SELF-ENSEMBLE ADVERSARIAL TRAINING FOR IMPROVED ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to numerous breakthroughs in real-world applications brought by machine intelligence, deep neural networks (DNNs) are widely employed in critical applications. However, predictions of DNNs are easily manipulated with imperceptible adversarial perturbations, which impedes the further deployment of DNNs and may result in profound security and privacy implications. By incorporating adversarial samples into the training data pool, adversarial training is the strongest principled strategy against various adversarial attacks among all sorts of defense methods. Recent works mainly focus on developing a new loss function or regularizers, attempting to find the unique optimal point in the weight space. But none of them taps the potentials of classifiers obtained from standard adversarial training, especially states on the searching trajectory of training. In this work, we are dedicated to the weight states of models through the training process and devise a simple but powerful *Self-Ensemble Adversarial Training* (SEAT) method for yielding a robust classifier by averaging weights of history models. That considerably improves the robustness of the target model against several state-of-the-art adversarial attacks, even merely utilizing the naive cross-entropy loss to supervise. We also discuss the relationship between an ensemble of predictions of different adversarially trained models and the prediction of weight-ensembled models, as well as provide theoretical and empirical evidence that the proposed self-ensemble method provides a smoother loss landscape and better robustness than both individual models and an ensemble of predictions from different classifiers. We further analyze a subtle but fatal issue in the general settings for the self-ensemble model, which causes the deterioration of the weight-ensembled method in the late phases.

1 INTRODUCTION

Deep learning techniques have showed promise in disciplines such as computer vision (Krizhevsky et al., 2012; He et al., 2016), natural language processing (Vaswani et al., 2017; Devlin et al., 2019), speech recognition (DBL, 2012; Sak et al., 2015) and even in protein structural modeling (Si & Yan, 2021; Jumper et al., 2021). However, even for those efforts surpassing human-level performance in several fields (Jia et al., 2021; Gleave et al., 2020; Xu et al., 2019; 2018b; Wang et al., 2020a), deep learning based methods are vulnerable to adversarial examples generated by adding small adversarial perturbations to natural examples (Szegedy et al., 2014; Goodfellow et al., 2015).

Following the discovery of this adversarial vulnerability, numerous defense approaches for protecting DNNs from adversarial attacks have been proposed (Xu et al., 2018a; Liao et al., 2018; Qin et al., 2020; Papernot et al., 2016; Ross & Doshi-Velez, 2018). Because it is a straightforward and effective data-driven strategy, among these methods, adversarial training has become the de-facto way for strengthening the robustness of neural networks. The core of adversarial training is proactively inducing a robust model by forcing DNNs to learn adversarial examples generated from the current model, which helps DNNs gradually make up for their deficiencies. Therefore, it is a practical learning framework to alleviate the impact of adversarial attacks.

As a saying goes, a person may go faster but a team can go farther. An ensemble of classifiers fused by output aggregates the collective wisdom of all participated classifiers, which outperform the decision made by a single classifier. Many ensemble methods (Tramèr et al., 2018; Pang et al., 2019; Kurakin et al., 2018) have shown their abilities to boost adversarial robustness. Despite averaged predictions of different models enables the aggregation against a large number of attacks, it is still a heavy computing burden to adversarially train all the member models. What is worse, the architectures and parameters of *every* member model, as well as their related contribution factors, should be safely

saved and need performing forward propagation for every member repeatedly, which is a nightmare when dealing with massive components. Despite the fact that the ensemble technique adopts a variety of architectures and parameters, these models may have similar decision-making boundaries (Tramèr et al., 2017). Therefore, it is natural to wonder: *Could we develop the potential of the classifier itself to build a robust model rather than rely on multiple seemingly disparate classifiers? Could we make use of states on the searching trajectory of optimization when performing adversarial training?* Motivated by these questions, we discuss the relationship between an ensemble of different classifiers fused by predictions and by parameters, and present a computation-friendly self-ensemble method named SEAT. In summary, this paper has the following contributions:

- By rethinking the standard adversarial training and its variants, instead of developing a well-designed loss to find a single optimal in the weight space, we design an efficient algorithm called *Self-Ensemble Adversarial Training* (SEAT) by uniting states of every history model on the optimization trajectory through the process of adversarial training. Unlike the traditional ensemble technique, which is extremely time-consuming and repeating adversarial training procedures for individuals, SEAT simply requires training once.
- We present a theoretical explanation of the differences between the two ensemble approaches discussed above and visualize the change of loss landscape caused by SEAT. Furthermore, while using data augmentation of external data can experimentally fix the deterioration during the late stages (Gowal et al., 2020; Rebuffi et al., 2021), we further investigate a subtle but fatal intrinsic issue in the general settings for the self-ensemble model, which causes the deterioration of the weight-ensembled method.
- We thoroughly compare adversarial robustness of SEAT with other models trained by the state-of-the-art techniques against several attack methods on CIFAR-10 and CIFAR-100 datasets. Results have shown that the SEAT method itself can efficiently search in the weight space and significantly improve adversarial robustness.

2 Review: Standard Adversarial Training and its Successors

In this section, we first introduce basic concepts of adversarial training. Then we briefly review some variants and enhancements of standard adversarial training.

2.1 Preliminaries

Consider a standard image task which involves classifying an input image into one of K classes, let $x \in [0,1]^n$ denote the input natural image and $y \in \{1, 2, \ldots, K\}$ be the ground truth label, sampled from a distribution $\mathbb{D}$. For adversarial attack, the goal of the adversary is to seek a malignant example in the vicinity of x, by adding a human imperceptible perturbation $\varepsilon \in \mathbb{R}^n$. To exactly describe the meaning of "imperceptible", this neighbor region $\mathbb{B}_\varepsilon(x)$ anchored at x with apothem ε can be defined as $\mathbb{B}_\varepsilon(x) = \{(x', y) \in \mathcal{D} \mid \|x - x'\| \leq \varepsilon\}$. We also define a loss function: $\ell \stackrel{def}{=} \Theta \times \mathbb{D} \to [0, \infty)$, which is usually positive, bounded, and upper-semi continuous (Blanchet & Murthy, 2019; Villani, 2003; Bartlett & Mendelson, 2001) for all $\boldsymbol{\theta} \in \Theta$. Instead of the original formulation of adversarial training (Szegedy et al., 2014; Goodfellow et al., 2015), which mixes natural examples x and adversarial examples x' together to train models, the adversarial risk minimization problem can be formulated as a two-player game that writes as below:

$$\underbrace{\inf_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{(x,y)\sim\mathbb{D}}}_{\text{Outer minimization}} \underbrace{\sup_{x' \in \mathbb{B}_\varepsilon(x)} \ell\left(\boldsymbol{\theta}; x', y\right)}_{\text{Inner maximization}}. \tag{1}$$

In general, the adversary player iteratively computes more and more harmful perturbations step by step, where the PGD method (Madry et al., 2018) is commonly applied:

$$x^{t+1} = \Pi_{\mathbb{B}(x,\epsilon)}\left(x^t + \kappa \operatorname{sign}\left(\nabla_x \ell(x, y)\right)\right), \tag{2}$$

where κ is the step size and the generated samples at each iteration will be projected back to the $\mathbb{B}_\varepsilon(x)$. When finding x' at their current optimal in the inner maximization, the model parameters will be updated according to these fixed samples only in the outer minimization.

2.2 RELATED WORK

Emerging adversarial training methods are primarily concerned with strengthening their robustness through empirical defense. Here we roughly separate these improvements into two categories: (1) *sample-oriented* (i.e. for x' in the inner maximization) and (2) *loss-oriented* (i.e. for ℓ in the outer minimization) improvement.

From the side of the former, Baluja & Fischer (2017) utilize a GAN framework to create adversarial examples via the generative model while FAT (Zhang et al., 2020) proposes a PGD-based method which control the length of searching trajectories of PGD through the lens of geometry.

For the latter, a stream of works have been carried out with the goal of establishing a new supervised loss for better optimization. Usually, the loss ℓ in the both inner maximization and outer minimization[1] are cross-entropy (Good, 1956) defined as: $\ell_{ce} = -\boldsymbol{y}^T \log \boldsymbol{q}$ when given the probability vector $\boldsymbol{q} = [\mathrm{e}^{f_{\boldsymbol{\theta}}(x')_1}, \cdots, \mathrm{e}^{f_{\boldsymbol{\theta}}(x')_c}] / \sum_{k=1}^{c} \mathrm{e}^{f_{\boldsymbol{\theta}}(x')_k}$. The addition of a regularizer to the vanilla loss is one branch of the modification. Representative works like ALP (Kannan et al., 2018), TRADES (Zhang et al., 2019), and VAT (Miyato et al., 2019) introduce a regularization term to smooth the gap between the probability output of natural examples and adversarial ones, which can be denoted as:

$$\ell_{reg} = \ell_{ce} + \eta \sum_{(x_i, y_i) \sim \mathcal{B}} \mathcal{R}(p(\boldsymbol{\theta}; x_i', y_i), p(\boldsymbol{\theta}; x_i, y_i)), \tag{3}$$

where η is the regularization hyper-parameter and $\mathcal{R}(\cdot, \cdot)$ is l_2 distance for ALP and VAT[2], and Kullback-Leibler divergence for TRADES. $\mathcal{B}$ stands for the sampled minibatch. Another branch of the loss modification is to reweight the contributions of each instance in the minibatch based on their intrinsic characteristics. This can be formulated by:

$$\ell_{rew} = \sum_{(x_i, y_i) \sim \mathcal{B}} w(\boldsymbol{\theta}; x_i, x_i', y_i) \ell_{ce}\left(\boldsymbol{\theta}; x_i, x_i', y_i\right), \tag{4}$$

where $w(\cdot)$ is a real-valued function mapping the input instances to an importance score. MMA (Ding et al., 2020) separately considers the correct and incorrect classification on natural cases and switches between them by learning a hard importance weight:

$$\ell_{\text{MMA}} = \sum_{(x_i, y_i) \sim \mathcal{B}} w_1(\boldsymbol{\theta}; x_i, y_i) \ell_{ce}\left(\boldsymbol{\theta}; x_i, y_i\right) + w_2(\boldsymbol{\theta}; x_i, y_i) \ell_{ce}\left(\boldsymbol{\theta}; x_i', y_i\right), \tag{5}$$

where w_1 and w_2 are the indicator function $\mathbb{1}\left(f_{\boldsymbol{\theta}}\left(x_i\right) \neq y_i\right)$ and $\mathbb{1}\left(f_{\boldsymbol{\theta}}\left(x_i\right) = y_i\right)$, respectively. Inspired by the geometry concern of FAT, GAIRAT (Zhang et al., 2021) evaluates $w(\boldsymbol{\theta}; x, y) = (1 + \tanh(c_1 - 10c_2 + 5))/2$ to adaptively control the contribution of different instances by their corresponding geometric distance of PGD, where c_1 is a hyper-parameter and c_2 is the ratio of the minimum successful iteration numbers of PGD to the maximum iteration step. Embracing both two schools of improvement, MART (Wang et al., 2020b) adds a new regularization term apart from KL divergence and explicitly assigns weights within regularizers:

$$\ell_{\text{MART}} = \ell_{ce} + \sum_{(x_i, y_i) \sim \mathcal{B}} w(\boldsymbol{\theta}; x_i, y_i) \mathcal{R}_{KL}(p(\boldsymbol{\theta}; x_i', y_i), p(\boldsymbol{\theta}; x_i, y_i)) + \mathcal{R}_{mag}(p(\boldsymbol{\theta}; x_i', y_i)), \tag{6}$$

where $\mathcal{R}_{KL}$ is Kullback-Leibler divergence same as TRADES and $w_i = (1 - p(\boldsymbol{\theta}; x_i, y_i))$ is a softer scheme when compared with MMA. The margin term $\mathcal{R}_{mag} = -\log\left(1 - \max_{k \neq y_i} p_k\left(\boldsymbol{\theta}; x_i', y_i\right)\right)$ aims at improving the decision boundary.

The above methods endeavor to solve the highly non-convex and non-concave optimization in Eqn 1. However, they assume that the optimal weights of a classifier appear in its prime and abandon all the states along the route of optimization, which are beneficial to approach the optimal.

3 METHODOLOGY

In this section, we first discuss the traditional ensemble methods based on predictions of several individual classifiers. Then, we propose a Self-ensemble Adversarial Training (SEAT) strategy fusing

[1]For simplification, here we ignore the implemented difference of loss in the inner maximization and the outer minimization (e.g. TRADES and MMA). The reference of ℓ can be contextually inferred easily.

[2]For VAT, θ in the first term and the second term are different.

Algorithm 1 Self-Ensemble Adversarial Training (SEAT)

Input: A DNN classifier $f_{\boldsymbol{\theta}}(\cdot)$ with initial learnable parameters $\boldsymbol{\theta}_0$ and loss function ℓ; data distribution $\mathbb{D}$; number of iterations N; number of adversarial attack steps K; magnitude of perturbation ε; step size κ; learning rate τ; exponential decay rates for ensembling α; constant factor c.
Initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$, $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$.
for t $\leftarrow 1, 2, \cdots, N$ **do**
  Sample a minibatch (x, y) from data distribution $\mathbb{D}$
  $x'_0 \leftarrow x + \varepsilon, \varepsilon \sim \text{Uniform}(-\varepsilon, \varepsilon)$.
  for k $\leftarrow 1, 2, \cdots, K$ **do**
    $x'_k \leftarrow \Pi_{x'_k \in \mathbb{B}_\varepsilon(x)} \left(\kappa \,\text{sign}\left(x'_{k-1} + \nabla_{x'_{k-1}} \ell(\boldsymbol{\theta}; (x'_k, y))\right)\right)$
  end for
  $\boldsymbol{g}_{\boldsymbol{\theta}_t} \leftarrow \mathbb{E}_{(x,y)}\left[\nabla_{\boldsymbol{\theta}_t} \ell(\boldsymbol{\theta}_t; (x'_k, y))\right]$ in every minibatch
  Calculate τ_t according to the current iterations
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t - \tau_t \boldsymbol{g}_{\boldsymbol{\theta}_t}$
  $\alpha' \leftarrow \min\left(\alpha, \frac{t}{t+c}\right)$
  $\tilde{\boldsymbol{\theta}} \leftarrow \alpha' \tilde{\boldsymbol{\theta}} + (1 - \alpha')\boldsymbol{\theta}_t$
end for
Return A self-ensemble adversarial training model $f_{\tilde{\boldsymbol{\theta}}}(\cdot)$

weights of individual models during different periods, which is both intrinsically and computationally convenient. Following that, we further provide theoretical and empirical analysis on why the prediction of such an algorithm can make a great difference from that of simple ensembling classifiers.

3.1 Prediction-oriented Ensemble

Ensemble methods are usually effective to enhance the performance (Hansen & Salamon, 1990; Caruana et al., 2004; Russakovsky et al., 2015; Garipov et al., 2018) and improve the robustness (Tramèr et al., 2018; Pang et al., 2019; Gowal et al., 2020; Rebuffi et al., 2021). For clarity, we denote $\mathcal{F}$ as a pool of candidate models where $\mathcal{F} = \{f_{\boldsymbol{\theta}_1}, \cdots, f_{\boldsymbol{\theta}_n}\}$. To represent the scalar output of the averaged prediction of candidate classifiers over K categories, we define the averaged prediction $\bar{f}$ involving all candidates in the pool $\mathcal{F}$ as:

$$\begin{aligned} \bar{f}_{\mathcal{F}}(x, y) &= \sum_{i=1}^{n} \beta_i f_{\boldsymbol{\theta}_i}(x, y) \\ s.t. \sum_{i=1}^{n} \beta_i &= 1, \end{aligned} \tag{7}$$

where β represents the normalized contribution score allotted to the candidates. It is logical to assume that $\forall \beta_i > 0$ since each candidate model must be well-trained, otherwise it will be expelled from the list. Note that here we only discuss whether the participated model performs well on its own, not whether it has a positive impact on the collective decision making, which is a belated action.

3.2 Self-ensemble Adversarial Training

In order to average weights of a model, it is natural to choose an analogous form as Eqn 7 to obtain the predicted weights: $\tilde{\boldsymbol{\theta}} = \sum_{t=1}^{T} \beta_t \boldsymbol{\theta}_t, s.t. \sum_{t=1}^{T} \beta_t = 1$. However, such a simple moving average cannot keenly capture the latest change, lagging behind the latest states by half the sample width. To address this problem, we calculate the optimal weights by characterizing the trajectory of the weights state as an exponential moving average (EMA) for measuring trend directions over a period of time. Intuitively, EMA not only utilizes recent proposals from the current SGD optimization but also maintains some influence from previous weight states:

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \alpha \tilde{\boldsymbol{\theta}_T} + (1 - \alpha)\boldsymbol{\theta}_T \\ &= (1 - \alpha)\left[\boldsymbol{\theta}_T + \alpha \boldsymbol{\theta}_{T-1} + \ldots + \alpha^{T-1}\boldsymbol{\theta}_0\right] \end{aligned} \tag{8}$$

Because α is usually assigned to a value approaching 1, it is clear that EMA places a greater weight and significance on the most recent weight states. Algorithm1 summarizes the full algorithm.

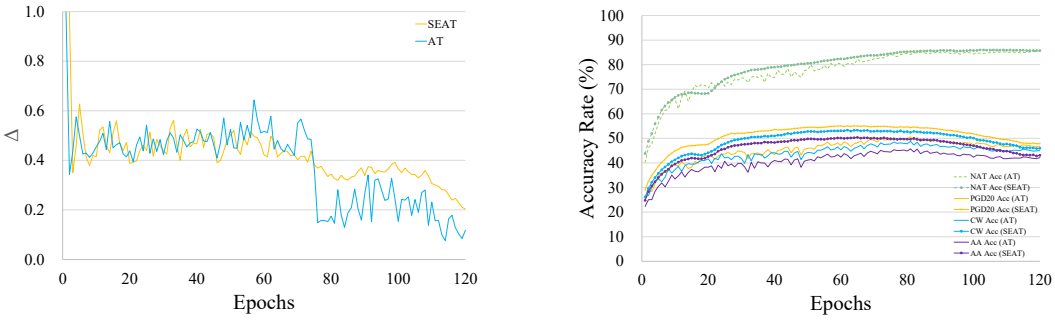
In consideration of both using the moving average technique, we wonder whether the averaged prediction has a link to the prediction of a weight-averaged model. In fact, they are indeed related to each other to some extent:

Proposition 1. *(Proof in Appendix B) Let $f_{\boldsymbol{\theta}}(\cdot)$ denote the predictions of a neural network parametrized by weights $\boldsymbol{\theta}$. Assuming that $\forall \boldsymbol{\theta} \in \Theta$, $f_{\boldsymbol{\theta}}(\cdot)$ is continuous and $\forall (x, y) \in \mathbb{D}$, $f_{\boldsymbol{\theta}}(x, y)$ is at least twice differentiable. Consider two points $\boldsymbol{\theta}_t, \tilde{\boldsymbol{\theta}} \in \Theta$ in the weight space and let $\boldsymbol{\xi} = \boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}$, for $t \in \{1, 2, \cdots, T\}$, the difference between $\bar{f}_{\mathcal{F}}(x, y)$ and $f_{\tilde{\boldsymbol{\theta}}}(x, y)$ is of the second order of smallness if and only if $\sum_{t=1}^{T}(\beta_t \boldsymbol{\xi}^{\mathrm{T}}) = \mathbf{0}$.*

Based on Proposition 1, we could immediately obtain the below conclusions:

Remark 1. *Note that it always constructs ensembles of the well-matched enhanced networks to obtain stronger defense, so it evenly assigns $\beta_1 = \beta_2 = \cdots = \beta_n = 1/n$. However, things change in the self-ensemble setting since models obtained at relatively late stages will be much robust than the beginning. Based on this assumption, it has a non-increasing sequence $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n, s.t. \sum_{i=1}^{n} \beta_i = 1$. The inequality is tight only when the initial weight reaches its fixed point.*

In this case, the averaged prediction will predispose to the models obtained at the end of optimization and such a predisposition loses the superiority of ensemble due to the phenomenon of homogenization in the late phases. To provide an empirical evidence of homogenization of models at latter periods, we visualize the effect of homogenization along the training process in Figure 1a. We define the homogenization of a model as: $\triangle_e = \frac{1}{||\mathcal{D}||} \min_{i \in [1,m]} |f_{\boldsymbol{\theta}_e}(x, y) - f_{\boldsymbol{\theta}_{e-i}}(x, y)|$, used for calculating the difference of the output of history models over a period of time m. Note that $\triangle_e$ becomes smaller and smaller along with epochs passing, which proves the existence of homogenization.



(a) The effect of homogenization of models.

(b) Robust accuracy of parameter-averaged ResNet18 against various attacks on CIFAR-10.

Figure 1: The overall analyse of AT and SEAT.

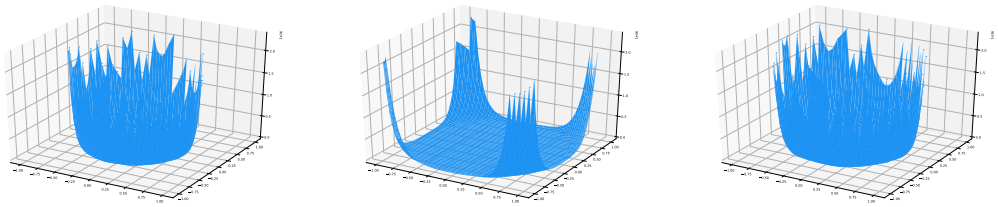
Here we formally state the gap between such an averaged prediction and the prediction given by SEAT in the following theorem:

Theorem 1. *(Proof in Appendix B) Assuming that for $i, j \in \{1, \cdots, T\}$, $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ if and only if $i = j$. The difference between the averaged prediction of multiple networks and the prediction of SEAT is of the second order of smallness if and only if $\beta_i = (1 - \alpha)\alpha^{i-1}$ for $i \in \{1, 2, \cdots, T\}$.*

Theorem 1 tells us that it is much harder for SEAT to approximate the averaged prediction of history networks than the vanilla EMA method. So the prediction of SEAT keeps away from the averaged prediction of models, which suffers from homogenization. To ensure whether the difference in Theorem 1 is benign, we provide empirical evidence of the improvement of the self-ensemble technique on the loss landscape.

It is difficult for the traditional 1D linear interpolation method to visualize non-convexities (Goodfellow & Vinyals, 2015) and biases caused by invariance symmetries (e.g. batch normalization) in the DNN. (Li et al., 2018) addresses these challenges, providing a scheme called filter normalization to explore the sharpness/flatness of DNNs on the loss landscape. Let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are two random direction vectors sampled from a random Gaussian distribution. So we plot the value of loss function around θ when inputting data samples:

$$L(\boldsymbol{\theta}; \boldsymbol{v}_1, \boldsymbol{v}_2) = \ell\left(\boldsymbol{\theta} + m\boldsymbol{v}_1 + n\boldsymbol{v}_2; x, y\right), \tag{9}$$



(a) The best individual (b) SEAT (c) Deteriorated SEAT

Figure 2: The loss surfaces of ResNet18 on CIFAR-10 test data. The loss surface of the standard adversarially trained models transitions up and down, which means the loss of such a model changes dramatically as moving along some directions. In contrast, SEAT has a fairly smooth landscape.

where $m = \frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{v}_1\|}$, $n = \frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{v}_2\|}$, and $\|\cdot\|$ denotes the Frobenius norm. Specifically, we apply the method to plot the surface of DNN within the 3D projected space. We visualize the loss values of ResNet18 on the testing dataset of CIFAR-10 in Figure 2. Each grid point in the x-axis represents a sampled gradient direction tuple $(\boldsymbol{v}_1, \boldsymbol{v}_2)$. It is clear that SEAT has a smooth and slowly varying loss function while the standard adversarially trained model visually has a sharper minimum with higher test error.

Safeguard for the Initial Stage: During the initial training period, the history models are less instrumental since they are not adequately learned. Therefore, it is not wise to directly make use of the ensembled weights, which may accumulate some naive weight states and do harm to approach the oracle. Therefore, we apply the protection mechanism for novice models at the beginning, denoted as $\alpha' = \min\left(\alpha, \frac{i}{i+c}\right)$, where c is a constant factor controlling the length of warm-up periods for history models.

3.3 Deterioration: The Devil in the Learning Rate

Having provided both theoretical and experimental support as well as a clear procedure of SEAT algorithm, we may take it for granted that this self-ensemble method could be integrated into the routine working flow as an off-shelf technique. But it is not the case.

As shown in Figure 2b and 2c, while the two self-ensemble models have *almost* nearly identical settings, their loss landscapes are completely different (quantitative results can be referred to Appendix A.4). Despite the fact that the landscape of deteriorated SEAT is a little mild than the standard adversarially trained model, it is still extremely steep when compared with SEAT. This phenomenon about deterioration is also discovered by (Rebuffi et al., 2021). They combine model weight averaging with data augmentation and external data to handle the deterioration. The intention behind such a solution is robust overfitting (Rice et al., 2020) of individuals. However, it is unclear why an ensemble model assembled from overfitting individuals performs poorly, while inferior models can still be aggregated to produce a superior one in the middle of training stages. In the rest of this section, we provide another solution from a quite new perspective: *the strategy of learning rate*.

Proposition 2. *(Proof in Appendix B) Assuming that every candidate classifier is updated by SGD-like strategy, meaning $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \tau_t \nabla_{\boldsymbol{\theta}_t} f_{\boldsymbol{\theta}_t}(x', y)$ with $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_T > 0$, the self-ensemble model will inevitably worsen and the degree depends on learning rate schedules.*

As illustrated in Figure 1a, we note that the phenomenon of homogenization exacerbates after each tipping point (e.g. Epoch 76, 91 and 111) when using the staircase strategy of learning rate. Combined with the trend of robust accuracy in Figure 1b, the result demonstrates that there exists a certain learning rate $\tau_{t'}$, the self-ensemble model cannot gain any benefits when the current learning rate drops below this bar.

4 Experimental Results

In this section, we first conduct a set of experiments to verify the advantages of the proposed method. Then, we investigate how each component of SEAT functions and how it affects robust accuracy.

4.1 EXPERIMENTAL SETUP

We mainly use ResNet18 and WRN-32-10 (Zagoruyko & Komodakis, 2016) for the experiments on CIFAR-10/CIFAR-100 and all images are normalized into $[0, 1]$. For simplicity, we only report the results based on L_∞ norm for the non-targeted attack. We train ResNet18 using SGD with 0.9 momentum for 120 epochs and the weight decay factor is set to $3.5e^{-3}$ for ResNet18 and $7e^{-4}$ for WRN-32-10. For SEAT, we use the piecewise linear learning rate schedule instead of the staircase one based on Proposition 2. The initial learning rate for ResNet18 is set to 0.01 and 0.1 for WRN-32-10 till Epoch 40 and then linearly reduced to 0.001, 0.0001 and 0.01, 0.001 at Epoch 60 and 120, respectively. The magnitude of maximum perturbation at each pixel is $\varepsilon = 8/255$ with step size $\kappa = 2/255$ and the PGD steps number in the inner maximization is 10. To evaluate adversarial defense methods, we apply several adversarial attacks including PGD (Madry et al., 2018), MIM (Dong et al., 2018), CW (Carlini & Wagner, 2017) and AutoAttack (AA) (Croce & Hein, 2020). We mainly compare the following defense methods in our experiments:

- **TRADES, MART, FAT and GAIRAT:** For TRADES and MART, we follow the official implementation of MART[3] to train both two models for better robustness. The hyper-parameter η of both TRADES and MART is set to 6.0. The learning rate is divided by 10 at Epoch 75, 90, 100 for ResNet18 and at 75, 90, 110 for WRN-32-10, respectively. For FAT and GAIRAT, we completely comply with the official implementation[4] to report results.
- **PoE:** We also perform the experiments on the Prediction-oriented Ensemble method (PoE). We select the above four advanced models as the candidate models. When averaging the output of these models, we assign different weights to individuals according to their robust accuracy. Specifically we respectively fix $\beta_{1,2,3,4} = 0.1, 0.2, 0.3, 0.4$ for FAT, GAIRAT, TRADES, and MART, named PoE. Considering the poor robust performance of FAT and GAIRAT, we also provide the PoE result by averaging the outputs of TRADES trained with $\lambda = 1, 2, 4, 6, 8$ and assign the same β to them, named PoE (TRADES).
- **CutMix (with WA):** Based on the experimental results of Rebuffi et al. (2021), Cutmix with a fixed window size achieves the best robust accuracy. We follow this setting and set the window size to 20. Since we do not have matched computational resources with Deepmind, we only train models for 120 epochs with a batch size of 128 rather than 400 epochs with a batch size of 512. Besides, we neither use Swish/SiLU activation functions (Hendrycks & Gimpel, 2016) nor introduce external data and samples crafted by generative models (Brock et al., 2019; Ho et al., 2020; Child, 2021) for fair comparison.

4.2 ROBUSTNESS EVALUATION OF SEAT

In this section, we fully verify the effectiveness of the SEAT method. We report both average accuracy rates and standard deviations. All results in Tables 1 and 2 are computed with 5 individual trials. Results on ResNet18 are summarized in Table 1. Here, we mainly report the results on the CIFAR-10 dataset due to the space limitation. For the results on CIFAR-100, please refer to Appendix A.2. From the table, we can see that the superiority of SEAT is apparent especially considering we only use candidate models supervised by the traditional cross-entropy loss without external data. Compared with two advanced loss-oriented methods (TRADES and MART), the results of SEAT are at least $\sim 2\%$ better than TRADES and MART against CW and AA. We further emphasize that although the boost of SEAT regarding PGD attacks is slight when compared with GAIRAT, SEAT achieves startling robust accuracy results when facing CW and AA attacks. Note that AA is an ensemble of various advanced attacks (the performance under each component of AA can be found in Appendix A.1) and thus the robust accuracy of AA reliably reflects the adversarial robustness, which indicates that the robustness improvement of SEAT is not biased. It's worth noting that PoE is unexpectedly much weaker in defending AA attack even than its members. We guess the poor performance of FAT and GAIRAT encumber the collective though we appropriately lower the contribution score of them. If we switch to an ensemble of the outputs of TRADES with different λs, the performance will be slightly improved. That demonstrates that the robustness of every candidate model will affect the performance of the ensemble. The negligible computational burden estimated in Appendix A.5 also demonstrates the superiority of our on-the-fly parameter-averaged method of SEAT.

We also evaluate the SEAT method on WRN-32-10. Results are shown in Table 2. We notice that the superiority of SEAT appears to be enhanced when using WRN-32-10. The gap between SEAT

[3]https://github.com/YisenWang/MART

[4]https://github.com/zjfheart/Geometry-aware-Instance-reweighted-Adversarial-Training

Table 1: Comparison of our algorithm with different defense methods using ResNet18 on CIFAR-10. The maximum perturbation is $\varepsilon = 8/255$. Average accuracy rates (in %) and standard deviations have shown that the proposed SEAT method greatly improves the robustness of the model.

Method	NAT	PGD^{20}	PGD^{100}	MIM	CW	AA
AT	84.32±0.23	48.29±0.11	48.12±0.13	47.95±0.04	49.57±0.15	44.37±0.37
TRADES	83.91±0.33	54.25±0.11	52.21±0.09	55.65±0.1	52.22±0.05	48.2±0.2
FAT	**87.72±0.14**	46.69±0.31	46.81±0.3	47.03±0.17	49.66±0.38	43.14±0.43
MART	83.12±0.23	55.43±0.16	53.46±0.24	**57.06±0.2**	51.45±0.29	48.13±0.31
GAIRAT	83.4±0.21	54.76±0.42	54.81±0.63	53.57±0.31	38.71±0.26	31.25±0.44
PoE	85.41±0.29	55.2±0.37	55.07±0.24	54.33±0.32	49.25±0.16	46.17±0.35
PoE (TRADES)	83.57±0.31	53.88±0.45	53.82±0.27	55.01±0.18	52.72±0.65	49.2±0.24
CutMix (with WA)	81.26±0.44	52.77±0.33	52.55±0.25	53.01±0.25	50.01±0.55	47.38±0.36
SEAT	83.7±0.13	**56.02±0.11**	**55.97±0.07**	**57.13±0.12**	**54.38±0.1**	**51.3±0.26**
SEAT+CutMix	81.53±0.31	55.3±0.27	54.82±0.18	56.41±0.17	53.83±0.31	49.1±0.44

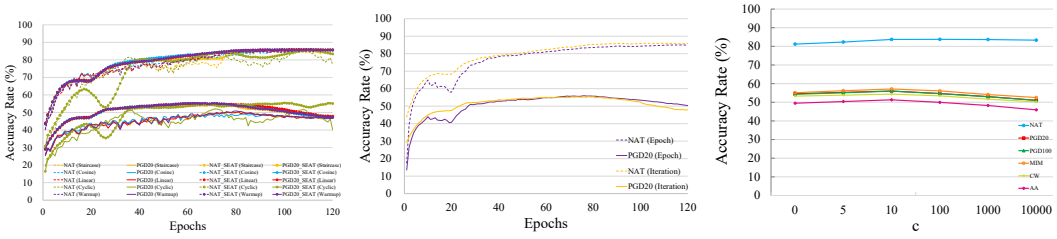
Table 2: Comparison of our algorithm with different defense methods using WRN-32-10 on CIFAR-10. The maximum perturbation is $\varepsilon = 8/255$. Average accuracy rates (in %) and standard deviations have shown that SEAT also shows great improvement on robustness.

Method	NAT	PGD^{20}	PGD^{100}	MIM	CW	AA
AT	87.32±0.21	49.01±0.33	48.83±0.27	48.25±0.17	52.8±0.25	48.17±0.48
TRADES	85.11±0.77	54.58±0.49	54.82±0.38	55.67±0.31	54.91±0.21	52.19±0.44
FAT	**89.65±0.04**	48.74±0.23	48.69±0.18	48.24±0.16	52.11±0.71	46.7±0.4
MART	84.26±0.28	54.11±0.58	54.13±0.3	55.2±0.22	53.41±0.17	50.2±0.36
GAIRAT	85.92±0.69	58.51±0.42	58.48±0.34	58.37±0.27	44.31±0.22	39.64±1.01
PoE	87.1±0.25	55.75±0.2	55.47±0.19	56.04±0.31	53.66±0.18	49.44±0.35
PoE (TRADES)	86.03±0.37	54.26±0.47	54.73±0.21	55.01±0.22	55.52±0.18	53.2±0.4
CutMix (with WA)	82.79±0.44	58.43±1.21	58.2±0.83	58.95±0.57	58.32±0.43	54.1±0.82
SEAT	86.44±0.12	59.84±0.2	59.8±0.16	**60.87±0.1**	58.95±0.34	55.67±0.22
SEAT+CutMix	84.81±0.18	**60.2±0.16**	**60.31±0.12**	60.53±0.21	**59.46±0.24**	**56.03±0.36**

and advanced defense methods enlarges to at least $\sim 4\%$. In the setting with data augmentation, when compared with the results of ResNet18, SEAT combined with CutMix gains much higher performance on robust accuracy against several attacks as the model becomes larger. The observation given by (Rebuffi et al., 2021) shows that additional generated data would overload the model with low capacity. When it comes to ours, we can further infer from a marginal and even a negative effect shown in Table 1 that the model with low capacity cannot benefit from both external and internal generated data. This phenomenon springs from the fact that CutMix techniques tend to produce augmented views that are far away from the original image they augment, which means that they are a little hard for small neural networks to learn.

4.3 Ablation Study

In this section, we perform several ablation experiments to investigate how different aspects of SEAT influence its effectiveness. If not specified otherwise, the experiments are conducted on CIFAR-10 using ResNet18.



(a) Different learning rate strategies (b) Epoch v.s. Iteration (c) Length of initial safeguard

Figure 3: The overall ablation study of SEAT. The dashed line represents the natural accuracy while the solid line represents the robust accuracy.

Different Learning Rate Strategies. The staircase learning rate technique is always applied to update weights in commonly used neural networks, which τ remains unchanged within the specific

epochs and suddenly decreases to a certain value at some points. We also have the cosine, cyclic, and warming up learning rate schedules to compare with, in addition to the staircase and piecewise linear one deployed to SEAT. Figure 3a depicts the results of this experiment using various learning rate methodologies. The natural accuracy under various learning rate schedules is reported by the top dashed lines, which all grow gradually over time. After reaching their maxima, the relative bottom solid lines fall more or less. Quantitative Results of above five strategies from best checkpoints are shown in Appendix A.3. However, it can be observed that the red line (piecewise linear) and the blue line (cosine) decrease more slowly than the orange line (staircase) and they achieve higher robust accuracy at much later epochs than the staircase. Apparently, the piecewise linear strategy makes τ approach the threshold $\tau_{t'}$ at a slow pace while the cosine one increases the valid candidate models that can make contributions. Likewise, the cyclic strategy gets better robust accuracy and further improves the performance in the late stage. But the effect of warming up learning rate is marginal since the homogenization of the candidate models we mentioned in Sec. 3.2 cannot be fixed by the warmup strategy. These facts bear out Proposition 2 from the side.

Epoch-based v.s. Iteration-based Ensembling. Next, we investigate whether the frequency of ensembling has a great effect on the robustness of the collective. Specifically, we select two scenarios to compare with. One is an epoch-based ensemble that only gathers the updated models after traversing the whole training dataset. The other is an iteration-based ensemble collecting models whenever SGD functions. Comparison between the epoch-based and the iteration-based ensembling is shown in Figure 3b. It can be seen that the curves of both two schemes are nearly the same in their prime except that the iteration-based method has a slightly greater natural accuracy and degrades more slowly.

Influence of the Initial Safeguard. We also study the effect of the length of the initial safeguard. We plot all results against different kinds of attacks with c ranging from 0 to 10000 (i.e. Epoch 32). The curves in Figure 3c show that the model ensembled with large c, which ignores more history models, results in weaker memorization and poor robustness. Actually, from the perspective of Proposition 2, t' is fixed if the scheme of learning rate updating and the total training epochs are unchanged. So using a large c is equivalent to squeeze the number of benign candidates before t'. Similarly, a small c can manually filter the early models without being well trained. However, such a drop is minor because EMA has a built-in mechanism for adjusting.

Mixed Strategies with SEAT for Improved Robustness. Finally, we report the results under different mixed strategies on CIFAR-10 dataset with ResNet18 in Table 3. Overall, SEAT+TRADES and SEAT+TRADES+CutMix do not perform well. We believe the reason is that TRADES cannot well adapt to learning strategies other than the staircase one for its history snapshots of TRADES trained under other strategies are much inferior than the ones trained under the staircase learning rate strategy. However, MART is in harmony with other strategies. The combination of SEAT+MART+CutMix achieves the best or the second-best robustness against almost all types of attacks.

Table 3: Average robust accuracy (%) and standard deviation under different mixed strategies on CIFAR-10 dataset with ResNet18.

Method	NAT	PGD20	PGD100	MIM	CW	APGDCE	APGDDLR	APGDT	FABT	Square	AA
SEAT	**83.7 ±0.13**	56.02 ±0.11	55.97 ±0.07	57.13 ±0.12	**54.38 ±0.1**	53.87 ±0.17	53.35 ±0.24	50.88 ±0.27	51.41 ±0.37	57.77 ±0.22	51.3 ±0.26
SEAT +TRADES	81.21 ±0.44	57.05 ±0.28	57.0 ±0.15	**57.92 ±0.12**	52.75 ±0.09	50.75 ±0.25	50.36 ±0.29	48.56 ±0.37	49.45 ±0.65	54.45 ±0.58	49.91 ±0.17
SEAT +MART	78.94 ±0.14	56.92 ±0.46	56.88 ±0.29	57.24 ±0.48	52.09 ±0.66	54.06 ±0.56	53.77 ±0.25	51.03 ±0.44	51.11 ±0.41	57.76 ±0.08	51.7 ±0.35
SEAT +TRADES+CutMix	78.22 ±0.33	57.14 ±0.21	57.09 ±0.45	57.11 ±0.39	53.17 ±0.41	52.67 ±0.82	51.86 ±0.5	50.91 ±0.17	50.23 ±0.29	55.12 ±0.46	50.01 ±0.44
SEAT +MART+CutMix	75.87 ±0.24	**57.3 ±0.16**	**57.29 ±0.43**	57.47 ±0.18	53.13 ±0.2	**54.33 ±0.33**	**53.98 ±0.61**	**51.2 ±0.73**	**51.42 ±0.17**	**58.01 ±0.08**	**52.1 ±0.22**

5 CONCLUSION

In this paper, we propose a simple but powerful method called Self-Ensemble Adversarial Training (SEAT), which unites states of every history model on the optimization trajectory through the process of adversarial training. Compared with the standard ensemble method, SEAT only needs training once and has a better reusability. Besides, we give a theoretical explanation of the difference between the above two ensemble methods and visualize the change of loss landscape caused by SEAT. Furthermore, we analyze a subtle but fatal intrinsic issue in the learning rate strategy for the self-ensemble model, which causes the deterioration of the weight-ensembled method. Extensive experiments validate the effectiveness of the proposed SEAT method.

REFERENCES

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012. doi: 10.1109/MSP.2012.2205597. URL https://doi.org/10.1109/MSP.2012.2205597.

Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *CoRR*, abs/1703.09387, 2017. URL http://arxiv.org/abs/1703.09387.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In David P. Helmbold and Robert C. Williamson (eds.), *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, volume 2111 of *Lecture Notes in Computer Science*, pp. 224–240. Springer, 2001. doi: 10.1007/3-540-44581-1_15. URL https://doi.org/10.1007/3-540-44581-1_15.

Jose H. Blanchet and Karthyek R. A. Murthy. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936. URL https://doi.org/10.1287/moor.2018.0936.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.49. URL https://doi.org/10.1109/SP.2017.49.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In Carla E. Brodley (ed.), *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/1015330.1015432. URL https://doi.org/10.1145/1015330.1015432.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. MMA training: Direct input space margin maximization through adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.

2018.00957. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8803–8812, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJgEMpVFwB.

I. J. Good. Some terminology and notation in information theory. 1956.

Ian J. Goodfellow and Oriol Vinyals. Qualitatively characterizing neural network optimization problems. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6544.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *CoRR*, abs/2010.03593, 2020. URL https://arxiv.org/abs/2010.03593.

Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, 1990. doi: 10.1109/34.58871. URL https://doi.org/10.1109/34.58871.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6709–6718. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Jia_IoU_Attack_Towards_Temporally_Coherent_Black-Box_Adversarial_Attack_for_Visual_CVPR_2021_paper.html.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli,

and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018. URL http://arxiv.org/abs/1803.06373.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Alexey Kurakin, Ian J. Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan L. Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. *CoRR*, abs/1804.00097, 2018. URL http://arxiv.org/abs/1804.00097.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1778–1787. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00191. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Liao_Defense_Against_Adversarial_CVPR_2018_paper.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821. URL https://doi.org/10.1109/TPAMI.2018.2858821.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4970–4979. PMLR, 2019. URL http://proceedings.mlr.press/v97/pang19a.html.

Nicolas Papernot, Patrick D. 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, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 582–597. IEEE Computer Society, 2016. doi: 10.1109/SP.2016.41. URL https://doi.org/10.1109/SP.2016.41.

Yao Qin, Nicholas Frosst, Sara Sabour, Colin Raffel, Garrison W. Cottrell, and Geoffrey E. Hinton. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Skgy464Kvr.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *CoRR*, abs/2103.01946, 2021. URL https://arxiv.org/abs/2103.01946.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8093–8104. PMLR, 2020. URL http://proceedings.mlr.press/v119/rice20a.html.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1660–1669. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17337.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Hasim Sak, Andrew W. Senior, Kanishka Rao, and Françoise Beaufays. Fast and accurate recurrent neural network acoustic models for speech recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 1468–1472. ISCA, 2015. URL http://www.isca-speech.org/archive/interspeech_2015/i15_1468.html.

Yunda Si and Chengfei Yan. Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. *bioRxiv*, 2021. doi: 10.1101/2021.05.10.443415. URL https://www.biorxiv.org/content/early/2021/05/11/2021.05.10.443415.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. *CoRR*, abs/1704.03453, 2017. URL http://arxiv.org/abs/1704.03453.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

C. Villani. Topics in optimal transportation. 2003.

Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 339–348. Computer Vision Foundation / IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00042. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Transferable_Controllable_and_Inconspicuous_Adversarial_Attacks_on_Person_Re-identification_With_CVPR_2020_paper.html.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=rklOg6EFwS.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018a. URL http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-4_Xu_paper.pdf.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4951–4961. Computer Vision Foundation / IEEE Computer Society, 2018b. doi: 10.1109/CVPR.2018.00520. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_Fooling_Vision_and_CVPR_2018_paper.html.

Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4135–4144. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00426. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Xu_Exact_Adversarial_Attack_to_Image_Captioning_via_Structured_Output_Learning_CVPR_2019_paper.html.

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

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11278–11287. PMLR, 2020. URL http://proceedings.mlr.press/v119/zhang20z.html.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=iAX0l6Cz8ub.