{\#1, \#2, \#3, \#4\}$. The other settings are the same as the InfoAT.

The best model robustness of mutual information estimated by MINE and surrogated by entropy is shown in...
Fig. 3. Comprehensive ablation experiments of InfoAT. In (a), \(l \in \{#1, #2, #3, #4\}\) represent the mutual information estimated by MINE on different layer latent representations \(z(l)\), and \(#5\) represents the mutual information surrogated by entropy based on softmax output on the last layer \(#5\). In (b)–(d), the red line represents our InfoAT method. The dataset is CIFAR-10. The model is ResNet-18. The attack method is PGD with 20 steps. (a) Mutual information estimation. (b) Different weighting methods. (c) Different divergences. (d) Different regularizations.

| Dataset | Model   | Natural | FGSM   | PGD\(^{20}\) | PGD\(^{100}\) | PGD\(^+\) | \(\text{CW}\)\(^{20}\) | \(\text{CW}\)\(^{100}\) | BPDA | InfoAT\(^{20}\) | AA  |
|---------|---------|---------|--------|-------------|-------------|-----------|----------------|----------------|------|----------------|-----|
| CIFAR-10| GanDef  | 86.31   | 63.63  | 49.25       | 44.71       | 45.09     | 49.44         | 45.62         | 48.18 | 47.57          | 47.13|
|         | AT      | 83.56   | 63.21  | 51.47       | 49.54       | 49.71     | 51.23         | 49.51         | 52.91 | 52.29          | 47.25|
|         | TRADES  | 82.21   | 63.40  | 53.92       | 52.66       | 52.57     | 51.81         | 50.56         | 54.42 | 54.49          | 49.59|
|         | MART    | 79.24   | 63.74  | 56.01       | 54.79       | 54.62     | 51.12         | 49.76         | 56.07 | 56.63          | 48.45|
|         | MART+   | 79.52   | 64.19  | 56.69       | 55.22       | 55.10     | 51.37         | 50.01         | 55.67 | 56.17          | 48.50|
|         | InfoAT (Ours) | 83.23 | 67.23  | 57.95       | 56.15       | 56.14     | 52.91         | 51.17         | 56.84 | 57.08          | 49.67|
| CIFAR-100| GanDef | 65.15   | 38.52  | 27.67       | 25.86       | 25.70     | 26.38         | 25.24         | 26.60 | 26.31          | 22.03|
|         | AT      | 57.28   | 35.10  | 26.46       | 25.10       | 25.10     | 25.25         | 24.47         | 30.80 | 30.58          | 22.89|
|         | TRADES  | 55.27   | 35.32  | 28.89       | 28.01       | 27.80     | 25.88         | 25.26         | 30.97 | 31.13          | 24.56|
|         | MART    | 53.91   | 38.72  | 33.03       | 32.61       | 32.46     | 29.01         | 28.45         | 33.10 | 33.48          | 26.67|
|         | MART+   | 47.84   | 36.96  | 32.30       | 31.93       | 31.82     | 27.00         | 26.62         | 31.94 | 32.73          | 26.50|
|         | InfoAT (Ours) | 58.23 | 40.62  | 34.08       | 33.39       | 33.06     | 29.51         | 28.68         | 33.43 | 33.92          | 26.76|
| SVHN    | GanDef  | 93.25   | 73.89  | 56.86       | 52.74       | 52.76     | 55.19         | 51.84         | 55.53 | 58.96          | 47.00|
|         | AT      | 89.97   | 69.63  | 54.78       | 51.01       | 51.76     | 50.87         | 48.78         | 53.53 | 53.02          | 46.99|
|         | TRADES  | 89.26   | 71.43  | 58.01       | 55.87       | 55.51     | 54.77         | 52.97         | 57.03 | 60.74          | 50.82|
|         | MART    | 92.76   | 76.67  | 62.46       | 59.42       | 58.87     | 56.95         | 53.98         | 63.56 | 65.40          | 49.25|
|         | MART+   | 90.28   | 73.58  | 62.67       | 60.62       | 60.39     | 57.15         | 54.54         | 61.84 | 64.68          | 50.15|
|         | InfoAT (Ours) | 92.37 | 76.91  | 63.16       | 60.72       | 60.14     | 57.22         | 54.64         | 64.37 | 67.05          | 51.37|

From Fig. 3(b) we can observe, surrogating mutual information with entropy (i.e., red bar) leads to a significant
TABLE III
BLACK-BOX ROBUSTNESS [ACCURACY (%)] ON CIFAR-10 WITH RESNET18 UNDER SPSA ATTACK WITH DIFFERENT BATCH SIZES AND $\epsilon = 8/255$

| Model   | SPSA$_{128}$ | SPSA$_{256}$ | SPSA$_{512}$ | SPSA$_{1024}$ |
|---------|--------------|--------------|--------------|--------------|
| GanDef  | 56.70        | 53.81        | 51.65        | 50.08        |
| AT      | 58.23        | 55.77        | 53.90        | 52.74        |
| TRADES  | 58.69        | 56.48        | 54.82        | 53.60        |
| MART    | 57.41        | 55.50        | 53.84        | 52.52        |
| MART+   | 57.73        | 55.62        | 53.20        | 52.58        |
| InfoAT  (Ours) | 59.56 | 57.40 | 55.79 | 54.51 |

Fig. 4. Weight loss surface across different methods on CIFAR-10 with ResNet18.

In this part, we evaluate the robustness of InfoAT on CIFAR-10, CIFAR-100, and SVHN datasets with ResNet18 against various white-box and black-box attacks. We consider five types of representative AT methods here: 1) AT [16]; 2) TRADES [15]; 3) MART [22]; 4) MART+ (replacing 1-prob in MART [22] with the proposed mutual information weight); and 5) GanDef [19], following their original articles.

1) White-Box Robustness: The robustness of all defense models is evaluated against nine types of attacks: FGSM, PGD$^{20}$, PGD$^{100}$ (20-step/100-step PGD with step size $\epsilon/4$), PGD$^{+}$ (PGD$^{+}$ is PGD with five random restarts, and each restart has 40 steps with step size 0.01, which keeps the same as [17]), CW$_{\infty}$, CW$_{\infty}$ (L$_{\infty}$ version of CW optimized through PGD$^{20}$/PGD$^{100}$), BPDA [43], InfoPGD$^{20}$ (20-step adaptive PGD attack [44] using (18) to find adversarial examples), and auto-attack (AA) [45], which is a powerful and reliable attack that integrates many parameter-free attacks to evaluate the robustness of the network. AA includes three white-box attacks (APGD-CE, APGD-DLR, and FAB [46]) and a black-box attack (Square Attack [47]). All attacks are subject to the same $L_{\infty}$ perturbation limit, specifically, $\epsilon = 8/255$. Table II shows the white-box robustness of all defense models, where “Natural” represents the accuracy of natural test images. From Table II, we can observe, InfoAT achieves the best robustness against all nine types of attacks on all CIFAR-10, CIFAR-100, and SVHN datasets. For example, on the CIFAR-10 dataset, when attacked by BPDA, InfoAT achieves 56.84% defense accuracy, which is much higher than that (48.18%) on GanDef. When we apply the adaptive InfoPGD$^{20}$ attack, we get an accuracy of 33.92% for InfoAT on CIFAR-100, which is still higher than the accuracy of 30.58% for AT. Furthermore, compared with MART, we can observe that MART+ further improves the robustness of the model (e.g., on the SVHN dataset, the accuracy of MART+ is 57.15% when attacked by CW$_{\infty}$, while the accuracy of MART is 56.95%). This verifies the effectiveness of using mutual information to weight examples. It should be noted that the InfoAT robustness improvement is not caused by the so-called “obfuscated gradients” [43]. This can be verified from three phenomena:

1) Strong test attacks (e.g., PGD$^{20}$) have higher success rates (lower accuracies) than weak test attacks (e.g., FGSM);
2) InfoAT can defend AA effectively (i.e., InfoAT achieves 49.67% accuracy when attacked by AA on the CIFAR-10 dataset);
3) Following [48], we plot the adversarial loss change by moving the weight $w$ along a random direction $d$ with magnitude $\delta$ to visualize the weight loss surface. As presented in Fig. 4, the weight loss surface is flatter (red line) than AT (blue line), TRADES (orange line), and no regularization (i.e., None with blue line) in Fig. 3(d). As can be observed, the robustness can be improved steadily by our proposed regularization (red line). This verifies the effectiveness of our regularization, that is, improving the mutual information of adversarial example with its associated representation during training.
Fig. 5. Plot of the loss surface of an InfoAT trained model on perturbed images $x'_i, i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The dataset is CIFAR-10. The model is ResNet-18. (a) Loss surface of $x'_1$. (b) Loss surface of $x'_2$. (c) Loss surface of $x'_3$. (d) Loss surface of $x'_4$. (e) Loss surface of $x'_5$. (f) Loss surface of $x'_6$. (g) Loss surface of $x'_7$. (h) Loss surface of $x'_8$. (i) Loss surface of $x'_9$. (j) Loss surface of $x'_{10}$.

TABLE IV

| Methods  | CIFAR-10 | SVHN |
|----------|----------|------|
|          | ResNet18 | WideResNet | ResNet18 | WideResNet |
|          | ADBD | Rob.Acc | ADBD | Rob.Acc | ADBD | Rob.Acc |
| GanDef   | 0.0328 | 49.60 | 0.0376 | 55.10 | 0.0367 | 50.00 |
| AT       | 0.0361 | 53.30 | 0.0371 | 54.20 | 0.0355 | 49.80 |
| TRADES   | 0.0379 | 51.90 | 0.0402 | 55.20 | **0.0383** | 52.80 |
| MART     | 0.0374 | 52.60 | 0.0399 | 56.50 | 0.0374 | 53.30 |
| InfoAT (Ours) | **0.0380** | **56.20** | **0.0402** | **56.50** | **0.0378** | **53.60** |

and MART (green line). This indicates the effectiveness and stability of our proposed InfoAT.

Fig. 5 shows the loss surfaces of our proposed InfoAT trained model on perturbed images of the form $x'_i = x_i + \delta_1 v + \delta_2 r$, obtained by varying $\delta_1$ and $\delta_2$. We craft PGD$^{10}$ perturbation as the normal direction $v$, and $r$ be a random direction, under the $L_\infty$ constraint of $8/255$. From Fig. 5, we can observe the loss surfaces of InfoAT trained model on perturbed images are smooth, which reveals the fact that the robustness improvement of our proposed InfoAT is not caused by gradient obfuscation.

2) Black-Box Robustness: We utilize SPSA [49] to perform query-based black-box attacks. To estimate the gradients, we set the batch size to 128, 256, 512, and 1024, the learning rate to 0.01, and the perturbation size to 0.001. SPSA attacks are run for 100 iterations, and early stop when they caused misclassification. Table III shows the experimental results. Our method obtains better robustness over the other five methods in each setting. The results verify that InfoAT reliably improves the robustness rather than causing gradient masking.

We further perform the more black-box attack using the hard-label attack RayS [50]. We record the average decision boundary distance (ADBD) and robust accuracy (Rob.Acc) against RayS in Table IV. On the CIFAR-10 dataset, InfoAT achieves maximum ADBD and the best robust accuracy within ResNet18 and WideResNet-34-10. On SVHN, InfoAT also achieves competitive ADBD and the best robust accuracy.

These analyses verify that InfoAT improves adversarial robustness effectively rather than gradient obfuscation or masking.

C. Benchmarking the State-of-the-Art Robustness

In this part, we conduct more experiments on the large-capacity network WideResNet to arrive at a state-of-the-art robustness benchmark. In addition, we also used unlabeled data for further robustness evaluation.

1) Performance on WideResNet: We evaluate the robustness of our proposed InfoAT defense method more comprehensively on the WideResNet-34-10 network and given the state-of-the-art robustness benchmarks on the CIFAR-10 and CIFAR-100 datasets. Table V shows the results of our experiment. As shown in Table V, the InfoAT we proposed is superior to all baseline methods in terms of robustness. Especially under the PGD$^{20}$ attack on CIFAR-10, which is the most common comparison setting. InfoAT improves $\sim7\%$ over standard AT, $\sim6\%$ over GanDef, $\sim4\%$ over TRADES, and $\sim2\%$ even over MART. For CIFAR-100, a similar improvement trend can also be observed.

2) Boosting With Additional Unlabeled Data: In this part, the effectiveness of the proposed InfoAT is evaluated on the semisupervised version. Following the exact settings in UAT++ [51], RST [17], and MART, we compare the robustness of InfoAT with them on WideResNet-28-10 against PGD$^{20}$ respectively. The dataset is CIFAR-10 with 500 K unlabeled data, which is extracted from the 80 Million Tiny Images...
dataset [52]. As confirmed in Table VI, our proposed InfoAT also achieves the best robustness (e.g., 65.4% accuracy on WideResNet-28-10 with 500 K unlabeled data). This again verifies the benefit of exploiting mutual information to select "hard" examples for improving robustness and further proves the superiority of our proposed method.

V. CONCLUSION

In this article, inspired by the IB principle, we showed that the example with high mutual information is more likely to be \textit{robust}, and it can also achieve the best robustness in the semi-supervised task. In the future, we plan to investigate the tradeoff between robustness and natural accuracy on AT and exploit the different impacts of examples to improve the robustness as well as the accuracy of models efficiently.

TABLE V

| Dataset | Model | Natural | FGSM | PGD$^{20}$ | PGD$^{100}$ | PGD$^+$ | CW$_{20}^+$ | CW$_{100}^+$ | BPDA | InfoPGD$^{20}$ | AA |
|---------|-------|---------|------|----------|-----------|---------|------------|------------|------|---------------|----|
| CIFAR-10 | GanDef | 88.78 | 68.04 | 55.68 | 52.84 | 52.98 | 54.34 | 51.19 | 54.43 | 53.83 | 48.20 |
| AT | 87.05 | 66.58 | 53.56 | 51.22 | 51.53 | 53.54 | 51.67 | 56.05 | 56.26 | 49.70 |
| TRADES | 84.93 | 66.95 | 56.48 | 54.72 | 54.63 | 54.61 | 53.48 | 57.30 | 57.53 | 52.35 |
| MART | 84.10 | 67.14 | 58.29 | 56.77 | 56.67 | 54.69 | 54.32 | 57.16 | 57.59 | 51.64 |
| MART+ | 84.22 | 68.43 | 59.50 | 57.75 | 57.73 | 55.26 | 53.34 | 58.60 | 58.91 | 51.70 |
| InfoAT (Ours) | 85.62 | 68.91 | 60.46 | 59.17 | 59.00 | 55.32 | 54.07 | 59.54 | 60.09 | 52.86 |

TABLE VI

| Model | Natural | PGD$^{20}$ |
|-------|---------|-----------|
| UAT++ | 86.21 | 62.76 |
| RST | 90.60 | 63.60 |
| MART | 90.20 | 64.80 |
| InfoAT (Ours) | 90.00 | 65.40 |

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Stateline, NV, USA, vol. 25, Dec. 2012, pp. 1097–1105.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1701–1708.

[3] C. Szegedy et al., "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent., 2014, pp. 1–10.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent., 2015, pp. 1–11.

[5] J. Levinson et al., “Towards fully autonomous driving: Systems and algorithms,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2011, pp. 163–168.

[6] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “DeepDriving: Learning affordance for direct perception in autonomous driving,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2722–2730.

[7] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, “Towards evaluating the robustness of deep diagnostic models by adversarial attack,” Med. Image Anal., vol. 69, Apr. 2021, Art. no. 101977.

[8] J. A. Marpaung, M. Sain, and H.-J. Lee, “Survey on malware evasion techniques: State of the art and challenges,” in Proc. 14th Int. Conf. Adv. Commun. Technol., Feb. 2012, pp. 744–749.

[9] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, “Countering adversarial examples using input transformations,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–12.

[10] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2017, pp. 155–157.

[11] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, “Thermometer encoding: One hot way to resist adversarial examples,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–22.

[12] Z. Zou, T. Shi, Z. Shi, and J. Ye, “Adversarial training for solving inverse problems in image processing,” IEEE Trans. Image Process., vol. 30, pp. 2513–2525, 2021.

[13] B. Sun, N.-H. Tsai, F. Liu, R. Yu, and H. Su, “Adversarial defense by stratified convolutional sparse coding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11447–11456.

[14] X. Zhang, J. Wang, T. Wang, R. Jiang, J. Xu, and L. Zhao, “Robust feature learning for adversarial defense via hierarchical feature alignment,” Inf. Sci., vol. 560, pp. 256–270, Jun. 2021.

[15] H. Zhang, Y. Yu, J. Jian, E. Xing, L. El Ghaoui, and M. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in Proc. Int. Conf. Mach. Learn., May 2019, pp. 7472–7482.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–28.

[17] Y. Carmon, A. Raghunathan, L. Schmidt, and J. Duchi, “Unlabeled data improves adversarial robustness,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1–12.

[18] H. Zhang and J. Wang, “Defense against adversarial attacks using feature scattering-based adversarial training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1831–1841.

[19] G. Liu, I. Khalil, and A. Khreishah, “ZK-GanDef: A GAN based zero knowledge adversarial training defense for neural networks,” in Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN), Jun. 2019, pp. 19–32.

[20] G. Liu, I. Khalil, and A. Khreishah, “ZK-GanDef: A GAN based zero knowledge adversarial training defense for neural networks,” in Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN), Jun. 2019, pp. 64–75.

[21] S. Park and J. So, “On the effectiveness of adversarial training in defending against adversarial example attacks for image classification,” in Proc. Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 2513–2523, 2021.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
