AUTE: Peer-Alignment and Self-Unlearning Boost Adversarial Robustness for Training Ensemble Models

Lifeng Huang[1], Tian Su[2], Chengying Gao[2], Ning Liu[2], Qiong Huang[1] *

[1]College of Mathematics and Informatics, South China Agricultural University
[2]School of Computer Science and Engineering, Sun Yat-sen University
{huanglf6, qhuang}@scau.edu.cn, {mcsgcy, liuning2}@mail.sysu.edu.cn

Abstract

Adversarial attacks poses a significant threat to the security of AI-based systems. To counteract these attacks, adversarial training (AT) and ensemble learning (EL) have emerged as widely adopted methods for enhancing model robustness. However, a counter-intuitive phenomenon arises where the simple combination of these approaches may potentially compromising adversarial robustness of ensemble models. In this paper, we propose a novel method called ***A**lignment and **U**nlearning for **T**raining **E**nsembles* (AUTE), aiming to effectively integrate AT and EL to maximize their benefits. Specifically, AUTE incorporates two key components. Firstly, AUTE divides the ensemble into a big peer model and a single member in a loop manner, aligning their outputs for boosting robustness of each member. Secondly, AUTE introduces the concept of unlearning, actively forgetting specific data with over-confident properties to preserve model capacity to learn more robust features. Extensive experiments across various datasets and networks illustrate that AUTE achieves superior performance compared to baselines. For instance, a 5-member AUTE with ResNet-20 networks outperforms state-of-the-art method by 2.1% and 3.2% in classifying clean and adversarial data. Additionally, AUTE can easily extend to non-adversarial training paradigm, surpassing current standard ensemble learning methods by a large margin.

Code — https://github.com/mesunhlf/AUTE

Introduction

Deep neural networks (DNNs) have been widely employed in essential systems, including classification (He et al. 2016), recognition (Qiao et al. 2021) and translation (Ouyang et al. 2022). Despite their excellent performance, DNNs are not robust to adversarial examples: adding human-imperceptible perturbations to the clean data can deceive DNNs into outputting unexpected predictions (Wu et al. 2020; Huang et al. 2020; Doan et al. 2022).

There has emerged a number of research on defenses. For example, adversarial training (AT) and ensemble learning (EL) are two promising methods to enhance adversarial robustness. Specially, AT trains DNNs on generated adversarial examples, while most of AT methods merely focus on the defensive capabilities of individual models (Zhang et al. 2019; Wang et al. 2019; Rade and Moosavi-Dezfooli 2021b; Qizhang Li 2023). In contrast, EL methods optimize multiple models on the clean data and then combine them together for joint predictions (Pang et al. 2019; Yang et al. 2020; Deng and Mu 2024; Zhuang et al. 2024; Huang et al. 2023b). In general, EL approaches exhibit desirable robustness against black-box attacks. Given the observations that EL is benefit to constitute stronger standard trained models, a natural question arises: *Can AT and EL be integrated compatibly to further boost accuracy and robustness?*

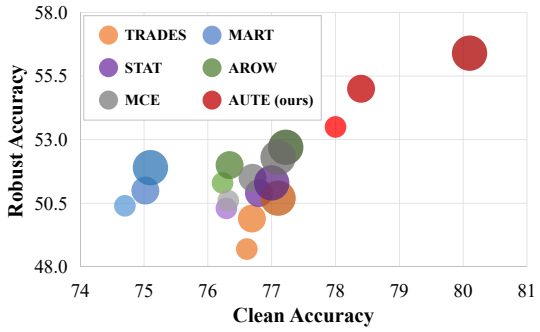

Figure 1: The comparison results between different ensemble models. The primary axes represent clean accuracy and adversarial robustness against the AA attack. The size of each circle (ensemble method) indicates the robustness level of ensemble models with 3, 5, and 8 members, ranging from small to large. Our proposed AUTE ensembles exhibit significantly better performance compared to existing methods.

Empirically, we observe a counter-intuitive phenomenon where the simple combination of AT and EL approaches sometimes offers marginal robustness improvements, while at other times, it inadvertently introduces side effects that can compromise the adversarial robustness. This phenomenon suggests that this simple combination may struggle to learn balanced features from both clean and adversarial data, which aligns the findings in related works (Xie and Yuille 2019). In light of these results, our forthcoming objective is to answer the question: *How can we maximize the benefits from EL for further boosting AT ensembles?*

*Corresponding author.

In this paper, we present a novel method aimed at establishing robust ensemble models, referred to as **A**lignment and **U**nlearning for **T**raining **E**nsembles (AUTE). AUTE consists of two crucial components: the Peer-Alignment scheme and the Self-Unlearning optimization, respectively.

- **Peer-Alignment**: Given that an ensemble of multiple AT models often demonstrates an improved defensive capacity than a single model, we partition the ensemble into two parts—a larger peer model characterized by higher robustness and a single member with weaker resistance. By aligning the robust features of the member with those of the peer model, its robustness is enhanced, and it is then integrated into the peer model to create a more robust ensemble. Since the alignment process between single member and its peer is carried out in a loop manner, the robustness of each ensemble member is gradually enhanced, leading to improved overall defensive capability from an ensemble learning perspective.
- **Self-Unlearning**: Most AT methods usually require network to capture robust features from all adversarial examples. However, it is challenging for models to entirely learn all adversarial data, which arises from the need for the consumption of a significant portion of the model capacity (Zhang et al. 2020). Particularly, we identify that certain adversarial examples are predicted with high confidence, resulting in over-confident predictions. This behavior can inadvertently have a negative impact on model performance (Müller, Kornblith, and Hinton 2019). Therefore, this category of data is designated for *forgetting* to free up model capacity and subsequently relearned by the AT model. As training progresses, the AT model can correctly classify more adversarial data, while simultaneously mitigating the over-confidence dilemma.

We conduct extensive experiments on various datasets and networks under a range of attack scenarios. The empirical results suggest that the proposed AUTE method perform signicantly higher clean accuracy as well as adversarial robustness compared to SOTA methods (see Fig. 1). Furthermore, AUTE exhibits well scalability, allowing it to further boost the performance as the ensemble size increases. For instance, training ResNet-20 networks on CIFAR-10 dataset, when AUTE trains an ensemble with 3 members, it improves average robustness from the SOTA 51.6% to 54.6%, and after increasing the group size to 8 members, AUTE boosts robustness from the SOTA 53.0% to 56.8%. In summary, the contributions of our work are three-fold:

- We introduce Peer-Alignment training scheme, a novel strategy designed to enhance the robustness of ensemble models. It guides each individual model within the ensemble to iteratively align with stronger peers, ultimately strengthening the overall robustness of the ensemble.
- We propose the Self-Unlearning optimization, which force ensembles to learn more robust features. Unlike most methods that continuously perform the learning process, we forces ensemble members to intentionally forget certain adversarial examples with high confidence and then attempt to relearn this type of data.
- Extensive experiments across various datasets and networks demonstrate that AUTE not only achieves the highest performance, but also showcases strong scalability and generalizability to the standard training paradigm.

Related Work

Adversarial Attacks. Given the susceptibility of DNNs to adversarial examples, there has been a significant surge in interest surrounding the development of attack techniques. Specially, an adversarial example is created by adding an imperceptible perturbation to the clean data, which can mislead the model to flip its label to a wrong prediction. A variety of attack methods has been developed recently, including Momentum-based Iterative Method (MIM) (Dong et al. 2018), Projecting Gradient Descent (PGD) (Madry et al. 2017), Parameters-freed Auto-Attack (AA) (Croce and Hein 2020), *etc.* These attacks have demonstrated their capability to achieve a high success rate in misleading DNNs, even when subjected to defensive mechanisms (Prakash et al. 2018; Athalye, Carlini, and Wagner 2018). They also exhibit ability to generalize adversarial effect across different networks and datasets under black-box scenarios (Huang, Gao, and Liu 2023; Huang et al. 2022). This poses a significant threat to the security of AI-controlled systems.

Adversarial Defenses. Adversarial Training (AT) treats adversarial examples as a form of augmentation data to train models (Madry et al. 2017). For example, TRADES (Zhang et al. 2019) is engineered to achieve better balance between clean accuracy and adversarial robustness. MART (Wang et al. 2019) distinguishes between misclassified and correctly classified data during optimization, aiming to improve overall model performance. HAT (Rade and Moosavi-Dezfooli 2021b) introduces helper examples to confine the excessive margin of decision boundaries. STAT (Qizhang Li 2023) creates collaborative examples (instead of adversarial examples) during training, fortifying its defensive capacity. AROW (Yang, Kong, and Kim 2023) pays more attention to less robust data, going a step further in boosting resistance. Although these methods perform well in single-model scenarios, effectively combining multiple AT models to enhance robustness presents an ongoing challenge.

Ensemble Learning (EL) was initially conceived to enhance overall performance in the context of classifying out-of-distribution data (Schapire 2013) or uncertainty estimation (Lakshminarayanan, Pritzel, and Blundell 2017). Recent studies show that EL methods can improve adversarial robustness, particularly in the settings of defending against black-box attacks (Pang et al. 2019; Yang et al. 2020). Several methods focus on training ensembles from an optimization standpoint, such as ADP (Pang et al. 2019) and GAL (Kariyappa and Qureshi 2019). Another way to strengthen the ensemble is employing augmentation and smoothing techniques, including DVERGE (Yang et al. 2020) and TRS (Yang et al. 2021b). However, their robustness is significantly degenerated under white-box attacks. Beyond standard EL methods, some advanced approaches, such as MCE (Zhang et al. 2022) and DRT (Yang et al. 2021a), incorporate both AT and EL concepts to build robust ensemble models, though their improvements are limited.

Methodology

Simple Combinations of AT and EL

We aim to answer two questions in the field: **(1)** Can adversarial training (AT) and ensemble learning (EL) be combined to improve adversarial robustness; and **(2)** If such a harmonious integration is feasible, how can we maximize the benefits derived from both EL and AT to further enhance clean accuracy and robustness?

To answer the first question, we start by building two simple combinations of AT and EL approaches. The first one directly incorporates the principles of AT into EL methods (AT2EL), which utilize adversarial examples for training ensemble models. The second approach involves the implementation of EL concepts into AT methods (EL2AT) that multiple AT models are optimized and subsequently consolidated into a large ensemble group. We test the robustness of these two combinations by using six methods, and detailed experimental results are shown in the Appendix. A.

We draw two key insights from the empirical results: **(1)** combining multiple AT models together to form an ensemble indeed exhibit higher adversarial robustness compared to a single AT model; and **(2)** Diversification regularization (Pang et al. 2019; Yang et al. 2020) sometimes unintentionally degrade the adversarial robustness of ensemble models. Building on these findings, we introduce a novel method to further enhance ensemble performance, termed Alignment and Unlearning for Training Ensembles (AUTE), addressing the second question. This method notably improves performance from both ensemble learning and adversarial training perspectives: Peer-Alignment and Self-Unlearning (Fig. 2).

Peer-Alignment Training Scheme

We introduce the Peer-Alignment (PA) scheme from the perspective of ensemble learning. Two crucial discoveries serve as the basis for it: **(1)** The experimental results above substantiate the conclusion that simply adding AT models in the ensemble can strengthen them to become a stronger defender. This finding is rooted in the intuition that stacking multiple members together to create a larger network provides increased capacity for learning more robust features from adversarial examples; and **(2)** Recent works (Zhao et al. 2022; Zhou et al. 2021; Huang et al. 2023a) have demonstrated that distillation is a valuable technique for training smaller robust networks by transferring defensive knowledge from an existing adversarially trained model. Thus, we can instruct each ensemble member to align features from its partners with higher robustness. These two observations motivate us to include three key steps in Peer-Alignment: separation, alignment, and reallocation. The intuitive mechanism of PA is illustrated in Fig. 2.

Separation, which creates adversarial examples and divides the entire ensemble into two groups. Specifically, given an ensemble model F comprising a total of n members $F = \{f_1, f_2, ..., f_n\}$, it is partitioned into two non-overlapping parts: one is a single member, denoted as f_i, which is presently the subject of optimization; another part is a *freezed* subset ensemble, namely the peer model of P_i, which consists of the rest of $n-1$ members.

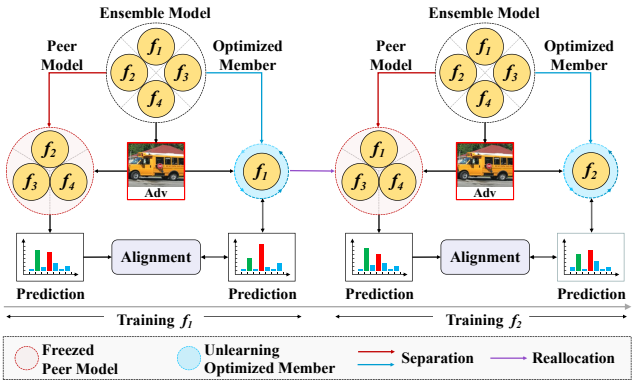

Figure 2: The pipeline of the proposed AUTE. At each iteration, the entire ensemble is divided into a large freezed peer model and an small optimized member. The member undergoes self-unlearning optimization (blue area) as well as aligns with the robust peer model. Afterward, the optimized member is reassigned within the peer model in preparation for alignment in the next iteration. These steps are executed in a round-robin manner.

Alignment, which instructs a optimized member to align with its peer model. Specifically, we use the Kullback-Leibler (KL) divergence to measure and minimize the differences between the two networks, guiding the alignment process, which is defined as:

$$\mathcal{L}_{i,\text{DIS}} = \text{KL}(f_i(x^{\text{adv}}), P_i(x^{\text{adv}})), \tag{1}$$

where x^{adv} is the adversarial example. It is evident that the capacity of the peer model become larger than this of the optimized member when the ensemble size $n \geq 3$.

Due to the increased model capacity, the peer model can effectively achieve higher robustness during training. However, we speculate that this does not necessarily make the peer model a suitable anchor for all adversarial examples. To test our hypothesis, we trained several ensembles using different AT methods and evaluated the robustness of individual members and their peer model (see Appendix. B). Surprisingly, we found that some clean data and adversarial examples were correctly classified by the smaller member but misclassified by the larger peer model, suggesting that the member may have acquired comprehensive knowledge than its peer on these specific data. This finding aligns with recent studies (Zhu et al. 2021). Thus, we employ a *selective* manner to instruct the member in alignment process. The optimization objective is reformulated Eq. (1) as:

$$\mathcal{L}_{i,\text{PA}} = (1 - f_i(x^{\text{adv}})) \cdot \text{KL}(f_i(x^{\text{adv}}), P_i(x^{\text{adv}})), \tag{2}$$

where the term $1 - f_i(x^{\text{adv}})$ is treated as an learning weight for encouraging the member to selectively align with the peer model based on the predicted confidence for the data.

Reallocation, which involves incorporating the optimized member into the peer model for making it stronger, while simultaneously excluding another member from the peer model. The excluded member then prepares for optimizing in the next iteration.

These steps are performed in a round-robin manner to progressively enhance the robustness of the ensemble. Intuitively, the peer model *dynamically* gains strength as each robustified member re-engages with the group, leading to improved accuracy on adversarial examples.

Self-Unlearning Optimization

While most methods typically treat all data equally during the optimization, recent studies emphasize the potential benefits of introducing instance-weighting to enhance robustness (Yang, Kong, and Kim 2023). Nevertheless, to the best of our knowledge, existing AT methods still force the model to *continuously* capture features from the dataset, including both adversarial and clean data. However, we discover this training paradigm may induce larger margin distance, which may harm the robustness of ensembles. To this end, we introduce the Self-Unlearning (SU) to optimize the model.

Rethink of Margin Distance. Most AT methods usually aims to achieve the maximum margin distance $\mathcal{D}$ between the groundtruth and others classes, which is defined as:

$$\mathcal{D}\big(f(x), y\big) = f(x)_y - \max_{k \neq y} f(x)_k. \tag{3}$$

Generally, $\mathcal{D}$ is employed to measure the gap between the data point and the nearest decision boundary in the latent space: the larger $\mathcal{D}\big(f(x), y\big)$, the farther away the data x is from the boundary, and vice versus. This concept is widely adopted in training robust models (Ding et al. 2019) or developing strong attacks (Carlini and Wagner 2017).

However, we observe a perplexing phenomenon where models like MART and AROW, despite having smaller margins on both clean and adversarial data, unexpectedly exhibit stronger defensive capacity compared to models with larger margins, such as SAT and TRADES (see Appendix. B). Two types of research shed light on explaining this behavior. Firstly, large margin distance is often linked to the overconfident property, potentially degrading the generalization of models (Müller, Kornblith, and Hinton 2019). Secondly, an excessive margin from the data to the decision boundary may impose a burden on the model capacity (Rade and Moosavi-Dezfooli 2021a). These observations inspire us to consider minimizing the margin distance, thereby conserving model capacity and improving adversarial robustness.

Forgetting Fewer for Learning More. Recent advanced studies have introduced regularization techniques in the training of AT models, which implicitly penalize overconfident data (Wang et al. 2019; Yang, Kong, and Kim 2023). To further reduce the margin between adversarial examples and the decision boundary, we incorporate the idea of machine unlearning (Sekhari et al. 2021) into the optimization of robust models, referred to as Self-Unlearning (SU).

In contrast to existing machine unlearning techniques that entirely eliminate the features of specific data, our approach emphasizes encouraging models to retain a subset of features from instances demonstrating overconfidence. Specifically, SU comprises two main components: firstly, the model continuously learns features from adversarial examples until they can be accurately classified, and secondly, the model progressively forgets data locates in low-loss regions with high confidence. Therefore, we reformulate standard Cross-Entropy by introducing an unlearning weight as $\mathcal{L}_{\text{UCE}}$:

$$\mathcal{L}_{\text{UCE}}(f(x), y) = -w_{\text{SU}} \cdot \sum_{k=1}^{C} y_k \cdot f(x)_k \tag{4}$$

$$w_{\text{SU}} = \begin{cases} 1 - \mathcal{D}\big(f(x), y\big) & \mathcal{D}\big(f(x), y\big) < \mathcal{M} \\ -\gamma \cdot \mathcal{D}\big(f(x), y\big) & \text{Otherwise} \end{cases} \tag{5}$$

where $\mathcal{M}$ is the threshold of margin distance, γ is a small constant. Intuitively, data with a greater margin has larger weight during the unlearning process, experiencing a faster displacement. As a consequence, both adversarial examples and clean data tend to move away from the low-loss regions.

AUTE Optimization

The proposed AUTE performs the Peer-Alignment (PA) and the Self-Unlearning (SU) for training the ensemble (Fig. 2). Specifically, we follow (Yang et al. 2020; Pang et al. 2019) to optimize members sequentially and then combing them to form a robust ensemble. In this process, each member simultaneously align with their peer model (PA) and captures (or forgets) the features from adversarial examples (SU). Therefore, the overall objective for a single member f_i is

$$\min_{\theta_i} \mathop{\mathbb{E}}_{(x,y)\sim D} \Big[\mathcal{L}_{i,\text{UCE}}(f_i(x^{\text{adv}}), y) + \beta \cdot \mathcal{L}_{i,\text{PA}}\Big], \tag{6}$$

where β is the balance weight, $\mathcal{L}_{i,\text{UCE}}$ and $\mathcal{L}_{i,\text{PA}}$ are objectives defined in SU and PA. Detailed pseudo-code for training an AUTE ensemble is shown in Appendix. C.

Experiments

Experimental Settings

Dataset. We mainly evaluate the ensemble models using the CIFAR-10 dataset. To illustrate the generalizability of the proposed method, we show that AUTE can consistently achieves superior performance across varying dataset scales—specifically, on smaller datasets like MNIST as well as complex datasets such as CIFAR-100 and Tiny-ImageNet.
Network. We align our ensemble setups with those of previous literature (Pang et al. 2019; Kariyappa and Qureshi 2019; Yang et al. 2020) during evaluations. Specifically, we utilize the light-weighting ResNet-20 architecture to develop robust ensemble models. Furthermore, we extend the experiments to include deeper and wider DNNs, such as VggNet-16, ResNet-18 and WideResNet-34. To demonstrate the scalability of AUTE, we build ensemble models with 3, 5, and 8 members together, respectively.
Baselines. We consider several ensemble versions of AT methods in comparisons: TRADES (Zhang et al. 2019), which aims to minimize the empirical risk and the robustness regularization. MART (Wang et al. 2019), which assigns larger weight to adversarial examples where the corresponding clean counterparts are wrongly predicted. HAT (Rade and Moosavi-Dezfooli 2021a) introduces a standard neural network as the helper to handle the overly perturbed adversarial images. STAT (Qizhang Li 2023) creates collaborative examples during the training of robust models. AROW (Yang, Kong, and Kim 2023) applies increased

Method	3 members						5 members						8 members					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	76.6	49.2	49.7	48.7	48.7	49.1	76.7	50.4	50.9	50.0	49.9	50.3	78.0	51.3	51.7	50.7	50.7	51.1
MART	74.7	51.1	51.4	50.6	50.4	50.9	75.0	51.6	52.0	51.2	51.0	51.5	75.0	52.6	52.8	52.1	51.9	52.4
HAT	77.2	50.0	50.6	49.3	49.2	49.8	77.9	50.5	51.3	49.9	49.8	50.4	79.2	51.2	51.7	51.0	50.7	51.2
STAT	76.3	50.6	50.9	50.3	50.3	50.5	76.8	51.2	51.5	51.0	50.9	51.2	77.0	51.7	51.9	51.4	51.3	51.6
MCE	76.3	51.1	51.5	51.0	50.6	51.1	76.7	51.9	52.2	51.8	51.5	51.9	77.1	52.5	52.9	52.5	52.3	52.6
AROW	76.2	51.7	52.0	51.3	51.3	51.6	76.3	52.4	52.7	52.1	52.0	52.3	77.4	53.1	53.5	52.8	52.7	53.0
AUTE	**78.0**	**54.8**	**56.1**	**54.1**	**53.5**	**54.6**	**78.4**	**55.7**	**57.0**	**54.4**	**55.0**	**55.5**	**80.1**	**57.1**	**58.1**	**55.6**	**56.4**	**56.8**

Table 1: The robust accuracy (%) of adversarially trained ensembles on CIFAR-10 dataset with ResNet-20 networks.

Method	VggNet-16						Resnet-18						WideResNet-34-10					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	82.9	52.9	54.0	52.3	52.1	52.8	81.6	55.7	56.4	55.5	55.2	55.7	84.4	56.7	56.3	56.2	56.0	56.3
MART	82.5	51.5	53.1	50.6	50.1	51.3	81.9	56.6	57.8	56.2	55.8	56.6	86.2	60.9	58.6	58.4	58.2	59.0
HAT	83.7	51.0	52.8	49.8	49.4	50.8	**84.8**	53.7	55.0	53.1	52.6	53.6	**86.3**	58.5	60.3	59.1	57.3	58.8
STAT	82.7	54.1	55.0	53.5	53.4	54.0	83.1	56.8	57.7	56.5	56.2	56.8	86.2	59.8	60.7	59.4	59.1	59.8
MCE	83.1	55.2	55.0	54.1	53.5	54.5	82.2	56.5	57.6	56.1	56.0	56.6	85.3	60.1	61.0	59.4	59.0	59.9
AROW	82.8	54.9	55.7	54.5	54.5	54.9	82.4	57.9	58.3	57.6	57.5	57.8	85.8	59.7	60.5	59.5	59.2	59.7
AUTE	**84.8**	**59.5**	**61.6**	**58.2**	**55.8**	**58.8**	82.1	**64.2**	**65.0**	**63.9**	**61.3**	**63.6**	85.1	**64.6**	**66.1**	**63.9**	**62.6**	**64.3**

Table 2: The robust accuracy (%) of adversarially trained ensembles on CIFAR-10 dataset with different structures.

regularization to data susceptible to adversarial attacks. MCE (Zhang et al. 2022), which is an advanced method for learning an ensemble with maximum margin.

Attack Models and Metrics. We test the robustness of ensembles under different attack scenarios. Particularly, we include four white-box attacks: **1)** 50-step Momentum-based Iterative attack Method (MIM) (Dong et al. 2018) with step size $\epsilon/5$; **2)** 10-step and 100-step PGD (Madry et al. 2017), denoted as PGD-10 and PGD-100, respectively; and **3)** Auto-Attack (AA) (Croce and Hein 2020), which is an ensemble of parameter-free attacks to fool classifiers. We evaluate the accuracy of ensembles on clean data (clean accuracy) and robustness against adversarial examples generated with a perturbation magnitude of $\epsilon = 8/255$. More experimental results of ensemble voting, smaller and larger perturbations are reported in Appendix.

Experimental Results of AUTE

We mainly assess the performance of ensemble models on the light-weighting ResNet-20 structure using CIFAR-10 dataset. We also extend our experiments to different datasets (*i.e.*, MNIST, CIFAR-100, and Tiny-ImageNet). We also explore the combination of various network architectures (*i.e.*, VggNet-16, ResNet-18, and WideResNet-34).

(1) Performance on CIFAR-10 Dataset. We demonstrate the clean accuracy and adversarial robustness under different white-box attacks in Tab. 1. In particular, we include the adversarially trained ensembles with various group size (*i.e.*, 3, 5, and 8 members) in the evaluations.

Referring to Table 1, it is consistent with our observations in Section that most ensembles exhibit similar trends. As the number of members in the ensemble models increases, their performance gradually improves. Among baselines, HAT consistently demonstrates superior performance in classifying clean data among the baselines. Conversely, MART emerges as a robust defender against adversarial examples, albeit at the cost of recording the lowest accuracy on clean data, thereby limiting its practical applicability. Similarly, methods like TRADES and STAT also exhibit varying degrees of bias towards either clean accuracy or robustness. The experimental results illustrate the challenge of balancing these two metrics simultaneously.

In comparison to baseline methods, AUTE demonstrates significantly enhanced performance in classifying both clean and adversarial examples. Notably, a 3-member AUTE accurately identifies 54.6% of adversarial examples on average, surpassesing the runner-up AROW ensemble by 3.5%. Moreover, the scalability of AUTE is remarkable, as evidenced by its favorable outcomes with larger ensemble groups. By training with more members, AUTE consistently enhances its performance. Particularly, an 8-member AUTE surpasses the nearest competitors AROW by a considerable margin of 3.9% in robustness against PGD-100 attacks. This improvement can be attributed to the fact that AUTE combines the alignment process with an unlearning paradigm to optimize ensemble members, thereby demonstrating better trade-off between robustness and accuracy.

(2) Performance on Complicated Structures. Instead of merely using the light-weighting ResNet-20 network to form ensemble models, we consider following settings: **1)** ResNet family but with different network depths and widths, *i.e.*, ResNet-18 and WideResNet-34-10; and **2)** the VggNet network family, *i.e.*, VggNet-16. The performance of these variants on CIFAR-10 dataset is reported in Tab. 2.

According to Tab. 2, there are three implications. Firstly, it is evident that incorporating deeper ResNets into ensemble models results in significant enhancements across both clean accuracy and adversarial robustness metrics. This observa-

Method	MNIST						CIFAR-100						Tiny-ImageNet					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	98.2	93.7	93.7	93.3	93.1	93.4	58.0	30.9	31.7	30.6	30.2	30.9	51.5	25.1	25.3	24.8	24.8	25.0
MART	99.1	93.3	93.6	92.3	89.7	92.2	56.1	31.0	32.1	30.3	30.0	30.9	50.1	24.2	24.6	23.9	23.7	24.1
HAT	**99.2**	93.6	93.9	93.2	93.4	93.5	58.2	31.4	31.3	30.8	30.3	31.0	50.3	24.4	25.0	24.1	24.0	24.4
STAT	98.7	93.6	94.3	93.5	90.3	92.9	57.3	31.8	32.0	31.6	31.3	31.7	50.2	24.8	25.0	24.8	24.8	24.9
MCE	98.5	93.1	94.0	93.0	92.9	93.3	58.1	31.1	32.3	32.0	31.6	31.8	50.0	25.0	25.6	25.0	24.8	25.1
AROW	97.8	93.3	93.3	93.1	92.8	93.1	58.4	32.1	32.4	31.9	31.0	31.9	50.6	25.2	25.6	25.1	25.0	25.2
AUTE	98.9	**94.9**	**94.9**	**94.3**	**94.2**	**94.6**	**59.3**	**32.6**	**33.2**	**32.3**	**32.0**	**32.5**	**52.5**	**25.9**	**26.3**	**25.4**	**25.3**	**25.7**

Table 3: The robust accuracy (%) of adversarially trained ensembles on different datasets.

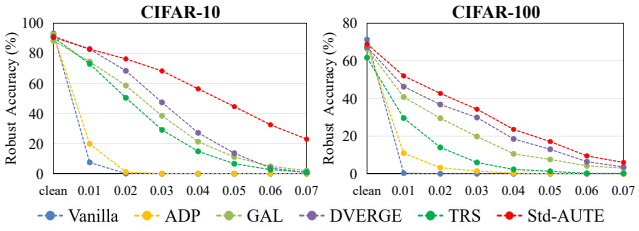

Figure 3: Robustness of standard (non-adversarial) trained ensembles on CIFAR-10 and CIFAR-100 datasets.

Settings	PA	SU	Clean	MIM	PGD-10	AA	Avg
AT	×	×	76.7 (-)	49.5	49.9	48.9	49.4 (-)
w/o PA	×	✓	78.4 (+1.7)	**55.2**	56.1	52.0	54.4 (+5.0)
w/o SU	✓	×	79.0 (+2.3)	50.4	51.2	49.4	50.3 (+0.9)
w/o Weight	⋆	✓	78.5 (+1.8)	54.9	55.9	52.5	54.4 (+5.0)
w/ ADP	⋆	✓	71.3 (-5.4)	45.0	45.4	44.6	45.0 (-4.4)
AUTE-LS (0.1)	✓	⋆	**79.2 (+2.5)**	48.6	49.3	47.6	48.5 (-0.9)
AUTE-LS (0.3)	✓	⋆	79.1 (+2.4)	49.0	49.8	48.3	49.0 (-0.4)
AUTE-LS (0.5)	✓	⋆	78.6 (+1.9)	49.1	50.0	48.4	49.2 (-0.2)
AUTE (**ours**)	✓	✓	78.0 (+1.3)	54.8	**56.1**	**53.5**	**54.8 (+5.4)**

Table 4: The robust accuracy (%) of adversarially trained ensembles trained by using different setups.

tion is consistent with prior studies that networks with more parameters usually brings larger model capacity, which supports models in achieving higher performance against adversarial attacks. Secondly, the choice of network family emerges as a crucial factor influencing learning preferences during model optimization. Specifically, VggNet ensembles significantly enhances clean data classification capabilities, *e.g.*, the improvements are 5.9% and 7.8% for TRADES and MART, respectively. However, the improvements in robustness are comparatively less pronounced. This fact reveals the accuracy and robustness bias related to the network structures. Thirdly, the proposed AUTE still demonstrate superior performance compared to current methods. Specially, leveraging both alignment and unlearning mechanisms, AUTE achieves significant advantages from larger model capacities. The average enhancements of these three complicated ensembles are 3.2%, 9.0% and 9.7% in adversarial robustness, surpassing those of baseline methods. This observation highlights the potential and effectiveness of AUTE.

(3) Performance on Different Datasets. We report the experimental results of various ensemble models with 3-member setup on MNIST, CIFAR-100 and Tiny-ImageNet datasets in Tab. 3. Specially, MNIST has resolutions of 28×28, which is smaller than the dimension of CIFAR-10 dataset. CIFAR-100 consists of 100 classes, and Tiny ImageNet comprises 200 classes with a resolution of 64×64. We train ResNet-20 ensembles on MNIST and ResNet-18 ensembles on CIFAR-100 and Tiny-ImageNet, respectively.

In Table 3, a noteworthy deviations are observed in the behaviors of MART, STAT, and HAT compared to their performances on the CIFAR-10 datasets. Specifically, the defensive capabilities of MART and STAT diminish significantly when confronted with different attacks. On the contrary, HAT exhibits an noticeable improvement in both clean accuracy and robustness on MNIST compared to its performance on CIFAR-10 (see Table. 1). We hypothesize that this success can largely be attributed to the fact that its helper model, trained using the standard paradigm, can achieve nearly 100% accuracy on these datasets. Moreover, it is clearly indicates that AUTE maintains state-of-the-art performance in defending against various attacks. Specifically, AUTE stands out as the frontrunner, outperforming the strongest competitor, AROW, by 0.9% in clean accuracy and 0.6% in average robustness on the CIFAR-100 dataset. Similarly, on the Tiny ImageNet dataset, AUTE achieves an average improvement of 1.9% in clean accuracy and 0.5% in robustness. This outcome provides compelling evidence to support the efficacy of employing AUTE for training more complex datasets.

(4) Performance on Standard Training. We explore the potential of AUTE by training ensembles exclusively on natural data, without incorporating any adversarial examples. We consider five standard ensemble methods in evaluations: Vanilla, ADP, GAL, DVERGE, and TRS. We note that these methods aim to achieve dual objectives: maintaining high clean accuracy while simultaneously enhancing adversarial robustness. More details are introduced in the Appendix. E.

We plot the black-box robustness of different methods in Fig. 3 and report detailed white-box robustness in Appendix. E. We can see that the proposed Std-AUTE also exhibits remarkable robustness against black-box attack with large perturbation compared to baselines. A similar tendency is observed under white-box attacks that AUTE surpasses the second-place baseline by a substantial margin.

The Effect of Alignment and Unlearning

We study the influence of Peer-Alignment (PA) and Self-Unlearning (SU) in AUTE. Concretely, we investigate each component in 3-member AUTE ensembles as following:

- Removing either PA or SU from the AUTE ensemble is denoted as *w/o PA* or *w/o SU*, respectively.
- Alignment process without a sample-selective manner in PA, where the weight in Eq. (2) is set equally for all adversarial data, denoted as *w/o Weight*.
- Replace the PA with a diversification ADP (Pang et al. 2019) to regularize the ensemble, denoted as *w/ ADP*.
- Replacing the SU with the Label Smoothing (LS) with a cofficient λ, denoted as *AUTE-LS (λ)*.
- The full version of the proposed method, *i.e.*, *AUTE*.

The comparison results for the aforementioned settings are presented in Table 4. The symbol $\star$ denotes the substitution of PA or SU with other mechanisms.

Observing the results in Table 4, both PA and SU contribute significantly to enhancing the performance of ensemble models (rows 2 and 3). Specifically, PA notably improves clean accuracy by 2.3%, while AU demonstrates a greater tendency to enhance adversarial robustness, with an average increase of 4.8%. Furthermore, removing the learning weights of PA results in a slight boost in clean accuracy, but at the expense of degraded robustness, particularly against attackers with large perturbations (row 4). Besides, the diversification regularizer may compromise the robustness, aligning with our discussions in Methodology (row 5). As for the label smoothing strategy, we observe that it indeed marginally improves clean accuracy. However, its defensive capacity diminishes significantly. An interesting observation is that as the label smoothing increases (rows 6-8), the model's robustness improves while its clean accuracy decreases. This aligns with findings from a related study (Yang et al. 2021b), suggesting that smoothing the model could be a viable defense against attacks. Consequently, the combination of PA and SU achieves desirable clean accuracy and the highest robustness compared to other setups (row 9).

Ablation Studies of AUTE

We train ensembles with 5-member on CIFAR-10 dataset, where quantitative results are presented in the Appendix.

Group Size of Peer Models. We consider to selectively combine fewer robust members to form the peer model. A common trend is observed: both clean accuracy and robustness gradually increase as the peer model becomes larger. This supports the conclusion that a larger capacity helps the model capture more robust features. Thus, we select all partner members within the peer model.

Threshold of Margin Distance. The threshold $\mathcal{M}$ determines unlearning behaviors of ensembles (Eq. (5)). We evaluate thresholds ranging from 0.01 to 1.0, where a smaller threshold indicates that adversarial data are positioned closer to the decision boundary. We find that variations in the threshold do not significantly affect clean accuracy, maintaining a stable performance of 78.0±0.5%. However, there is a decline in robustness with increasing thresholds.

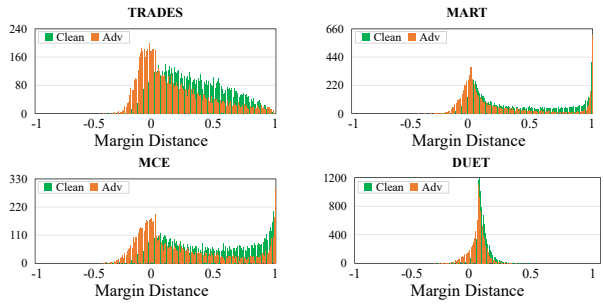

Figure 4: Statistic of margin distance on different ensembles.

Balance Weight. We train the AUTE with different weights β (Eq. (6)). We can see that clean accuracy and robustness initially improve and then decline after reaching the peaks. This suggests that emphasizing the learning from the peer model too much may lead to inverse consequences. Therefore, we select a medium weight to strike a better balance between these competing objectives.

Statistic of Margin Distance. We conducted a statistical analysis on the margin distances of 10000 clean data and adversarial examples (Fig. 4). We observe that MART and MCE ensembles tend to memorize all data, resulting in a phenomenon where a portion of data are classified with absolute confidence. Conversely, TRADES exhibited a relatively more uniform distribution trend. In comparison to baselines, AUTE showcased different statistical patterns that most of data points are concentrated within a small region, which is largely attributed to the unlearning process.

Conclusion

In this paper, we focused on enhancing the robustness of ensemble models. We introduced a novel learning method, termed AUTE, aimed at further improving ensemble robustness. AUTE comprises Peer-Alignment (PA) and Self-Unlearning (SU), which enhance performance from the perspectives of ensemble learning and adversarial training, respectively. Specifically, PA employing a selective alignment process to fortify the ensemble member in a iterative manner, and SU facilitates the ensemble in forgetting adversarial examples with overconfidence property. Extensive experiments show that AUTE not only achieves higher accuracy and robustness across different scenarios, including large datasets, complicated structures, and challenging attacks, but also showcases scalability, enabling extension to larger group and compatibility with standard training paradigms.

Acknowledgments

This work was supported by the National Key Research and Development Plan in China (2023YFC3306100), the National Natural Science Foundation of China (62472182), the Guangdong Basic and Applied Basic Research Foundation (2023A1515110075, 2024A1515010950), the Science and Technology Program of Guangzhou (2024A04J6542), and Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang) (ZJW-2023-04).

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Deng, Y.; and Mu, T. 2024. Understanding and improving ensemble adversarial defense. *Advances in Neural Information Processing Systems*, 36.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2019. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. In *International Conference on Learning Representations*.

Doan, B. G.; Xue, M.; Ma, S.; Abbasnejad, E.; and Ranasinghe, D. C. 2022. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems. *IEEE Transactions on Information Forensics and Security*, 17: 3816–3830.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, B.; Chen, M.; Wang, Y.; Lu, J.; Cheng, M.; and Wang, W. 2023a. Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24668–24677.

Huang, L.; Gao, C.; and Liu, N. 2023. Erosion Attack: Harnessing Corruption To Improve Adversarial Examples. *IEEE Transactions on Image Processing*.

Huang, L.; Gao, C.; Zhou, Y.; Xie, C.; Yuille, A. L.; Zou, C.; and Liu, N. 2020. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 720–729.

Huang, L.; Huang, Q.; Qiu, P.; Wei, S.; and Gao, C. 2023b. FASTEN: Fast Ensemble Learning For Improved Adversarial Robustness. *IEEE Transactions on Information Forensics and Security*.

Huang, L.; Wei, S.; Gao, C.; and Liu, N. 2022. Cyclical Adversarial Attack Pierces Black-box Deep Neural Networks. *Pattern Recognition*, 108831.

Kariyappa, S.; and Qureshi, M. K. 2019. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, 4970–4979. PMLR.

Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; and Storer, J. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8571–8580.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8681–8690.

Qizhang Li, W. Z. H. C., Yiwen Guo. 2023. Squeeze Training for Adversarial Robustness. In *International Conference on Learning Representations*.

Rade, R.; and Moosavi-Dezfooli, S.-M. 2021a. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Rade, R.; and Moosavi-Dezfooli, S.-M. 2021b. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*.

Schapire, R. E. 2013. Explaining adaboost. In *Empirical inference*, 37–52. Springer.

Sekhari, A.; Acharya, J.; Kamath, G.; and Suresh, A. T. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34: 18075–18086.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

Wu, J.; Chen, B.; Luo, W.; and Fang, Y. 2020. Audio steganography based on iterative adversarial attacks against convolutional neural networks. *IEEE transactions on information forensics and security*, 15: 2282–2294.

Xie, C.; and Yuille, A. 2019. Intriguing Properties of Adversarial Training at Scale. In *International Conference on Learning Representations*.

Yang, D.; Kong, I.; and Kim, Y. 2023. Improving Adversarial Robustness by Putting More Regularizations on Less Robust Samples. In *International conference on machine learning*. PMLR.

Yang, H.; Zhang, J.; Dong, H.; Inkawhich, N.; Gardner, A.; Touchet, A.; Wilkes, W.; Berry, H.; and Li, H. 2020. DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles. *arXiv preprint arXiv:2009.14720*.

Yang, Z.; Li, L.; Xu, X.; Kailkhura, B.; Xie, T.; and Li, B. 2021a. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*.

Yang, Z.; Li, L.; Xu, X.; Zuo, S.; Chen, Q.; Rubinstein, B.; Zhang, C.; and Li, B. 2021b. Trs: Transferability reduced ensemble via encouraging gradient diversity and model smoothness. *arXiv preprint arXiv:2104.00671*.

Zhang, D.; Zhang, H.; Courville, A.; Bengio, Y.; Ravikumar, P.; and Suggala, A. S. 2022. Building robust ensembles via margin boosting. In *International Conference on Machine Learning*, 26669–26692. PMLR.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2020. Geometry-aware Instance-reweighted Adversarial Training. In *International Conference on Learning Representations*.

Zhao, S.; Yu, J.; Sun, Z.; Zhang, B.; and Wei, X. 2022. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, 585–602. Springer.

Zhou, S.; Wang, Y.; Chen, D.; Chen, J.; Wang, X.; Wang, C.; and Bu, J. 2021. Distilling holistic knowledge with graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10387–10396.

Zhu, J.; Yao, J.; Han, B.; Zhang, J.; Liu, T.; Niu, G.; Zhou, J.; Xu, J.; and Yang, H. 2021. Reliable Adversarial Distillation with Unreliable Teachers. In *International Conference on Learning Representations*.

Zhuang, W.; Huang, L.; Gao, C.; and Liu, N. 2024. LAFED: Towards robust ensemble models via Latent Feature Diversification. *Pattern Recognition*, 150: 110225.