# A Mutual Information Regularization for Adversarial Training

**Modeste Atsague** MODESTE@IASTATE.EDU
*Iowa State University*

**Olukorede Fakorede** FAKOREDE@IASTATE.EDU
*Iowa State University*

**Jin Tian** JTIAN@IASTATE.EDU
*Iowa State University*

**Editors:** Vineeth N Balasubramanian and Ivor Tsang

## Abstract

Recently, a number of methods have been developed to alleviate the vulnerability of deep neural networks to adversarial examples, among which adversarial training and its variants have been demonstrated to be the most effective empirically. This paper aims to further improve the robustness of adversarial training against adversarial examples. We propose a new training method called mutual information and mean absolute error adversarial training (MIMAE-AT) in which the mutual information between the probabilistic predictions of the natural and the adversarial examples along with the mean absolute error between their logits are used as regularization terms to the standard adversarial training. We conduct experiments and demonstrate that the proposed MIMAE-AT method improves the state-of-the-art on adversarial robustness.

**Keywords:** Adversarial Examples, Adversarial Training, Mutual Information.

## 1. Introduction

The progress of deep neural networks is remarkable. Often said to be capable of achieving human-level intelligence, deep learning models have outperformed human-level intelligence in some complex tasks. They have defeated human grandmasters in Go, a game so difficult that it was once thought to be beyond the reach of AI because there are more possibilities for an algorithm to explore in a single game than there are atoms in the universe (Silver et al., 2017). In computer vision, this progress significantly improves the performance of self-driving car (Chen and Huang, 2017) and medical image processing. However, deep learning models have been shown to be vulnerable to adversarial examples, which lead the model to predict, with high confidence, a wrong class (Goodfellow et al., 2014).

A number of defense strategies have been proposed in response to the vulnerability of deep learning models. The strategies can be classified into two categories: adversarial detection and adversarial defense. Adversarial detection seeks to detect malicious samples before they are fed to the model (Li and Li, 2017; Feinman et al., 2017; Meng and Chen, 2017; Xu et al., 2018). To detect malicious examples, Xu et al. (2018) suggested detection using feature squeezing, while Li and Li (2017) proposed detecting adversarial examples by analyzing whether they come from the same distribution as the normal examples, instead of training a deep neural network to detect them. Meng and Chen (2017) proposed MagNet that includes one or more separate detector networks and a reformer network where the detector networks learn to differentiate between normal and adversarial examples by approximating the manifold of normal examples.

On the other hand, adversarial defense can be classified into two subgroups: certified and empirical defenses. Certified defense provides a provable guarantee of adversarial robustness to norm bounded attacks (Balunovic and Vechev, 2019; Cohen et al., 2019; Wong and Kolter, 2018; Raghunathan et al., 2018; Zhang et al., 2019). However, the state-of-the-art certified defenses at this point achieve lower natural as well as robust accuracy when compared to state-of-the-art empirical defenses. Among the empirical defense methods, adversarial training and its variants have been demonstrated to be the most successful (Athalye et al., 2018). Adversarial training incorporates adversarial examples during the training of deep learning models to make them less vulnerable to adversarial examples. Let $D = \{(x_i, y_i)\}_{i=1}^{n}$ be a set of data points where $x_i$ are feature vectors and $y_i$ class labels. The objective function of the standard adversarial training (Madry et al., 2018) is defined as follows:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{x' \in B_\epsilon[x_i]} l(f_\theta(x_i'), y_i), \tag{1}$$

where $f_\theta(.)$ is the prediction of neural networks with parameters $\theta$, $l(.)$ is the loss function and $B_\epsilon[x] = \{x' |\ \|x' - x\|_p < \epsilon\}$ is a neighborhood of $x$. The inner maximization is used to generate adversarial examples given inputs $x_i$, which are then used to train the model by minimizing the loss (outer minimization ).

Inspired by the empirical success of adversarial training, several variants have been proposed (Cai et al., 2018; Zhang et al., 2019; Wang et al., 2019, 2020). In this paper, we work along the line of adversarial training and propose a simple, yet very effective training objective consisting of two regularization terms to the adversarial training loss function in Eq. (1). Our main contributions are summarized as follows:

- We propose a new training objective, termed MIMAE-AT, for improving the robustness of deep neural networks against adversarial examples that consists of two regularization terms to the standard adversarial training: mutual information between the probabilistic predictions of the natural example and its adversarial version and the mean absolute error between their logits.
- We experimentally demonstrate that the proposed MIMAE-AT method improves the state-of-the-art on adversarial robustness against common attacks and achieves a better trade-off between the natural accuracy and adversarial robustness.

## 2. Related Work

A number of adversarial defense methods have been proposed based on utilizing adversarial examples during the training (Madry et al., 2018; Kannan et al., 2018; Cai et al., 2018; Zhang et al., 2019; Wang et al., 2019, 2020; Ding et al., 2020). In particular, encouraged by the success of the adversarial training (AT) method proposed by Madry et al. (2018), several variants have been proposed with varying regularization terms and training objective functions that improve adversarial robustness. Adversarial Logit Pairing (ALP) (Kannan et al., 2018) proposes a regularization term that minimizes the mean square error loss between the logits of natural and adversarial examples. However, Engstrom et al. (2018) raises doubts about the effectiveness of ALP. TRADES (Zhang et al., 2019) theoretically characterizes the trade-off between accuracy and robustness of classification problems and proposes a regularization term that trades adversarial robustness off against accuracy and is shown experimentally to improve adversarial robustness over the standard AT. MMA (Ding et al., 2020) proposes a training objective that directly maximizes the input margin for each data point to achieve adversarial robustness and is shown to have improved performance over TRADES. MART (Wang et al., 2020) proposes a regularization term that explicitly differentiates misclassified and correctly classified examples and achieves the state-of-the-art adversarial robustness.

In this work, we aim to further improve the adversarial training and propose to use the mutual information between the probabilistic predictions of the natural and the adversarial examples as a regularization term to the standard AT. Mutual information has been heavily used in information theory for various applications, e.g. clustering and image segmentation. Several recent works have used mutual information to train deep neural networks. For example, Hu et al. (2017) maximizes the mutual information between data and its representation to encourage the predicted representations of augmented data points to be close to those of the original data, and Hjelm et al. (2018) leverage mutual information estimation for representation learning. A mutual information objective function has also been used in unsupervised clustering along the same line in Ji et al. (2019) aiming to group data points into classes by training without any labels a randomly initialized neural network into a classification function. A recent method proposes unsupervised learning to obtain robust representations by maximizing the worst-case mutual information between the input and output distributions (Zhu et al., 2020).

## 3. Preliminaries

We will use capital letters $X$, $Y$ to represent randoms variables and lower-case letters $x$ and $y$ to represent their realisation.

Mutual information is often used to measure the mutual dependency between two random variables and is defined for discrete random variables as follows:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} P(x,y) log(\frac{P(x,y)}{P(x)P(y)}), \tag{2}$$

where $P(X,Y)$ is the joint probability mass function of $X$ and $Y$.

### 3.1. Notation

We study a classification problem with $C$ classes. Let the data set be $D = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i$ is a (natural) input example associated with the label $y_i \in \{1, ......, C\}$. Let $f_c(x_i, \theta)$ be the *logit* output of the deep neural network with model parameters $\theta$ corresponding to class $c$. Let the output of the network be interpreted as a distribution of a discrete random variable $Z$ over $C$ classes. The probability that the network predicts a class "c" given the input example $x_i$ is given by the following softmax function:

$$p_c(x_i, \theta) = P(Z = c | X = x_i, \theta) = \frac{e^{f_c(x_i,\theta)}}{\sum_{c'=1}^{C} e^{f_{c'}(x_i,\theta)}}. \tag{3}$$

Given input $x_i$, the output $Z$ of the network follows the distribution $p(x_i, \theta)$. The prediction of the network is given by $f_\theta(x_i) = argmax_c \; p_c(x_i, \theta)$. The loss of the model over the dataset $D$ for a loss function $l(.)$ is defined as follows:

$$E[l(.)] = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i). \tag{4}$$

### 3.2. Adversarial Data Generation

A lot of works have been done seeking to generate (with small perturbations) adversarial examples that fool a model with high confidence, including Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), BIN (Kurakin et al., 2016), R-FGSM (Tramèr et al., 2018), FFGSM (Wong et al., 2020), DeepFool (Moosavi-Dezfooli et al., 2016), and strong attacks PGD (Madry et al., 2018) and CW (PGD optimized with CW loss) (Carlini and Wagner, 2017).

Let an $\epsilon$ *neighborhood* of $x$ be defined as $B_\epsilon[x] = \{x' | \, \|x' - x\|_p < \epsilon\}$ where $\| \cdot \|_p$ denotes the $L_p$ norm. To generate strong adversarial examples, PGD performs a multiple-step maximization of the inner part in Eq. (1) by projected gradient descent (PGD) on the negative loss:

$$x'_i \leftarrow \prod_{B_\epsilon[x]} (x'_i + \alpha \cdot sgn(\nabla_{x'_i} l(f_\theta(x_i), y_i))) \tag{5}$$

where $\prod(.)$ is the projection operator.

## 4. Proposed Defense

### 4.1. Empirical Risk

Consider the adversarial risk on dataset $D = \{(x_i, y_i)\}_{i=1}^n$ of the classifier model $f_\theta(.)$ with the 0-1 loss formulated as (Madry et al., 2018; Zhang et al., 2019):

$$Risk(f_\theta(.)) = \frac{1}{n}\sum_{i=1}^{n} \max_{x'_i \in B_\epsilon[x]} \mathbb{1}(f_\theta(x'_i) \neq y_i) \tag{6}$$

where $\mathbb{1}(.)$ is the indicator function.

Standard AT (Madry et al., 2018) aims to minimize the adversarial risk in Eq. (6). We propose to add a regularization term to the adversarial risk to encourage learning a model that makes the same classification decisions on the natural input example and its perturbed version. Similar ideas have been used for improving the robustness of deep neural networks (Zheng et al., 2016; Wang et al., 2020). We believe a perfectly robust classifier should produce exactly the same prediction on an input and its corresponding adversarial version, which is not reflected in the adversarial risk in Eq. (6). We hypothesize minimizing $\mathbb{1}(f_\theta(x_i) \neq f_\theta(x'_i))$ as a regularization term may help improve adversarial robustness. We therefore propose to train a model that minimizes the following regularized adversarial risk

$$Risk_r(f_\theta(.)) = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}(f_\theta(x'_i) \neq y_i) + \mathbb{1}(f_\theta(x_i) \neq f_\theta(x'_i))], \tag{7}$$

where

$$x'_i = \arg\max_{x'_i \in B_\epsilon[x]} \mathbb{1}(f_\theta(x'_i) \neq y_i). \tag{8}$$

### 4.2. MIMAE-AT Training Objective

Directly minimizing the empirical risk defined in Eq. (7) with 0-1 loss is intractable. The 0-1 loss is usually replaced by an appropriate convex surrogate loss in practice.

The most commonly used surrogate loss for the $\mathbb{1}(f_\theta(x'_i) \neq y_i)$ term in Eq. (7) is the cross-entropy (CE) loss defined by

$$CE(p(x'_i, \theta), y_i) = -\log p_{y_i}(x'_i, \theta), \tag{9}$$

where $p_{y_i}(x'_i, \theta)$ is the probability that the network predicts class $y_i$ given the input example $x'_i$. We will instead use the following boosted cross-entropy (BCE) proposed in Wang et al. (2020) as the surrogate loss for the $\mathbb{1}(f_\theta(x'_i) \neq y_i)$ term:

$$BCE(p(x'_i, \theta), y_i) = -\log p_{y_i}(x'_i, \theta) - log(1 - \max_{k \neq y_i} p_k(x'_i, \theta)), \tag{10}$$

where the first term is the standard CE loss and the second term is used to improve the decision margin of the classifier. The main reason to use the BCE loss is that it has been argued that correct classification of adversarial examples requires a stronger classifier than needed for natural data classification (Madry et al., 2018; Wang et al., 2020).

For the regularization term $\mathbb{1}(f_\theta(x_i) \neq f_\theta(x'_i))$ in Eq. (7), we propose a surrogate loss that uses the mutual information between the class assignments of the natural example $x$ and its disturbed version $x'$, denoted by $MI(Z, Z')$, where $Z$ and $Z'$ are the outputs of the network given inputs $x$ and $x'$ respectively (interpreted as a distribution of a random variable over $C$ classes). Maximizing $MI(Z, Z')$ will maximize the predictability of $Z$ from $Z'$ and vice versa, which encourages learning a model that preserves the common characteristics between $x$ and $x'$ while discarding details specific to each instance. Maximizing mutual information between the class assignments of pairs of images have been used in the invariant information clustering (IIC) method with great success (Ji et al., 2019).

We will often write $MI(Z, Z')$ as $MI(p(x_i, \theta), p(x'_i, \theta))$ as $Z$ and $Z'$ follow the distributions $p(x_i, \theta)$ and $p(x'_i, \theta)$ respectively. We will describe how to compute $MI(p(x_i, \theta), p(x'_i, \theta))$ in Section 4.2.1.

In addition, we propose to use the mean absolute error (MAE) between the logits of the natural example $x$ and its adversarial version $x'$, given by $\|f(x'_i, \theta) - f(x_i, \theta)\|_1$, as another surrogate loss for $\mathbb{1}(f_\theta(x_i) \neq f_\theta(x'_i))$. This loss between logits offers an extra regularization encouraging learning a model with better embedding of data such that the natural example and its adversarial version obtain similar internal representation. This idea of regularizing with a loss between logits of natural and adversarial examples has been used in the ALP method for adversarial robustness (Kannan et al., 2018). We note that both MAE and mean squared error (MSE) loss are reasonable choices. We have used MAE because experiments showed it had slightly better performance than MSE.

In summary, we propose the following *Mutual Information and Mean Absolute Error Adversarial Training (MIMAE-AT)* objective:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \left[ BCE(p(x'_i, \theta), y_i) - \lambda\, MI(p(x_i, \theta), p(x'_i, \theta)) + \beta \|f(x'_i, \theta) - f(x_i, \theta)\|_1 \right], \tag{11}$$

where $\lambda$ and $\beta$ are the regularization hyperparameters, and the adversarial examples $x_i'$ will be generated following the standard AT method (Madry et al., 2018) as follows:

$$x_i' = \underset{x_i' \in B_\epsilon[x]}{\arg\max}\, CE(p(x_i', \theta), y_i). \tag{12}$$

Our proposed MIMAE-AT algorithm for training deep neural networks is presented in the following Algorithm 1.

**Algorithm 1** MIMAE-AT

**Input:** Training data $D = \{x_i, y_i\}_{i=1}^{n}$, step size $\mu_1$ and $\mu_2$ for the inner and the outer optimization respectively, the batch size $m$, the number of outer iteration $T$, and the number of inner iteration $K$.
**Initialization:** Initialize $f_\theta$

```
for $t = 1, 2, ...., T$ do
    At random, uniformly sample a mini batch of training data $B_{(t)} = \{x_1, ..., x_m\}$
    for each $x_i \in B_{(t)}$ do
        $x_i^{'} = x_i + 0.001 \times k; k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        for $k = 1, 2, ...., K$ do
            $x_i^{'} = \prod_{B_\epsilon[x_i]}(x_i^{'} + \mu_1 sgn(\nabla_{x_i^{'}} CE(p(x_i', \theta), y_i)))$
        end for
    end for
    $\theta = \theta - \frac{\mu_2}{m} \sum_{i=1}^{m} \nabla_\theta \left[BCE(p(x_i', \theta), y_i) - \lambda\, MI(p(x_i, \theta), p(x_i', \theta)) + \beta \|f(x_i', \theta) - f(x_i, \theta)\|_1\right]$
end for
```

### 4.2.1. COMPUTATION OF MUTUAL INFORMATION

Computing mutual information $MI(Z, Z')$ requires the joint distribution $P(Z, Z')$ (see Eq. (2)) while we only have $P(Z|x, \theta) = p(x, \theta)$ and $P(Z'|x', \theta) = p(x', \theta)$. We will estimate the joint $P(Z, Z')$ and in turn mutual information $MI(Z, Z')$ using a batch of natural and adversarial example pairs $B = \{x_i, x_i'\}_{i=1}^{m}$, where $m$ is the batch size, following the method in Ji et al. (2019).
We have that the conditional joint distribution is given by $P(Z = y, Z' = y'|x, x', \theta) = p_y(x, \theta)p_{y'}(x', \theta)$, which states that $Z$ and $Z'$ are independent when conditioned on specific inputs $x$ and $x'$. However, they are not independent after marginalization over a set of input pairs $\{x_i, x_i'\}$. If we marginalize over the batch $B$ of natural and adversarial example pairs, the joint probability distribution is given by a $C$ by $C$ matrix $\mathcal{P}$ where the elements at rows $y$ and column $y'$ constitutes $\mathcal{P}_{yy'} = P(Z = y, Z' = y')$ :

$$\mathcal{P} = \frac{1}{m} \sum_{i=1}^{m} p(x_i, \theta) p(x_i', \theta)^T. \tag{13}$$

$\mathcal{P}$ can be symmetrized using $(\mathcal{P} + \mathcal{P}^T)/2$. Finally, the mutual information $MI(Z, Z')$ is computed as follows:

$$MI(p(x, \theta), p(x', \theta)) = MI(Z, Z') = \sum_{y=1}^{C} \sum_{y'=1}^{C} \mathcal{P}_{yy'} \ln \frac{\mathcal{P}_{yy'}}{\mathcal{P}_y \mathcal{P}_{y'}}, \tag{14}$$

where $\mathcal{P}_y = P(Z = y)$ and $\mathcal{P}_{y'} = P(Z' = y')$ are marginal probabilities and can be computed by summing over the rows and the columns of the matrix $\mathcal{P}$.

### 4.3. Related Work

We summarize in Table 4.3 the differences between our proposed MIMAE-AT training objective and the existing variants of the standard adversarial training (AT) (Madry et al., 2018), including ALP (Kannan et al., 2018), TRADES (Zhang et al., 2019), MMA (Ding et al., 2020), and MART (Wang et al., 2020).

Table 1: Loss function comparison with related work

| Method | Loss Function |
|---|---|
| Standard AT | $CE(p(x_i',\theta),y_i)$ |
| ALP | $CE(p(x_i',\theta),y_i)+\lambda\vert\vert f(x_i',\theta)-f(x_i,\theta)\vert\vert_2^2$ |
| TRADES | $CE(p(x_i,\theta),y_i)+\lambda\cdot KL(p(x_i,\theta)\vert\vert p(x_i',\theta))$ |
| MMA | $CE(p(x_i',\theta),y_i)\cdot\mathbb{1}(f_\theta(x_i)=y_i)+CE(p(x_i,\theta),y_i)\cdot\mathbb{1}(f_\theta(x_i)\neq y_i)$ |
| MART | $BCE(p(x_i',\theta),y_i)+\lambda\cdot KL(p(x_i,\theta)\vert\vert p(x_i',\theta))\cdot(1-p_{y_i}(x_i,\theta))$ |
| **MIMAE-AT** | $BCE(p(x_i',\theta),y_i)-\lambda\cdot MI(p(x_i,\theta),p(x_i',\theta))+\beta\Vert f(x_i',\theta)-f(x_i,\theta)\Vert_1$ |

Standard AT minimizes the adversarial risk via the CE loss over adversarial examples. ALP introduces a loss between logits of clean and adversarial examples as a regularization term. However, Engstrom et al. (2018) raises doubts about the effectiveness of ALP. TRADES uses standard loss and introduces KL-divergence between output probabilities of clean and adversarial examples as a regularization. MMA treats correctly and incorrectly classified examples differently, wherein only correctly classified examples were adversarially perturbed. Like TRADES, MART also uses the KL-divergence between clean and adversarial output distributions as a regularization but treats incorrectly classified examples differently like MMA.

One notable difference of our proposed MIMAE-AT with the existing work is that we regularize the adversarial loss with the mutual information between the class assignments of the clean example and its adversarial version. The mutual information term has not been used in the setting of adversarial robustness before. It encourages learning a model that preserves the common characteristics between the natural image and its disturbed version while discarding details specific to each instance, which has different practical effects from KL-divergence.

## 5. Experiments

We conducted a series of experiments to evaluate the effectiveness of the proposed method. We compare MIMAE-AT with the state-of-the-art defenses on benchmark datasets in both white-box and black-box settings. We report the experimental results in this section.

### 5.1. Settings

**Baselines:** We compare MIMAE-AT with the top performing defenses to date Standard AT (Madry et al., 2018), TRADES (Zhang et al., 2019), and MART (Wang et al., 2020) (MMA (Ding et al., 2020) is shown to be outperformed by MART in Wang et al. (2020) and hence is not compared with). For TRADES, regularization parameter $\lambda$ is set to 4.0, and for MART, 5.0, as set in their papers.

**Threat models:** We evaluate the effectiveness of the defense methods under the following $L_\infty$ attacks: Fast Gradient Sign Method FGSM (Goodfellow et al., 2014),strong attacks PGD (Madry

et al., 2018) and CW (PGD optimized with CW loss, confidence level $K = 50$) Carlini and Wagner (2017). In addition, we evaluate the defenses against $L_2$ attacks, specifically PGD-$L_2$ attack.
We evaluate the effectiveness of the defense methods on commonly used benchmark datasets CIFAR-10 (Krizhevsky and Hinton, 2009), CIFAR-100 (Krizhevsky and Hinton, 2009), and MNIST (LeCun et al., 1998).
**CIFAR-10**: We tested on two model architectures : ResNet-18 (He et al., 2016) and a larger capacity network WideResNet-34-10 (Zagoruyko and Komodakis, 2016). Both models were trained for 100 epochs, using SGD with a momentum of 0.9. We initialized the learning rate at 0.1, then decayed to $0.01$ at 75th, and further decayed to $0.001$ at the 90th epoch. For both models, we used the batch size of 128. Our regularization parameter $\lambda$ is set to .4 and .6 for ResNet-18 and WideResNet-34-10 respectively, and $\beta$ is set to 1.
The weight decay is set to $2$ x $10^{-4}$ and $5$ x $10^{-4}$ for ResNet-18 and WideResNet respectively. It is observed in Pang et al. (2021) that weight decay plays a significant role in robust accuracy and a fair comparison of defenses should take this factor into consideration. Therefore, for fair comparison, we re-evaluated all the defenses using the same weight decay. We note that the original implementation of Standard AT and TRADES used the same weight decay of $2 \times 10^{-4}$ when training ResNet-18, while MART used a weight-decay of $3.5 \times 10^{-3}$ which slightly improved the accuracy they reported. Adversarial data used in training are generated by PGD with random start and maximum perturbation $\epsilon$ set to 0.031 and the step size of 0.007 while the number of steps is 10. $L_\infty$ attacks are generated with perturbation $\epsilon = 0.031$ and the step size of 0.003. PDG-$L_2$ attacks are generated with perturbation 0.5, step size of 0.003, and 20 iterations.

**CIFAR-100**: We tested on ResNet-18 with the same settings as for CIFAR-10.

**MNIST**: The results on MNIST are based on a 4-Layer CNN from Carlini and Wagner (2017). We trained our model using SGD for 100 epochs with momentum 0.9 and the batch size $m = 128$. The initial learning rate is set to 0.01 then decay by a factor of 10 at the 55th, 75th, and 90th epochs. We used weight decay of $2 \times 10^{-4}$ for all defense methods. We used the same defense and attack settings as in Zhang et al. (2019): all adversarial examples are generated by PGD using a maximum perturbation $\epsilon = 0.3$ and the perturbation step size is 0.01 for training and 0.005 for attacking. We used 10 iterations for training and tested 20 and 40 iterations for attacking.

### 5.2. Sensitivity to Regularization Hyperparameters

We studied the influence of the regularization hyperparameters on the natural and robust accuracy. Based on the experimental results, the regularization parameter $\lambda$ is set to 0.4 and 0.6 for ResNet-18 and Wide-ResNet respectively, and $\beta$ is chosen to take the value of 1. Tables 2 and 3 show experimental results on CIFAR-10 with varying $\lambda$ values for ResNet-18 and WideResNet-34-10 respectively. Table 4 shows experimental results on CIFAR-10 for ResNet-18 with varying values of $\beta$. In Section 5.5, we further explore the impact of each of the (mutual information and MAE) regularization term on the adversarial robustness.

Table 2: The performance (results in %) under different values of regularization parameter $\lambda$ with $\beta = 1$. The model architecture is ResNet-18, and the dataset is CIFAR-10.

| $\lambda$ | 0.1 | 0.2 | 0.3 | **0.4** | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $Natural$ | 82.00 | 81.54 | 81.96 | **82.24** | 82.03 | 81.99 | 81.97 | 81.55 | 81.72 |
| PGD-20 | 54.31 | 53.35 | 53.81 | **54.38** | 53.81 | 54.08 | 53.65 | 53.81 | 53.54 |

Table 3: The performance (results in %) under different values of regularization parameter $\lambda$ with $\beta = 1$. The model architecture is WideResNet-34-10, and the dataset is CIFAR-10.

| $\lambda$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | **0.6** | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $Natural$ | 85.98 | 85.64 | 85.72 | 85.85 | 85.45 | **86.06** | 85.60 | 85.78 | 85.43 |
| PGD-20 | 57.33 | 57.91 | 57.94 | 57.63 | 57.97 | **58.03** | 57.49 | 58.08 | 57.81 |

Table 4: The performance (results in %) under different values of regularization parameter $\beta$ with $\lambda = 0.4$. The model architecture is ResNet-18, and the dataset is CIFAR-10.

| $\beta$ | 0.1 | 0.3 | 0.4 | 0.6 | 0.8 | 0.9 | **1.0** |
|---|---|---|---|---|---|---|---|
| $Natural$ | 83.74 | 83.22 | 83.15 | 83.52 | 83.58 | 83.06 | **82.24** |
| PGD-20 | 52.59 | 52.68 | 52.67 | 52.56 | 53.63 | 54.30 | **54.38** |

### 5.3. White-Box Robustness

White-box attacks assume that the attacker has full access to the model parameters.
**On CIFAR-10**:
The results on CIFAR-10 for ResNet-18 and WideResNet-34-10 are presented in Tables 5 and 6 respectively.

Table 5: Comparing white-box attack robustness (accuracy %) on CIFAR-10. The model architecture is **ResNet-18**.

| DEFENCE METHOD | Natural | FGSM | PGD-20 | PGD-50 | CW$_\infty$ | PGD-$L_2$ |
|---|---|---|---|---|---|---|
| STANDARD-AT | **84.15** | 58.71 | 51.40 | 47.65 | 46.15 | 61.98 |
| TRADES | 81.18 | 58.80 | 52.74 | 51.88 | 50.19 | 62.59 |
| MART | 82.17 | 58.50 | 54.04 | 52.13 | 48.94 | 62.30 |
| **MIMAE-AT** | 82.24 | **58.92** | **54.38** | **52.27** | **51.43** | **62.80** |

On CIFAR-10 for ResNet-18, our approach (MIMAE-AT) achieved a robust accuracy of 58.92% on FGSM, 54.38% on PGD-20, 51.43% on CW, and 62.80% on PGD-$L_2$, all better than or on par with TRADES, MART, and Standard AT.The improvements over existing defenses are relatively more significant on the strongest attack CW. However, Standard AT recorded the best accuracy on natural data. Engstrom et al. (2018) showed that the performance of models trained using adversarial logit pairing (ALP) degrades significantly with the increase in the number of steps of PGD attack. To verify that such behavior is not observed, we evaluated our defense also under PGD-50 and observed that there is no significant degradation in the performance of models trained using MIMAE-AT methods as well as TRADES and MART.

Table 6: Comparing white-box attack robustness (accuracy %) on CIFAR-10. The model architecture is **WideResNet-34-10**.

| DEFENCE METHOD | Natural | FGSM | PGD-20 | $\text{CW}_\infty$ |
|---|---|---|---|---|
| STANDARD-AT | **86.08** | 61.75 | 56.12 | 54.20 |
| TRADES | 84.66 | 61.36 | 56.90 | 54.20 |
| MART | 84.17 | 61.62 | **58.25** | 54.55 |
| **MIMAE-AT** | 86.06 | **63.64** | 58.03 | **55.87** |

On CIFAR-10 for WideResNet-34-10, MIMAE-AT achieved better or on bar performances compared with the baselines on both natural and robustness accuracy. Specifically, MIMAE-AT outperformed Standard AT on all attacks with a natural accuracy on par with Standard AT; MIMAE-AT outperformed TRADES on all attacks as well as natural accuracy; and MIMAE-AT outperformed MART on FGSM and CW attacks as well as natural accuracy with on par performance on PGD-20 attack. We also observe that the performances of all the WideResNet-34-10 models are better than the corresponding ResNet-18 models, echoing the belief that robust classification needs more complex classifiers (Nakkiran, 2019).

**On CIFAR-100**

The results on CIFAR-100 for ResNet-18 are presented in Table 7.

Table 7: Comparing white-box attack robustness (accuracy %) on CIFAR-100. The model architecture is **ResNet-18**.

| DEFENCE METHOD | Natural | FGSM | PGD-20 | $\text{CW}_\infty$ | PGD-$L_2$ |
|---|---|---|---|---|---|
| STANDARD-AT | **56.91** | 29.26 | 26.46 | 25.72 | 34.70 |
| TRADES | 53.60 | 30.45 | 28.65 | 25.47 | 35.94 |
| MART | 52.71 | 28.86 | 26.58 | 23.70 | 35.58 |
| **MIMAE-AT** | 56.51 | **31.19** | **28.92** | **26.28** | **38.50** |

Our approach (MIMAE-AT) achieved a robust accuracy of $31.19\%$ on FGSM, $28.92\%$ on PGD-20, $26.28\%$ on CW, and $38.50\%$ on PGD-$L_2$, outperforming TRADES, MART, and Standard AT on all attacks. In addition, MIMAE-AT achieved a natural accuracy on par with Standard AT while outperforming TRADES and MART.

**On MNIST**:

The results on MNIST are presented in Table 8.

Table 8: Comparing white-box attack robustness (accuracy %) on MNIST.

| DEFENCE METHOD | Natural | FGSM | PGD-20 | PGD-40 | CW$_\infty$ | PGD-$L_2$ |
|---|---|---|---|---|---|---|
| STANDARD-AT | **99.40** | 98.60 | 88.65 | 45.57 | 87.02 | 98.25 |
| TRADES | 98.92 | 98.59 | 91.57 | 69.11 | 88.59 | 97.58 |
| MART | 98.70 | 98.83 | **92.40** | **71.46** | 89.11 | 98.09 |
| **MIMAE-AT** | 99.29 | **99.11** | 92.09 | 71.15 | **89.45** | **98.34** |

On MNIST, all the defense methods achieved good performances on FGSM, PGD-20, CW, and PGD-$L_2$ attacks as well as natural accuracy relatively to CIFAR-10 and CIFAR-100 datasets due to the simplicity of MNIST data compared with CIFAR-10 and CIFAR-100. The performances of MIMAE-AT, TRADES, and MART are similar on all attacks and are better than Standard AT on strong PGD-20 and CW attacks. We also performed PGD-40 attack on MNIST, and observed all defense methods had a significant drop in accuracy from PGD-20 attack, particularly Standard AT (from 88.65% to 45.57%).

In summary, the experimental results in Tables 5-8 show that our proposed MIMAE-AT consistently outperforms Standard AT on white-box attacks, and consistently outperforms or at least is on par with TRADES and MART on white-box attacks while outperforming TRADES and MART on the natural accuracy. The results show that *MIMAE-AT achieves a better trade-off between the natural accuracy and adversarial robustness than these state-of-the-art defense methods*.

### 5.4. Black-Box Robustness.

We evaluated MIMAE-AT against black-box attacks on CIFAR-10 over ResNet-18. In the black-box attack setting, since the attacker has no access to the model parameters, adversarial examples are crafted using a substitute model (Papernot et al., 2017). We have used a more powerful network ResNet-50 as the substitute model.

The results are presented in Table 9. MIMAE-AT achieved $83.31\%$, $83.27\%$, and $83.45\%$ accuracy on FGSM, PGD-20, and CW attacks respectively, outperforming Stardard AT and TRADES and slight improving over MART.

Table 9: Comparing black-box attack robustness (accuracy %) on CIFAR-10 over ResNet-18 with substitute model ResNet-50. Except for our MIMAE-AT results, the results in the table are from Wang et al. (2020). We used the same settings specified in Wang et al. (2020) for a fair comparison.

| DEFENCE METHOD | FGSM | PGD-20 | CW$_\infty$ |
|---|---|---|---|
| STANDARD-AT | 79.98 | 80.01 | 80.85 |
| TRADES | 81.52 | 81.53 | 82.11 |
| MART | 82.75 | 82.70 | 82.95 |
| **MIMAE-AT** | **83.81** | **83.27** | **83.45** |

### 5.5. Ablation on Regularization Terms

We proposed two regularization terms to the adversarial risk training objective BCE loss: a mutual information (MI) term between the class assignments of the natural example $x$ and its adversarial version $x'$ and a mean absolute error (MAE) loss between the logits of $x$ and $x'$. We conducted experiments to study the impact of different combinations of the MI and MAE regularization terms on the natural and robust accuracy. The results on CIFAR-10 over WideResNet-34-10 are presented in Table 10. From Table 10, we observe that the combination (BCE + MAE) performed slightly better than the combination (BCE - $\lambda$ MI), and the best performance is achieved by using both MI and MAE terms.

Table 10: Impact of different combinations of regularization terms on white-box attack robustness (accuracy %). The results are on CIFAR-10 over WideResNet-34-10.

| Loss terms | Natural | FGSM | PGD-20 | CW$_\infty$ |
|---|---|---|---|---|
| BCE -$\lambda$ MI | 84.33 | 62.85 | 57.81 | 54.78 |
| BCE + MAE | 85.60 | 63.30 | 57.60 | 55.81 |
| BCE -$\lambda$ MI + MAE | **86.06** | **63.64** | **58.08** | **55.87** |
| BCE -$\lambda$ MI + MSE | 86.11 | 62.69 | 57.0 | 55.09 |
| CE -$\lambda$ MI + MAE | 86.60 | 62.71 | 56.58 | 55.70 |

In addition, we note that both MAE and mean squared error (MSE) are reasonable choices as regularization. Table 10 also shows the results on using MSE. We observe that the combination (BCE - $\lambda$MI + MAE) performed slightly better than the combination (BCE - $\lambda$ MI + MSE). We further compare BCE with the standard CE and report the results in Table 10, which shows that BCE performed slightly better.

### 5.6. Sensitivity of Mutual Information to Batch Size

One might be concerned that the estimation of the mutual information term may be sensitive to the batch size. We perform experiments using varying batch sizes with the mutual information term as the only regularizer, i.e., the training objective is BCE -$\lambda$ MI. The results are shown in Table 11. We observe that the performance of the mutual information term is not sensitive to the batch size.

Table 11: White-box attack robustness (accuracy %) of BCE -$\lambda$ MI with varying batch sizes on CIFAR-10. The model architecture is ResNet-18.

| Baches | 64 | 100 | 128 | 256 |
|---|---|---|---|---|
| Natural | 81.61 | 82.07 | 82.73 | 82.98 |
| PGD-20 | 53.79 | 53.39 | 53.23 | 53.80 |
| FGSM | 58.05 | 58.70 | 58.60 | 58.73 |
| $CW_\infty$ | 50.41 | 50.08 | 50.12 | 53.20 |

### 5.7. Obfuscated Gradients

It has been shown that some defense methods gave a false sense of security against adversarial examples by intentionally or inadvertently using obfuscated gradients (Athalye et al., 2018). Our experimental results indicate that MIMAE-AT does not have any characteristic behavior of obfuscated gradients specified in Athalye et al. (2018). Specifically, the results in Section 5.3 show that MIMAE-AT achieved better robustness on a one-step attack FGSM than on a multi-step attack PGD (see Tables 5-8). In addition, white-box attacks are more successful than black-box attacks (comparing Table 5 with Table 9). Therefore, we believe the adversarial robustness achieved by MIMAE-AT is not due to obfuscated gradients.

## 6. Conclusion

We propose a new training objective MIMAE-AT for adversarial training to defend deep neural networks against adversarial examples. We empirically evaluate the effectiveness of MIMAE-AT method on benchmark datasets in both white-box and black-box settings and compare it with the state-of-the-art defenses. The experimental results indicate that MIMAE-AT improves the state-of-the-art on adversarial robustness. In particular, MIMAE-AT achieves a better trade-off between the natural accuracy and adversarial robustness than the state-of-the-art defense methods.

## Acknowledgments

We thank the reviewers for their feedback helping to improve this manuscript. Jin Tian was partially supported by ONR grant N000141712140.

## References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2019.

Qi-Zhi Cai, Chang Liu, and Dawn Song. Curriculum adversarial training. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3740–3747. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/520. URL https://doi.org/10.24963/ijcai.2018/520.

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.

Zhilu Chen and Xinming Huang. End-to-end learning for lane keeping of self-driving cars. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1856–1860, 2017. doi: 10.1109/IVS.2017.7995975.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts. In *International Conference on Machine Learning*, 2017.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning*, pages 1558–1567. PMLR, 2017.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1607.html#KurakinGB16.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5772, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Xb8xvrtB8Ce.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354—359, October 2017. ISSN 0028-0836. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, volume 1, page 2, 2019.

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. URL https://openreview.net/forum?id=rklOg6EFwS.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings of the 2018 Network and Distributed Systems Security Symposium (NDSS)*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

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.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

Sicheng Zhu, X. Zhang, and D. Evans. Learning adversarially robust representations via worst-case mutual information maximization. In *ICML*, 2020.