REVISITING INSTANCE REWEIGHTED ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Instance Reweighted Adversarial Training (IRAT) is one of the adversarial training that assigns large weights to high-importance examples and then minimizes weighted loss. The importance often uses margins between decision boundaries and each example. In particular, IRAT can alleviate robust overfitting and obtain excellent robustness by computing margins with an estimated probability. However, previous works implicitly deal with binary classification even though multi-class because of computing margins with only true and the most confusing class. Computed margins probably become equal even different true probability examples because of drawing the complex decision boundaries in multi-class classification. In this paper, firstly, we clarify as aforementioned with specific examples. Then, we propose *reweighting for the margin* that can transform previous margins into appropriate representations for multi-class classification by leveraging the relations between the most confusing class and other classes. The experiment demonstrates proposed method is effective in boosting the robustness against several attacks of previous methods on CIFAR-10/100 datasets.

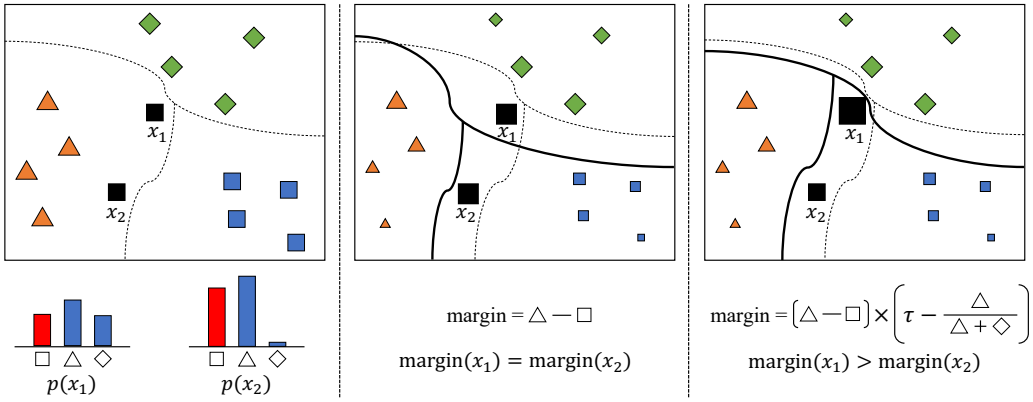
1 INTRODUCTION

While convolutional neural networks (CNNs) achieve excellent performance in various tasks, it endangers that CNNs are vulnerable to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015) that malicious perturbed samples. This perturbation is a threat against AI systems based on CNNs (e.g., autonomous driving or medical diagnosis) because of imperceptible to humans. Thus, there are various efficacy approaches to alleviate the negative impact of the perturbation (Papernot et al., 2016; Samangouei et al., 2018; Xu et al., 2018; Madry et al., 2018). Among them, adversarial training (AT) (Madry et al., 2018) is well known as an attractive defense strategy because of its clarity and efficacy.

Instead of benign samples, AT trains adversarial examples generated by projected gradient descent (PGD) (Madry et al., 2018). Although AT can realize excellent robustness, it endangers that occurring robust overfitting or the degenerating performance of benign samples (Zhang et al., 2019; Rice et al., 2020). Instance Reweighted Adversarial Training (IRAT) (Zeng et al., 2021; Kim et al., 2021; Zhang et al., 2021; Wang et al., 2021; Gao et al., 2021) is one of the effective methods in many approaches tried to overcome these issues.

IRAT computes margins between the decision boundaries and each example as the importance and transforms it to weights with a nonlinear increasing function. Then, IRAT minimizes weighted classification loss by assigning these weights to each example. Geometry-aware instance-reweighted adversarial training (GAIRAT) proposed by (Zhang et al., 2021) decides the importance for each example by the least PGD steps (LPS). GAIRAT represents the margin in an input space because LPS is the number of steps to make the adversarial example cross the decision boundaries, starting from a benign instance. Smaller margin examples indicate closer to the decision boundaries and assign larger weights. Although GAIRAT realizes excellent robustness compared to standard AT, it is vulnerable against other than PGD because it defines the importance depending on LPS.

Meanwhile, margin-aware instance reweighting learning (MAIL) (Wang et al., 2021) successfully overcomes the weakness of GAIRAT by defining the importance with estimated probabilities. Specifically, MAIL transforms the difference between two probabilities, i.e., true class and the most confusing class, to the weight with the nonlinear increasing function. Weighted minimax risk



(a) Before assigning weights to each sample.

(b) After assigning weights to each sample with MAIL or WMMR.

(c) After assigning weights to each sample with ours.

Figure 1: Difference between both previous methods (i.e., MAIL and WMMR) and our method. The size of each shape in (b) and (c) indicates the amount of weight, and solid and dashed lines are the decision boundaries before and after updates, respectively. Each shape is a different class and we describe the difference from (b) previous and (c) our method focused on black squares $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Previous methods assign equal margins to equal weights because of using only the difference between the true class and the most confusing class probability. However, we should define larger weight for $\boldsymbol{x}_1$ than $\boldsymbol{x}_2$ because $\boldsymbol{x}_1$ closes to the intersection of multiple class boundaries. Our method can transform margins for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ to appropriate forms in multi-class classification.

(WMMR) (Zeng et al., 2021) uses the same approach. The difference between these methods is the use of different weighting functions and MAIL achieves more excellent performance than WMMR. Different from GAIRAT with the discrete representation for weights, MAIL alleviates robust overfitting by using the continuous. However, MAIL and WMMR have a concern that the importance cannot adequately represent in multi-class classification because it does not consider other than the most confusing class and true class. As shown in Figure 1(a), we assume that previous approaches occur the issue for instances that have the same margin. Intuitively, among examples that have the same margin, the importance should define closer around the intersection of the decision boundaries of multiple classes but previous methods neglect its representation.

This paper reveals the problem of previous margin computation with a specific example to resolve this doubt. Then, as shown in Fig. 1(b), we propose *reweighting for the margin*, which enables us to transform previous margins to appropriate representation by also considering other than the most confusing class and true class. Although the straightforward approach to accomplish our purpose arise computing margins between true class and the others (i.e., using multi-class margin), it is hard to design the weighting function to aggregate the multi-class margin to the single weight. Thus, we propose a novel metric which is the ratio of the top2 in the sum of all probabilities except for true class (the incorrect rate). In the same previous idea, suppose each class probability well represents the center of the class, this metric can identify whether examples are close to the intersection of multi-class boundaries. Therefore, we do not require to design a special weighting function and can get appropriate representation just by multiplying this measure by previous margins. Experiments demonstrate that the proposed method is sufficient to boost the performance of previous methods in some attacks. In summary, our work makes the following contributions:

- We clarify that previous computing the margin using predicted probabilities is insufficient. Specifically, we show that the case where computes the same margins to some examples even though both true and the most confusing probabilities are different.
- We propose *reweighting for the margin* that enables previous margins to transform appropriate ones by leveraging a relation between the most confusing class probability and the incorrect rate.
- In experiments, we show our approach is efficacy for boosting the previous robustness against adversarial attacks.

2 PRELIMINARY AND RELATED WORKS

In this section, we give an overview standard training and adversarial training (AT), and then describe about instance reweighted adversarial trainig (IRAT) as related works.

2.1 STANDARD TRAINING VS. ADVERSARIAL TRAINING

Standard training: Let $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ be a training dataset where $\boldsymbol{x}_i \in \mathbb{R}^{c\times h\times w}$ is an input example and y_i is a ground truth label. In standard training, a deep neural net $f : \mathbb{R}^{c\times h\times w} \to \mathbb{R}^K$ parameterized by $\boldsymbol{\theta}$ minimizes the loss $\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)], \tag{1}$$

where K is the number of classes, and the loss function $\ell(\cdot)$ uses the cross-entropy loss.

Adversarial training: AT (Madry et al., 2018) aims to obtain the model robustness against adversarial attacks by training computed adversarial examples during the training. Except for image classification, this framework activates before deep learning becomes popular (Dalvi et al., 2004; Lowd & Meek, 2005). Different from standard training, AT consists of two processes which are inner-maximization and outer-minimization. First, the inner-maximization computes a perturbation $\boldsymbol{\delta}_i$ such as maximizing the loss within a radius ϵ centered $\boldsymbol{x}_i$. After that, the outer-minimization updates weight parameters $\boldsymbol{\theta}$ for minimizing the loss with adversarial examples $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{\delta}_i$ obtained by the inner-maximization. The following equation indicates these processes:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}_i\|_p \le \epsilon} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i), y_i) \right], \tag{2}$$

where $p = \{0, 1, 2, \infty\}$ and we use $p = \infty$ in this paper.

The inner-maximization often computes adversarial examples by a multi-step attack (e.g. PGD (Madry et al., 2018) with a step size $\alpha \le \epsilon$):

$$\hat{\boldsymbol{x}}_i^{(t+1)} := \Pi_{\mathcal{B}[\boldsymbol{x}_i^{(0)}]} \left(\hat{\boldsymbol{x}}_i^{(t)} + \alpha \cdot \mathrm{sign} \left(\nabla_{\hat{\boldsymbol{x}}_i^{(t)}} \ell(f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_i^{(t)}), y_i) \right) \right), \tag{3}$$

where $\mathcal{B}[\boldsymbol{x}_i^{(0)}] := \{\hat{\boldsymbol{x}}_i \in \mathcal{X} \mid \|\boldsymbol{x}_i^{(0)} - \hat{\boldsymbol{x}}_i\|_\infty \le \epsilon\}$ in the input space $\mathcal{X}$ and Π is a projection function which brings an outlier into $\mathcal{B}[\boldsymbol{x}_i^{(0)}]$.

Although AT is not hard to obtain robustness, it has drawbacks causing degeneration of classification performance of benign examples or the robust overfitting. These issues begins to overcome by various breakthrough approaches (Cai et al., 2018; Zhang et al., 2019; Uesato et al., 2019; Zhang & Wang, 2019; Wu et al., 2020; Wang et al., 2020; Zhang et al., 2020; Cheng et al., 2022; Ding et al., 2020; Song et al., 2021; Cui et al., 2021; Yu et al., 2022). Among them, we deal with IRAT (Zeng et al., 2021; Kim et al., 2021; Zhang et al., 2021; Wang et al., 2021; Gao et al., 2021).

2.2 INSTANCE REWEIGHTED ADVERSARIAL TRAINING

IRAT is adversarial training that defines weights based on the importance rate computed for each example and minimizes the weighted loss. IRAT has various approaches that are different computations of importance.

Geometry-aware instance-reweighted adversarial training (GAIRAT) (Zhang et al., 2021) determines the importance of each example using the least PGD steps (LPS). LPS indicates the number of steps to make the adversarial example cross the decision boundaries, starting from a benign instance. While GAIRAT can achieve excellent performance against the PGD attack, it is vulnerable to other strong attacks because of the importance depending on the PGD attack of the inner-maximization. To overcome the weakness of GAIRAT, Gao et al. (2021) proposed to consider other attacks, such as CW (Carlini & Wagner, 2017) or AutoAttack (AA) (Croce & Hein, 2020), into the inner-maximization, but this approach is costly. Meanwhile, entropy-weighted adversarial training (EWAT) (Kim et al., 2021) improves robustness against CW (Carlini & Wagner, 2017) and AutoAttack (AA) (Croce & Hein, 2020) by assigning an entropy to the input as weights to each example.

Different from EWAT, weighted minimax risk (WMMR) (Zeng et al., 2021) and margin-aware instance reweighting learning (MAIL) (Wang et al., 2021) define margins to each example using true class and the most confusing class probabilities as the following equation:

$$m(\boldsymbol{x}_i, y_i) = \arg\max_{j \neq y_i} p_j(\boldsymbol{x}_i) - p_{y_i}(\boldsymbol{x}_i), \tag{4}$$

where $p_{y_i}(\boldsymbol{x}_i)$ indicates true class probability since $p(\boldsymbol{x}_i)$ is the predicted distribution. These approaches are different weighting functions (MAIL uses sigmoid, WMMR uses exponential). MAIL resolves weakness of GAIRAT and realizes more excellent performance.

We aim to define the margin with Eq. (4) referred to same as the MAIL and WMMR in this paper. However, previous methods are hardly appropriate margins in multi-class classification because of ignoring other than relations between true class and the most con fusing class. Thus, we firstly reveal that representations of Eq. (4) are insufficient in multi-class classification and aim to improve adversarial robustness by reweighting to margins computed by Eq. (4).

3 Reweighting for the Margin

In this section, we describe the weakness of previous methods and details of the proposed method. First, we reveal the weakness of the previous IRATs in Section 3.1. Specifically, we show that Eq. (4) may obtain the same margins with some examples even though true class and the most confusing class probabilities are different. Next, we propose a novel metric to represent relations between the most confusing class and the others in Section 3.2. Finally, we introduce weighting functions that incorporated the proposed method into previous methods.

Definition 1 (top2 probability)**.** *Let $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}$ be the input data and let $p(\boldsymbol{x}_i) \in [0,1]^K$ be the posterior distribution to $\boldsymbol{x}_i$. Then, the top2 probability is defined as $p_2(\boldsymbol{x}_i) = \arg\max_{k \neq y_i} p_k(\boldsymbol{x}_i)$.*

Remark: Strictly speaking, the top2 probability is the highest when $p_{y_i}(\boldsymbol{x}_i) < p_2(\boldsymbol{x}_i)$. However, note we deal with the top2 probability in this situation in this paper because of clarification of a word.

3.1 Weakness of previous IRAT

Although previous IRATs with class probabilities achieve excellent performance, they do not take other than the most confusing class into account. Except for an easy case, some samples should lie around the intersection of multi-class of decision boundaries. Thus, we can easily suppose to be insufficient representations when WMMR and MAIL deal with multi-class classification.

Lemma 1. *For the margin calculated by Eq.(4) and let $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training dataset, there exist samples satisfy bellow*

$$m(\boldsymbol{x}_i, y_i) = m(\boldsymbol{x}_j, y_j). \tag{5}$$

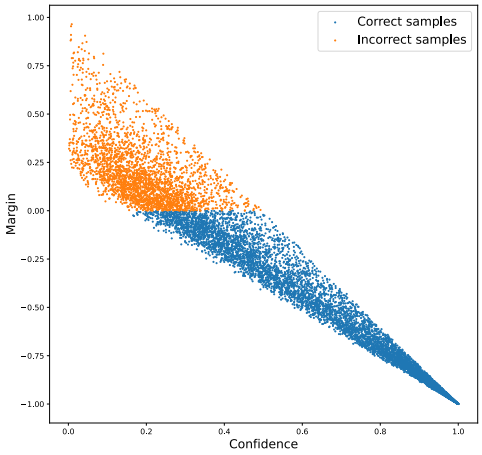


Figure 2: Correlations between true class probabilities $p_{y_i}(\boldsymbol{x}_i)$ and margins computed by Eq. 4. These results show standard adversarial trained ResNet-18 on the CIFAR-10 dataset. Orange and blue dots are incorrect and correct samples, respectively.

Based on the Lemma 1, we discuss computed margins by Eq. 4. First, examples that have high confidence in either true class or top2 probability represent far away from other classes. In other words, these examples locate around the center of class of either true or top2. In this case, Eq. 5 does not hold true between high and low confidence examples (see Fig. 2). Therefore it is not a big deal even if we define margins with Eq.(4). On the other hand, low confidence examples in both true and top2 probability may be around the decision boundaries of multi-class. These examples satisfy Eq. 5 for not only both probability extremely equal examples but also different probability examples (e.g., $p_{y_i}(\boldsymbol{x}_i) < p_{y_j}(\boldsymbol{x}_j)$). For instance, as shown in Fig. 2, we can obtain approximately equal margins for examples within [0.2, 0.6]. This situation represents

$\boldsymbol{x}_i$ is closer to the intersection of multiple class boundaries and has a higher attack risk than $\boldsymbol{x}_j$. Hence, we argue that making equal the margin to these examples is inappropriate. Although we expect to improve robustness if we could carefully distinguish these margins, both WMMR and MAIL do not take into account the aforementioned.

3.2 Quantifying the importance for margins

Previous margin determination is insufficient when we recall that Lemma 1 holds for different probabilities among some examples. Intuitively, a straightforward approach for resolving this issue computes margins between true class and other classes with Eq. (4). However, this approach is hard to realize because we need to carefully design the weighting function to transform multi-class margins to the appropriate weight.

Definition 2 (Incorrect rate)**.** *Let $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}$ be the input data and let $p_{y_i}(\boldsymbol{x}_i) \in [0,1]$ be the correct class probability. Then, the incorrect rate is defined as*

$$\begin{aligned} \bar{p}_{y_i}(\boldsymbol{x}_i) &= 1 - p_{y_i}(\boldsymbol{x}_i) \\ &= \sum\nolimits_{k \neq y_i} p_k(\boldsymbol{x}_i). \end{aligned} \tag{6}$$

Thus, we propose a metric, which quantifies relations among the most confusing class and other classes by using top2 probability and the incorrect rate. Suppose Lemma 1 holds for different confidence examples, we can derive following the inequality for a ratio of top2 probability in the incorrect rate.

Proposition 1. *For input data $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$, let $m(\boldsymbol{x}_i, y_i) = m(\boldsymbol{x}_j, y_j)$. Moreover, let $p_{y_i}(\boldsymbol{x}_i) > p_{y_j}(\boldsymbol{x}_j)$ and $p_{y_j}(\boldsymbol{x}_j) > \frac{1}{K}$. Then the following inequality for the top2 ratio in the incorrect rate of each sample holds:*

$$o(\boldsymbol{x}_i, y_i) > o(\boldsymbol{x}_j, y_j), \ \text{ where } o(\boldsymbol{x}, y) = \frac{p_2(\boldsymbol{x})}{\bar{p}_y(\boldsymbol{x})}. \tag{7}$$

Eq.(19) represents different for the ratio of top2 in the incorrect rate depending on the probabilities of each example when margins are equal with different true/top2 probability. According to the previous idea (i.e., Eq.(4) implies the margin between the examples and the decision boundaries), we identify that a place of example with a low ratio of top2 distributes around the intersection of multiple classes. Thus, we expect that our doubt resolves by applying the ratio of top2 to the margin computed by Eq.(4).

Meanwhile, we require to discuss a range of the ratio of top2 $o(\boldsymbol{x}_i, y_i)$. Let $P_{\text{top2}} := \left(\frac{\bar{p}_{y_i}(\boldsymbol{x}_i)}{K-1}, \bar{p}_{y_i}(\boldsymbol{x}_i)\right)$ be range of $p_2(\boldsymbol{x}_i)$, and the upper/lower bound with directly using $o(\boldsymbol{x}_i, y_i)$ defines as bellow:

$$\begin{cases} \sup\limits_{p_2(\boldsymbol{x}_i) \in P_{\text{top2}}} o(\boldsymbol{x}_i, y_i) = 1 & p_2(\boldsymbol{x}_i) = \bar{p}_{y_i}(\boldsymbol{x}_i), \\ \inf\limits_{p_2(\boldsymbol{x}_i) \in P_{\text{top2}}} o(\boldsymbol{x}_i, y_i) = \frac{1}{K-1} & p_2(\boldsymbol{x}_i) = \frac{\bar{p}_{y_i}(\boldsymbol{x}_i)}{K-1}. \end{cases} \tag{8}$$

According to Eq.(8), directly using the ratio of top2 is contrary behavior to our motivations because we confirm $o(\boldsymbol{x}_i, y_i)$ approaches the upper/lower bound when $p_2(\boldsymbol{x}_i)$ is large/small. Hence, we can obtain the appropriate range by inverting the ratio of top2 with an arbitrary coefficient $\tau \in \mathbb{N}$. However, we cannot say the appropriate representation because $\tau = 1$ computes the upper/lower bound as shown below:

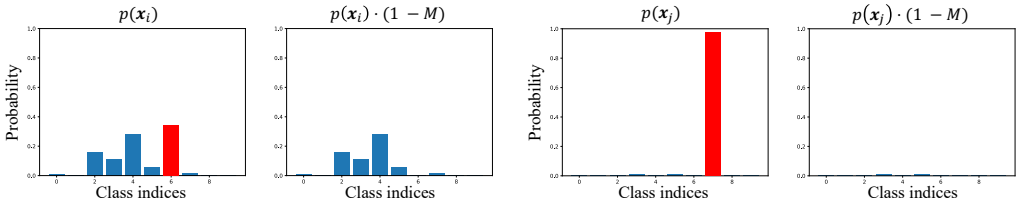
$$\begin{cases} \sup\limits_{p_2(\boldsymbol{x}_i) \in P_{\text{top2}}} (\tau - o(\boldsymbol{x}_i, y_i)) = 1 - \frac{1}{K-1} & p_2(\boldsymbol{x}_i) = \frac{\bar{p}_{y_i}(\boldsymbol{x}_i)}{K-1}, \\ \inf\limits_{p_2(\boldsymbol{x}_i) \in P_{\text{top2}}} (\tau - o(\boldsymbol{x}_i, y_i)) = 0 & p_2(\boldsymbol{x}_i) = \bar{p}_{y_i}(\boldsymbol{x}_i). \end{cases} \tag{9}$$

Eq.(9) implies that examples with almost top2 probability in the incorrect rate (i.e., significantly misclassified adversarial examples) no longer involve the updating parameters.

Therefore, we define weights to margins computed by Eq.(4) with the following equation:

$$s(\boldsymbol{x}_i, y_i) = \tau - o(\boldsymbol{x}_i, y_i), \ \text{ s.t. } \tau \in \mathbb{N}_{\geq 2}, \tag{10}$$

where $\mathbb{N}_{\geq 2}$ indicates the natural number set over 2.



(a) Low confidence example, $c(\boldsymbol{x}_i, y_i) = 0.7083$ (b) High confidence example, $c(\boldsymbol{x}_j, y_j) = 0.0143$

Figure 3: These results show computing cosine similarity with Eq.(11) for predicted distributions of two examples. The red bar indicates the true class of each example. (a) and (b) are the successful classification examples with low and high probability, respectively.

3.3 Representation the importance for the incorrect rate

Although we enable to represent whether examples are close to the intersection of multiple classes or not in Section 3.2, Eq.(10) cannot transform to weights of our expectation yet. Specifically, we obtain approximately the same amount of weights as around the boundaries when incorrect class probabilities are uniformly equal in spite of successful classification with high confidence. Because these examples require the definition of different weights, the former argument implies the importance of the incorrect rate is also different. Therefore, we represent the importance of the incorrect rate leveraging predicted probability distributions by the following equation:

$$c(\boldsymbol{x}_i, y_i) = \frac{< p(\boldsymbol{x}_i), p'(\boldsymbol{x}_i) >}{\|p(\boldsymbol{x}_i)\|_2 \cdot \|p'(\boldsymbol{x}_i)\|_2} \tag{11}$$

where $p(\boldsymbol{x}_i) \in [0, 1]^K$ and $< \cdot, \cdot >$ are the predicted probability distribution of $\boldsymbol{x}_i$ and the dot product, respectively. $p'(\boldsymbol{x}_i)$ indicates the distribution eliminated the true class and computes using the binary mask, which except for the true class is 1:

$$p'(\boldsymbol{x}_i) := p(\boldsymbol{x}_i) \cdot (1 - M), \ \text{ s.t. } \sum\nolimits_{k=1}^{K} p'_k(\boldsymbol{x}_i) < 1. \tag{12}$$

As shown in Fig. 3(a), the similarity tends to be large for the example with low confidence. Meanwhile, as shown in Fig. 3(b), we confirm that the similarity tends to be small for the example with high confidence. The trend obtains almost equal results on all validation examples and we confirm that the similarity has correlated to the incorrect rate (see Fig. 4 in Appendix). Thus, we enable to assign appropriate weights to margins computed by Eq.(4) by applying Eq.(12) to Eq.(10). Based on these reasons, we transform Eq.(10) to the appropriate representation with

$$s'(\boldsymbol{x}_i, y_i) = s(\boldsymbol{x}_i, y_i) \cdot c(\boldsymbol{x}_i, y_i). \tag{13}$$

3.4 Incorporating our approach to previous IRAT

In previous sections, we reveal the previous approach is insufficient and propose *reweighting for the margin* to transform margins computed by Eq.(4) to adequate representations with the ratio of top2. We describe how to incorporate the proposed method into MAIL and WMMR in this section.

We multiply the ratio of top2 computed by Eq.(13) and the margins of Eq.(4);

$$\tilde{m}(\boldsymbol{x}_i, y_i) = m(\boldsymbol{x}_i, y_i) \cdot s'(\boldsymbol{x}_i, y_i). \tag{14}$$

Then, we transform margins to boost the importance of examples around the intersection of the decision boundaries of multiple classes with Eq.(14) to weights with previous functional designs;

$$\omega_{\text{WMMR}} = \exp(-\tilde{m}(\boldsymbol{x}_i, y_i)), \tag{15}$$

$$\omega_{\text{MAIL}} = \text{sigmoid}(-\gamma \cdot (\tilde{m}(\boldsymbol{x}_i, y_i) + \beta)). \tag{16}$$

β and γ are hyper-parameters, and we use constants of the previous method in the experiment. Therefore, we should determine only τ in the proposed method.

We omit the constant of WMMR by replacing the proposed method because it just multiplies margins computed by Eq.(4). We are a warmup the model with $\omega = 1$ for all examples during the initial

training. In other words, the training within a set number of epochs is standard AT. This technique uses in GAIRAT or MAIL, which aims to avoid inappropriate weight representations during the initial training. The training procedure incorporated the proposed method show in Algorithm 1 in Appendix.

4 EXPERIMENTS

This section demonstrates performances incorporating the proposed method to WMMR and MAIL. In this experiment, we evaluate the standard accuracy of benign data and adversarial robustness against several attacks on the CIFAR-10/100 (Krizhevsky & Hinton, 2009) dataset. First, we describe training settings and evaluation of adversarial robustness details in Section 4.1. After that, we report quantitative evaluation results in Section 4.2, and then discuss the hyper-parameter and necessity of the similarity computation in Section 4.3.

4.1 EXPERIMENTAL DETAILS

This section describes experimental details. We use standard AT (Standard) (Madry et al., 2018), WMMR, and MAIL as comparison methods. Evaluation results use the average and the standard deviation of 5 models trained with different random seeds. Moreover, we evaluate with the highest performance model during the training (Best model) and at the end of the training model (Last model). The results of other methods and best models show in the Appendix due to the limited number of pages. For a fair evaluation, the training settings with WRN34-10 on the CIFAR-10 dataset integrate based on (Pang et al., 2021).

Training Settings: We use ResNet-18 (He et al., 2016) and WideResNet34-10 (WRN34-10) (Zagoruyko & Komodakis, 2016) as base network. We train models during 120 epochs using the batch size is 128 and SGD with a momentum is 0.9. Initial learning rates for each network are 0.01 (for ResNet-18)/0.1 (for WRN34-10) and divide by 1/10 at $\{75, 90, 100\}$ epochs. The weight decay sets 3.5×10^{-3} (for ResNet-18)/0.5×10^{-4} (for WRN34-10). Moreover, we use PGD with the perturbation tolerance $\epsilon = 8/255$, the step size $\alpha = \epsilon/4$, and the number of steps of PGD $N = 10$ for all methods.

Hyper-parameters: MAIL uses $\beta = 0.5$ and $\gamma = 10$ for all models and datasets. Meanwhile, the constant of WMMR uses $\alpha = 2$. These constants use the same settings regardless of and without the proposed method. The coefficient of the proposed method uses $\tau = 2$ for WRN34-10 on both CIFAR-10/100 dataset, and $\tau = 2$ on CIFAR-10 and $\tau = 3$ on CIFAR-100 for ResNet-18. Additionally, the number of warmups sets 75 epochs, and then we activate the weighted loss minimization.

Adversarial Robustness: We evaluate not only standard accuracy to benign examples (Clean) but also adversarial robustness against PGD-K with $K = 100$ (PGD-100), PGD-100 using the loss function of CW, AutoPGD with the cross entropy (APGD-CE) (Croce & Hein, 2020), and AutoAttack (AA) (Croce & Hein, 2020). As mentioned in the article (Wang et al., 2021), we evaluate the performance focused on the white-box attack because the black-box attack supposes easy to defend.

4.2 RESULTS

We list the results on CIFAR-10 in Table 1 and on CIFAR-100 in Table 2.

CIFAR-10: Results of ResNet-18 listed in the upper of Table 1 show improved adversarial robustness against PGD-100 and APGD-CE by incorporating the proposed method to MAIL. Moreover, the result of Clean improved the performance a bit. The results incorporated into WMMR observed the same trend. Then, focused on the results of WRN34-10 listed at the bottom of Table 1, incorporating the proposed method to WMMR get more excellent adversarial robustness against all attacks than the original WMMR. In particular, the average robust accuracy of AA improved by 0.13 and the standard deviation of 5 models was significantly small. Meanwhile, the classification accuracy of benign examples of ours+MAIL is comparable and it significantly improved the adversarial robustness of the PGD family, such as PGD-100 and APGD-CE.

Table 1: Performances of the last model of each method on CIFAR-10 (%). The upper results are ResNet-18, and the bottom results are WRN34-10. The bold symbol indicates the best performing method and ↑ indicates improved performance by incorporating the proposed method.

	Clean	PGD-100	CW-PGD	APGD-CE	AA
Standard	85.69±0.15	48.86±0.22	**49.26±0.20**	48.34±0.23	**46.38±0.26**
WMMR	**85.71±0.17**	49.60±0.41	48.31±0.36	48.99±0.38	45.23±0.37
MAIL	82.53±0.30	56.66±0.17	47.71±0.26	55.75±0.13	45.22±0.31
Ours + WMMR	85.65±0.13	50.31±0.25↑	47.94±0.13	49.59±0.29↑	44.79±0.18
Ours + MAIL	82.66±0.10↑	**57.96±0.13**↑	47.24±0.16	**56.95±0.19**↑	44.78±0.29
Standard	87.89±0.17	48.11±0.50	49.33±0.45	47.66±0.47	46.82±0.38
WMMR	**88.03±0.12**	49.33±0.41	49.95±0.33	48.88±0.43	47.37±0.38
MAIL	86.40±0.13	59.30±0.21	**51.82±0.22**	58.57±0.26	**49.29±0.14**
Ours+WMMR	87.99±0.14	49.39±0.19↑	49.97±0.22↑	48.96±0.18↑	47.50±0.17↑
Ours+MAIL	86.48±0.24↑	**60.56±0.42**↑	51.55±0.30	**59.64±0.45**↑	48.92±0.30

Table 2: Performances of the last model of each method on CIFAR-100 (%). The upper results are ResNet-18, and the bottom results are WRN34-10. The bold symbol indicates the best performing method and ↑ indicates improved performance by incorporating the proposed method.

	Clean	PGD-100	CW-PGD	APGD-CE	AA
Standard	60.10±0.22	28.65±0.21	**27.89±0.17**	28.12±0.18	**25.01±0.13**
WMMR	60.42±0.28	28.14±0.16	26.66±0.18	27.47±0.15	23.59±0.14
MAIL	56.77±0.19	**31.23±0.08**	25.68±0.26	30.45±0.09	22.99±0.09
Ours + WMMR	**60.50±0.18**↑	28.24±0.12↑	26.96±0.20↑	27.58±0.14↑	23.86±0.15↑
Ours + MAIL	57.50±0.23↑	**31.23±0.16**	25.30±0.09	**30.46±0.17**↑	22.61±0.10
Standard	62.63±0.33	24.82±0.12	25.80±0.17	24.53±0.13	23.70±0.13
WMMR	63.32±0.12	25.87±0.22	26.29±0.20	25.45±0.19	23.81±0.19
MAIL	62.56±0.21	34.29±0.08	**29.31±0.19**	**33.58±0.09**	**26.50±0.16**
Ours+WMMR	**63.76±0.09**↑	25.45±0.24	26.51±0.15↑	25.14±0.21	23.94±0.18↑
Ours+MAIL	63.19±0.21↑	**34.32±0.13**↑	28.94±0.18	33.47±0.18	26.18±0.14

CIFAR-100: Focused on the results of ResNet-18 listed at the upper of Table 2, ours+MAIL improved the accuracy to benign examples by 0.73 and the robust performance of PGD-100 and APGD-CE were comparable. Meanwhile, ours+WMMR achieved excellent performance to not only Clean but also all attacks over 0.10. For WRN34-10, as listed at the bottom of Table 2, ours+WMMR achieved an improvement of the accuracy to benign examples by 0.44. In addition, ours+WMMR improved adversarial robustness against both CW-PGD and AA, but the robustness of the PGD family slightly degenerated, unfortunately. Meanwhile, the results of ours+MAIL improved Clean by 0.63, and the robustness of PGD-100 was comparable. These results showed the proposed method was efficacy to improve the accuracy of benign examples and obtain comparable or outperformed adversarial robustness on the CIFAR-100.

The results of the best model of each dataset and base network discuss in Appendix E due to limited the number of pages. In the nutshell, their results observed the same trend as the results of the last model.

4.3 Ablation Study

This section discusses how the effect of the performance when we eliminate the similarity of Eq.(11) or use different hyper-parameters τ. This experiment compares five models of different random seeds the average and the standard deviation for WRN34-10 with configure of Section 4.1.

First, Table 3 shows the results of performance comparison with different hyper-parameter τ. This experiment compares the performance of $\tau = \{2, 3, 4\}$. The results on CIFAR-10 listed at the upper of Table 3 showed that the large τ significantly increased the robustness of APGD-CE. Meanwhile, Clean and CW-PGD degenerated the performance according to the increasing τ. We observed that the large τ increased classical PGD robustness and degenerated robustness of AA (see Appendix E). As listed at the bottom of Table 3, the results on CIFAR-100 showed the large τ increased accuracy of

Table 3: Ablation study for the hyper-parameter τ. The left and right 3 results are on CIFAR-10/100, respectively.

	CIFAR-10			CIFAR-100		
	Clean	CW-PGD	APGD-CE	Clean	CW-PGD	APGD-CE
$\tau = 2$	**86.48±0.24**	**51.55±0.30**	59.64±0.45	63.19±0.21	**28.94±0.18**	**33.47±0.18**
$\tau = 3$	86.39±0.14	50.65±0.35	60.77±0.48	63.44±0.22	27.79±0.25	32.65±0.17
$\tau = 4$	86.37±0.25	50.49±0.50	**62.09±0.47**	**63.62±0.29**	26.87±0.23	31.38±0.19

benign examples but it degenerated robustness against adversarial attacks. In particular, robustness degenerated by over 1.0 on CIFAR-10 and 2.0 on CIFAR-100. Thus, $\tau = 2$ was appropriate with WRN34-10 on both datasets. The results of ResNet-18 showed Table 8 in Appendix E.

Table 4: Ablation study for the similarity of Eq.(11) on the CIFAR-10 dataset. These results indicate the performance of the last model. sim.: similarity.

	w/ sim.	w/o sim.
Clean	**86.48±0.24**	86.40±0.09
CW-PGD	**51.55±0.30**	51.07±0.23
APGD-CE	59.64±0.45	**59.91±0.42**

Next, the results (i.e., Clean, CW-PGD, and APGD-CE) on CIFAR-10 lists in Table 4 when eliminated the similarity of Eq.(11) from the proposed method. While the proposed method without the similarity degenerated Clean and CW-PGD, it slightly improved the robustness of APGD-CE. As shown in Table 7 in Appendix E, the robustness of PGD was comparable by eliminating the similarity but the robustness of AA did not decrease. Although classical PGD is one of the strong attacks, evaluating the quality of the model performance just PGD links overestimation of robustness. Therefore, for the proposed method it is appropriate to use the similarity because of the comparable robustness of PGD despite improving the performance of CW-PGD or AA.

5 Discussion and Limitation

Surprisingly, standard AT achieved the most excellent robustness against AA and CW-PGD in the small capacity network such as ResNet-18. This result implies that MAIL makes vulnerable to strong attacks when the network does not have enough capacity. Therefore, our approach may accelerate the inborn vulnerability of MAIL in return for dramatically improving the robustness of the PGD attack. Meanwhile, WMMR with the proposed method improved the performance for almost the attacks, including in AA or CW-PGD, compared with the original despite just multiplying Eq.(13 to margins. Instead of the constant coefficient, our approach can apply different coefficients to the margins of each instance during the training. This result show necessity to transform margins to appropriate representation because the margins transformed by our approach directly contribute to the training without the constant coefficient.

For the results of each dataset, the performance improvement on CIFAR-100 was limited in the experiment compared to CIFAR-10. Intuitively, although increasing classes should consider our idea because it has complex decision boundaries, the proposed method just indirectly deals with relations among multi-class yet. Hence, considering the margin of multi-class inherently may make networks robust on CIFAR-100 or more large-scale datasets and we leave this as future work.

6 Conclusion

In this paper, we reveal a weakness of the previous instance-reweighted adversarial training (IRAT) and propose a novel metric to resolve the weakness. Unlike previous methods that define weight by considering only two probabilities, our approach successfully boosts the performance by applying the ratio of the top2 probability in the incorrect rate to the previous method. However, robustness improvement on CIFAR-100 compared with CIFAR-10 results is limited. Image classification in the real world supposes dealing with more complex data and vast classes than CIFAR-10/100. Therefore, we aim the performance improvement on large-scale datasets by carefully designing the weighting function in the future.

REFERENCES

Qi-Zhi Cai, Chang Liu, and Dawn Song. Curriculum adversarial training. In *International Joint Conference on Artificial Intelligence*, pp. 3740–3747, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness. In *International Joint Conference on Artificial Intelligence*, 2022.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, volume 119, pp. 2206–2216, 2020.

Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *IEEE/CVF International Conference on Computer Vision*, pp. 15721–15730, 2021.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

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*, pp. 1–28, 2020.

Ruize Gao, Feng Liu, Kaiwen Zhou, Gang Niu, Bo Han, and James Cheng. Local reweighting for adversarial training. arXiv preprint, 2021.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, pp. 1–11, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Minseon Kim, Jihoon Tack, Jinwoo Shin, and Sung Ju Hwang. Entropy weighted adversarial training. In *International Conference on Machine Learning Workshop*, 2021.

Alex Krizhevsky and Geoffery Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Daniel Lowd and Christopher Meek. Adversarial learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 641–647, 2005.

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*, pp. 1–23, 2018.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, pp. 1–21, 2021.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pp. 582–597, 2016.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, volume 119, pp. 8093–8104, 2020.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, pp. 1–17, 2018.

Chuanbiao Song, Yanbo Fan, Yichen Yang, Baoyuan Wu, Yiming Li, Zhifeng Li, and Kun He. Regional adversarial training for better robust generalization. arXiv preprint arXiv: 2109.00678, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, pp. 1–10, 2014.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, pp. 12214–12223, 2019.

Qizhou Wang, Feng Liu, Bo Han, Tongliang Liu, Chen Gong, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Probabilistic margins for instance reweighting in adversarial training. In *Advances in Neural Information Processing Systems*, pp. 1–12, 2021.

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*, pp. 1–14, 2020.

Dongxian Wu, Shu tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pp. 2958–2969, 2020.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*, 2018.

Chaojian Yu, Bo Han, Mingming Gong, Li Shen, Shiming Ge, Bo Du, and Tongliang Liu. Robust weight perturbation for adversarial training. In *International Joint Conferences on Artificial Intelligence*, pp. 3688–3694, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pp. 87.1–87.12, 2016.

Huimin Zeng, Chen Zhu, Tom Goldstein, and Furong Huang. Are adversarial examples created equal? a learnable weighted minimax risk for robustness under non-uniform attacks. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 10815–10823, 2021.

Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1829–1839, 2019.

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*, volume 97, pp. 7472–7482, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, volume 119, pp. 11278–11287, 2020.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, pp. 1–29, 2021.

A PROOF OF LEMMA 1

Lemma 1. *For the margin calculated by Eq.(4) and let $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training set, there exist samples satisfy bellow*

$$m(\boldsymbol{x}_i, y_i) = m(\boldsymbol{x}_j, y_j). \tag{17}$$

Proof:

We divide the training dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ to the successful classified samples set $\mathcal{S}^+ := \{(\boldsymbol{x}_i, y_i) \in \mathcal{D} \mid \arg\max p(\boldsymbol{x}_i) = y_i\}$ and the misclassified samples set $\mathcal{S}^- := \{(\boldsymbol{x}_i, y_i) \in \mathcal{D} \mid \arg\max p(\boldsymbol{x}_i) \neq y_i\}$. First, in common both $\mathcal{S}^+$ and $\mathcal{S}^-$, Lemma 1 obviously holds when $p_{y_i}(\boldsymbol{x}_i) = p_{y_j}(\boldsymbol{x}_j)$. Next, we prove that Lemma 1 also satisfies for different probabilities, i.e., $p_{y_i}(\boldsymbol{x}_i) \neq p_{y_j}(\boldsymbol{x}_j)$. Because all samples including in $\mathcal{S}^+$ and $\mathcal{S}^+$ absolutely satisfy $p_{y_i}(\boldsymbol{x}_i) > p_2(\boldsymbol{x_i})$ and $p_{y_i}(\boldsymbol{x}_i) < p_2(\boldsymbol{x_i})$, Lemma 1 holds for satisfying the following equation:

$$|p_{y_i}(\boldsymbol{x}_i) - p_{y_j}(\boldsymbol{x}_j)| = |p_2(\boldsymbol{x}_i) - p_2(\boldsymbol{x}_j)|, \tag{18}$$

Hence, equal or approximate equal margins exist in spite of different true class and top2 probability. □

B PROOF OF PROPOSITION 1

Proposition 1. *For input data $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$, let $m(\boldsymbol{x}_i, y_i) = m(\boldsymbol{x}_j, y_j)$. Moreover, let $p_{y_i}(\boldsymbol{x}_i) > p_{y_j}(\boldsymbol{x}_j)$ and $p_{y_j}(\boldsymbol{x}_j) > \frac{1}{K}$. Then the following inequality for the top2 ratio in the incorrect rate of each sample holds:*

$$o(\boldsymbol{x}_i, y_i) > o(\boldsymbol{x}_j, y_j), \text{ where } o(\boldsymbol{x}, y) = \frac{p_2(\boldsymbol{x})}{\bar{p}_y(\boldsymbol{x})}. \tag{19}$$

Proof: Suppose Lemma 1 holds under $p_{y_i}(\boldsymbol{x}_i) > p_{y_j}(\boldsymbol{x}_j)$ for each true class probabilities of input data $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$, top2 probabilities obviously holds the following inequality:

$$p_2(\boldsymbol{x}_i) > p_2(\boldsymbol{x}_j). \tag{20}$$

Hence, the incorrect rate of each data satisfy

$$\bar{p}_{y_i}(\boldsymbol{x}_i) < \bar{p}_{y_j}(\boldsymbol{x}_j). \tag{21}$$

Because we can guarantee that top2 probability does not overshoot the incorrect rate in any case, the following inequality holds:

$$\frac{p_2(\boldsymbol{x}_i)}{\bar{p}_{y_i}(\boldsymbol{x}_i)} > \frac{p_2(\boldsymbol{x}_j)}{\bar{p}_{y_j}(\boldsymbol{x}_j)}, \tag{22}$$

Hence, we obtain $o(\boldsymbol{x}_i, y_i) > o(\boldsymbol{x}_j, y_j)$ by using $o(\boldsymbol{x}, y) = \frac{p_2(\boldsymbol{x})}{\bar{p}_y(\boldsymbol{x})}$. □

C TRAINING PROCESS PROPOSED ADVERSARIAL TRAINING

AT incorporating the proposed method trains along Algorithm 1. We activate the assigning weights to loss after Ω epochs because the initial model is inappropriate for weighting. Therefore, we use $\omega = 1$ (i.e., the line 15 in Algorithm 1) less than Ω epochs. Moreover, computed weights multiply the coefficient M instead of directly applying it to loss. Despite multiplying the coefficient M by weights in (Wang et al., 2021), the authors do not discuss the necessity of this process. Experiments in the paper use $M = 3$.

Algorithm 1 Adversarial training incorporated our approach

Require: Training dataset $\mathcal{D}$, batch size n, training epochs T, learning rate η, model parameter $\boldsymbol{\theta}$, the number of warm-up Ω
Require: The function deriving adversarial perturbation $\mathcal{A}$
Require: hyper-parameters β, γ, τ

for $t = 1, \ldots, T$ **do**
  for $\{\boldsymbol{x}_i, y_i | i = 1, \ldots, n\} \sim \mathcal{D}$ **do**
    $\hat{\boldsymbol{x}}_i \leftarrow \mathcal{A}(\boldsymbol{x}_i, y_i; \boldsymbol{\theta})$
    $s(\hat{\boldsymbol{x}}_i, y_i) = \tau - o(\hat{\boldsymbol{x}}_i, y_i)$
    $c(\hat{\boldsymbol{x}}_i, y_i) = \frac{<p(\hat{\boldsymbol{x}}_i), p'(\hat{\boldsymbol{x}}_i)>}{\|p(\hat{\boldsymbol{x}}_i)\|_2 \cdot \|p'(\hat{\boldsymbol{x}}_i)\|_2}$
    $s'(\boldsymbol{x}_i, y_i) = s(\boldsymbol{x}_i, y_i) \cdot c(\boldsymbol{x}_i, y_i)$
    $\tilde{m}(\hat{\boldsymbol{x}}_i, y_i) = m(\hat{\boldsymbol{x}}_i, y_i) \cdot s'(\hat{\boldsymbol{x}}_i, y_i)$
    if $T > \Omega$ **then**
      if MAIL **then**
        $\omega_i = \mathrm{sigmoid}(\gamma \cdot (\tilde{m}(\hat{\boldsymbol{x}}_i, y_i) + \beta)) \times M$
      else if WMMR **then**
        $\omega_i = \exp(-\tilde{m}(\hat{\boldsymbol{x}}_i, y_i)) \times M$
      end if
    else
      $\omega_i = 1$
    end if
    model update:
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \omega_i \cdot \ell(f_{\boldsymbol{\theta}_t}(\tilde{\boldsymbol{x}}_i), y_i)$
  end for
end for
return model parameter $\boldsymbol{\theta}$

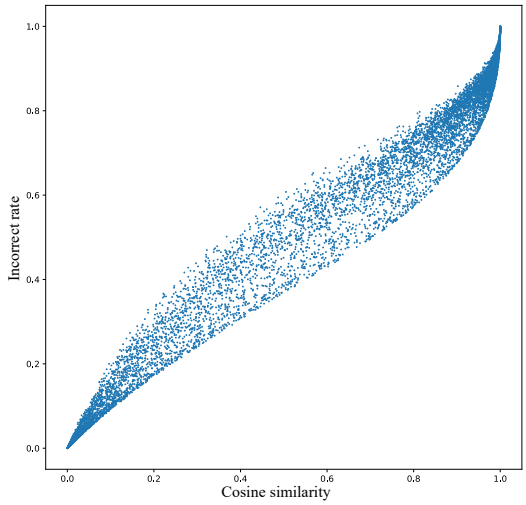


Figure 4: Plotting the similarity computed by Eq.(12) and the incorrect rate for WRN34-10 trained with standard AT on CIFAR-10. Between these two measures have extremely large correlation because the correlation coefficient is 0.98 computed by Eq.(23).

D Correlation between incorrect rate and similarity

This section discusses the correlation between similarity and the incorrect rate introduced in Section D. In this experiment, we compute the correlation between two measures for WRN34-10 trained

Table 5: Performance comparison results of the best model on CIFAR-10. The upper and bottom results are ResNet-18 and WRN34-10, respectively. The bold symbol is best performed method and the underline is successful improvement robustness by introducing the proposed method.

		Clean	FGSM	PGD-100	CW-PGD	APGD-CE	AA
ResNet-18	Standard	83.84±0.171	57.71±0.17	50.98±0.21	50.50±0.31	50.55±0.22	47.86±0.13
	GAIRAT	83.63±0.06	61.13±0.35	56.62±0.46	37.72±0.66	53.37±0.47	34.76±0.66
	EWAT	**83.97±0.52**	57.41±0.51	50.56±0.31	**50.51±0.50**	50.14±0.34	**48.03±0.48**
	WMMR	83.74±0.23	58.05±0.16	52.33±0.21	49.12±0.12	51.70±0.25	46.57±0.12
	MAIL	82.37±0.23	60.41±0.08	56.99±0.20	47.90±0.19	56.07±0.23	45.44±0.18
	Ours + WMMR	83.87±0.27	58.37±0.17	52.91±0.16	48.52±0.24	52.17±0.16	46.04±0.24
	Ours + MAIL	82.66±0.09	**61.74±0.25**	**58.36±0.30**	47.43±0.23	**57.24±0.30**	44.82±0.36
WRN34-10	Standard	86.67±0.21	61.33±0.15	54.17±0.23	54.05±0.15	53.79±0.22	51.73±0.32
	GAIRAT	85.74±0.90	63.50±0.38	60.45±0.18	45.85±2.48	57.29±0.85	43.12±2.42
	EWAT	85.93±0.46	60.98±0.62	54.17±0.24	**54.27±0.19**	53.79±0.25	**51.91±0.16**
	WMMR	86.57±0.18	61.30±0.23	55.16±0.37	52.76±0.39	54.64±0.35	50.36±0.33
	MAIL	86.28±0.12	63.60±0.02	59.77±0.21	51.91±0.29	58.93±0.26	49.47±0.33
	Ours + WMMR	**86.91±0.31**	61.36±0.32	55.00±0.25	53.27±0.08	54.49±0.30	50.73±0.12
	Ours + MAIL	86.45±0.19	**64.66±0.14**	**61.05±0.25**	51.87±0.17	**60.08±0.21**	49.26±0.22

Table 6: Performance comparison results of the best model on CIFAR-100. The upper and bottom results are ResNet-18 and WRN34-10, respectively. The bold symbol is best performed method and the underline is successful improvement robustness by introducing the proposed method.

		Clean	FGSM	PGD-100	CW-PGD	APGD-CE	AA
ResNet-18	Standard	57.48±0.50	33.09±0.16	29.51±0.06	**27.76±0.13**	28.71±0.55	**25.16±0.14**
	GAIRAT	56.91±0.29	32.66±0.06	29.16±0.10	27.30±0.09	28.61±0.13	24.69±0.10
	EWAT	57.75±0.32	32.87±0.40	29.35±0.19	27.20±0.17	28.88±0.18	25.03±0.16
	WMMR	57.82±0.45	32.51±0.25	29.03±0.16	26.84±0.20	28.34±0.21	24.02±0.27
	MAIL	56.65±0.18	**33.49±0.15**	**31.36±0.07**	25.72±0.18	**30.62±0.09**	23.02±0.17
	Ours + WMMR	**58.07±1.35**	32.82±0.49	28.94±0.09	27.03±0.28	28.30±0.13	24.19±0.04
	Ours + MAIL	57.59±0.12	33.47±0.11	31.25±0.16	24.80±0.24	30.44±0.17	22.06±0.24
WRN34-10	Standard	62.21±0.19	35.58±0.31	31.43±0.31	**30.51±0.27**	31.00±0.27	**27.96±0.10**
	GAIRAT	61.83±0.63	35.49±0.20	31.48±0.15	30.44±0.19	31.00±0.14	27.85±0.14
	EWAT	62.18±0.28	34.60±0.27	29.79±0.22	29.07±0.38	29.39±0.28	26.75±0.30
	WMMR	62.08±0.16	34.64±0.23	30.99±0.12	29.09±0.17	30.50±0.17	26.61±0.24
	MAIL	62.36±0.21	**37.26±0.12**	**34.64±0.12**	29.42±0.13	**33.94±0.14**	26.67±0.08
	Ours + WMMR	62.51±0.64	35.30±0.20	31.15±0.29	29.91±0.22	30.73±0.31	27.22±0.35
	Ours + MAIL	**63.00±0.11**	37.12±0.24	34.56±0.18	29.07±0.16	33.78±0.19	26.38±0.21

by standard AT with the following equation:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{p}_{y_i}(\boldsymbol{x}_i) - \bar{p}')(c(\boldsymbol{x}_i, y_i) - c')}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{p}_{y_i}(\boldsymbol{x}_i) - \bar{p}')^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c(\boldsymbol{x}_i, y_i) - c')^2}}, \tag{23}$$

where n is the number of samples, $\bar{p}' = \frac{1}{n}\sum_{i=1}^{n}\bar{p}_{y_i}(\boldsymbol{x}_i)$, and $c' = \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{x}_i, y_i)$.

As shown in Fig. 4, the correlation exists between the two measures visually. Quantitatively, we observed a strong correlation because of $r = 0.98$. Therefore, we enable to mitigate the inappropriate behavior of the ratio of the top2 for correct samples with high confidence by relying on the similarity.

E Additional results

This section discusses the experimental results excepted from the main paper due to limited the number of pages. Experimental settings are the same as Section 4.1. Additionally, we add GAIRAT and EWAT as comparison methods and FGSM as the attack.

Table 5, 6 lists adversarial robustness and classification accuracy for the best model results of each method on CIFAR-10/100. The best model in the experiment is the checkpoint that achieved the most excellent robustness against PGD-20 with the validation data. For adversarial robustness, we

Table 7: Ablation study for the similarity of Eq.(11) on CIFAR-10. The upper and bottom results are the last and best model, respectively. sim.: similarity.

		Clean	FGSM	PGD-100	CW-PGD	APGD-CE	AA
Last	w/ sim.	**86.48±0.24**	64.52±0.20	60.56±0.42	**51.55±0.30**	59.64±0.45	**48.92±0.30**
	w/o sim.	86.40±0.09	64.52±0.28	**60.88±0.36**	51.07±0.23	**59.91±0.42**	48.73±0.23
Best	w/ sim.	**86.45±0.19**	64.66±0.14	61.05±0.25	**51.87±0.17**	60.08±0.21	**49.26±0.22**
	w/o sim.	86.27±0.17	**64.88±0.22**	**61.35±0.17**	51.45±0.21	**60.39±0.17**	48.77±0.26

Table 8: Ablation study for the hyper-parameter τ for ResNet-18. These results are the performance of the last model. The upper and bottom 3 results are on CIFAR-10 and CIFAR-100, respectively.

		Clean	FGSM	PGD-100	CW-PGD	APGD-CE	AA
CIFAR-10	$\tau = 2$	82.66±0.10	61.46±0.20	57.96±0.13	**47.24±0.16**	56.95±0.19	**44.78±0.29**
	$\tau = 3$	**82.78±0.16**	61.52±0.47	57.81±0.49	47.07±0.18	56.71±0.48	44.54±0.30
	$\tau = 4$	82.35±0.32	**63.66±0.19**	**60.64±0.21**	45.54±0.39	**59.30±0.20**	42.64±0.39
CIFAR-100	$\tau = 2$	56.80±0.26	33.43±0.14	31.14±0.18	**25.81±0.13**	30.36±0.14	**23.09±0.16**
	$\tau = 3$	57.50±0.23	**33.47±0.11**	**31.23±0.16**	25.30±0.09	**30.46±0.17**	22.61±0.10
	$\tau = 4$	**57.78±0.16**	33.38±0.08	30.93±0.05	24.68±0.11	30.15±0.06	21.83±0.05

improved the performance of the standard accuracy and some attacks by 0.32-0.58 (for ResNet-18) and 0.06-0.51 (for WRN34-10) by ours+WMMR. Ours+MAIL improved the performance of the standard accuracy and some attacks by 1.17-1.37 (for ResNet-18) and 1.06-1.28 (for WRN34-10). Surprisingly, while the results of WRN34-10 of GAIRAT were comparable or more excellent performance than MAIL, ours+MAIL outperformed GAIRAT.

Focused on the results on CIFAR-100 of Table 6, ours+MAIL was a comparable or degenerated performance of MAIL against some attacks. Meanwhile, ours+MAIL improved Clean by 0.94 (for ResNet-18) and 0.64 (for WRN34-10). Ours+WMMR improved the performance against many attacks and outperform Clean accuracy. The results of WRN34-10 improved robustness against all attacks, especially, ours+WMMR succeed to improve robustness by 0.82 at best for CW-PGD and AA.

Table 7 shows performances of the best and last models with and without the similarity on CIFAR-10. As described in Section 4.3, using the similarity not only made robust against CW-PGD and AA but also improved Clean accuracy. The same results were obtained in the best model.

Table 8 shows the results of ResNet-18 trained with different τ, i.e., $\tau = \{2, 3, 4\}$. According to increasing τ, while robustness against CW-PGD and AA significantly degenerated, PGD or FGSM dramatically improved robustness on CIFAR-10. On CIFAR-100, large τ improved Clean accuracy but PGD or FGSM were comparable. Meanwhile, robustness against CW-PGD or AA degenerated on both datasets. From there results, we determined $\tau = 2$ (for CIFAR-10) and $\tau = 3$ (for CIFAR-100).