Chinese Journal of Electronics
vol. 33, no. 4, pp. 979–988, July 2024
https://doi.org/10.23919/cje.2022.00.449

Special Focus on Explainability, Robustness, and Security in AI Systems

RESEARCH ARTICLE

Efficient Untargeted White-Box Adversarial Attacks Based on Simple Initialization

Yunyi ZHOU, Haichang GAO, Jianping HE, Shudong ZHANG, and Zihui WU

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Corresponding author: Haichang GAO, Email: hchgao@xidian.edu.cn

Manuscript Received December 27, 2022; Accepted August 7, 2023

Copyright © 2024 Chinese Institute of Electronics

Abstract — Adversarial examples (AEs) are an additive amalgamation of clean examples and artificially malicious perturbations. Attackers often leverage random noise and multiple random restarts to initialize perturbation starting points, thereby increasing the diversity of AEs. Given the non-convex nature of the loss function, employing randomness to augment the attack's success rate may lead to considerable computational overhead. To overcome this challenge, we introduce the one-hot mean square error loss to guide the initialization. This loss is combined with the strongest first-order attack, the projected gradient descent, alongside a dynamic attack step size adjustment strategy to form a comprehensive attack process. Through experimental validation, we demonstrate that our method outperforms baseline attacks in constrained attack budget scenarios and regular experimental settings. This establishes it as a reliable measure for assessing the robustness of deep learning models. We explore the broader application of this initialization strategy in enhancing the defense impact of few-shot classification models. We aspire to provide valuable insights for the community in designing attack and defense mechanisms.

Keywords — Adversarial examples, White-box attacks, Image classification.

Citation — Yunyi ZHOU, Haichang GAO, Jianping HE, *et al.*, "Efficient Untargeted White-Box Adversarial Attacks Based on Simple Initialization," *Chinese Journal of Electronics*, vol. 33, no. 4, pp. 979–988, 2024. doi: 10.23919/cje.2022.00.449.

I. Introduction

Deep neural networks (DNNs) have caused paradigm shifts and revolutionized the way we approach computer tasks. However, due to the insufficient understanding of deep learning mechanisms, DNNs are subject to several hidden security risks in practice, the most well-known of which is the adversarial attack during the inference phase. Attackers can mislead a classifier by adding specific numerical vectors to the input. To enhance model robustness, several defense methods, including image preprocessing [1] and adversarial training (AT) [2], have been proposed. Adversarial attacks and defenses complement each other, and powerful attack methods can help identify shortcomings in existing model designs and defense algorithms. AT, which uses adversarial examples (AEs) for data augmentation to train models, remains the most effective defense method.

To break through increasingly diverse, sophisticated, and effective defense strategies, several white-box adversarial attacks strive to diversify perturbation directions. An essential design for improving attack performance is the use of multiple random restarts. By doing this, AEs are generated from different initial points each time, repeating the entire attack process several times. However, this exploration of the input space comes with a computational overhead. Moreover, Tashiro *et al.* [3] claim that random initial perturbations may not actively influence the output space after forward propagation through the neural network, thus not altering model predictions for attack purposes. In contrast, output diversified initialization (ODI) in [3] applies random restarts to the output space and uses the information in the output space to guide the generation of initial points at each restart for better attack results. However, according to subsequent comparison experiments, random noise causes relatively few changes in the predicted values. Therefore, we believe that random noise in the output space is also not

Guest Editor: Prof. Qi LI, Tsinghua University.

the best tactic as a starting point for exploration.

We believe there is still room for optimization in the design of the initialization process, given that random sampling in the initialization stage is sub-optimal. On the one hand, there is a lack of loss functions that directly guide the initialization optimization process, and standard loss functions (e.g., cross-entropy (CE) loss) perform poorly in the initialization process. Sriramanan *et al.* [4] claim that the initial gradient direction that maximizes CE loss may not provide a valid reference for the optimal solution. Ma *et al.* [5] assert that imbalanced gradients accompany the use of the margin loss early in the attack and push the attack towards a suboptimal direction. Moreover, reference [6] states poor surrogate selection leads to potentially weak projected gradient descent (PGD) performance, and switching between different objectives helps the algorithm expand its search space. On the other hand, there is still a research gap in finding suitable attack classes in the initialization phase, making it challenging to avoid attack failures and consuming a large number of computational resources.

Based on these two points, this paper proposes a one-hot mean square error (MSE) loss for the initialization phase, allowing for better consideration of each class's attack potential. To validate the improvement of the novel objective on the attack effectiveness, we combine it with PGD sequentially, resulting in a double-stage attack (DSA). We have also adopted a cosine annealing strategy to adjust the step size throughout the attack process, including the initialization phase, to mitigate the oscillation of perturbations near the solution due to excessive step size. Extensive experiments and analysis regarding the proposed attack scheme have been conducted, and we hope our study can draw researchers' attention to initialization and serve as a reference for evaluating model robustness and enhancing AT's defensive effectiveness.

The rest of the paper is organized as follows. Section II discusses the adversarial attack mechanism and reviews recent developments in attacks and defenses. Section III describes the motivation behind our proposed attack and its specific implementation. Section IV presents experimental results and analysis. Section V discusses this study's implications.

II. Preliminaries and Related Work

1. Attack principle

We consider a multi-classifier $f(\cdot)$ and an input image x with its ground-truth label y. Here, x and label y belong to the sets $[0,1]^D$ and $\mathbb{R}^C$, respectively, where D denotes the number of channel dimensions and C indicates the total number of classes. The classifier's prediction can be expressed as $f\left(x\right)=\underset{c=1,2,\dots,C}{\operatorname{argmax}}\ \boldsymbol{z}_c(x)$, where $\boldsymbol{z}$ signifies the non-normalized prediction vectors produced by the classifier, and the subscript c corresponds to the predicted value for each class. Adversarial regions are typically defined as a high-dimensional space $B_\epsilon(x)=\left\{x^{\text{adv}}:\left\|x^{\text{adv}}-x\right\|_p\leq\epsilon\right\}$ centered on the image x, where the perturbation threshold is ϵ under the constraints of l_p norm (for instance, l_1, l_2, and l_∞). Untargeted attacks aim to generate a perturbed example x^{adv} that confuses the classifier, thereby producing a probability vector not aligned with the original label $f\left(x^{\text{adv}}\right)\neq y$. Conversely, targeted attacks strive to alter the prediction to a specific class $f\left(x^{\text{adv}}\right)=t$, where $t\neq y$.

Adversarial perturbations are artificially imposed noise by the attacker to achieve the goal of reducing the test accuracy of classifiers. To successfully launch an untargeted attack, the attacker should construct a suitable loss function $\mathcal{L}$ such that the loss value $\mathcal{L}(\boldsymbol{z}(x^{\text{adv}}),y)$ is large enough when $f\left(x^{\text{adv}}\right)\neq y$ to generate the perturbation. The better feasible solution of the perturbation is found by maximizing the loss value $\mathcal{L}(\boldsymbol{z}(x^{\text{adv}}),y)$ through gradient ascent while satisfying the constraint on the perturbation size. The PGD-based attack, the most frequently used method for evaluating robustness, can be described as follows:

$$x_{k+1}^{\text{adv}}=\mathit{\Pi}_{B_\epsilon(x)}(x_k^{\text{adv}}+\alpha_k\cdot\operatorname{sign}(\nabla_x\mathcal{L}(\boldsymbol{z}(x_k^{\text{adv}}),y)))\tag{1}$$

Here, α_k and x_k^{adv} represent the optimization step size and the corresponding generated AEs in the kth iteration, respectively. $B_\epsilon(x)(\cdot)$ is the operation that confines the perturbation to the adversarial region. Typical loss functions in adversarial attacks often include CE loss and maximum marginal (MM) loss.

2. White-box attacks

Following the discovery of DNNs' vulnerability to AEs, researchers have devised various white-box attack methods. Fast gradient sign method (FGSM) [7] uses one-step gradient ascent with CE loss, and iterative fast gradient sign method (I-FGSM) [8] is its multi-step variant. PGD [2] adds random noise in initial starting points. Carlini *et al.* [9] propose an optimization-based attack approach, C&W, while introducing MM loss to produce better attacks. Fast adaptive boundary (FAB) attack [10] utilizes geometric motivation to find minimal perturbations. Guided adversarial margin attack (GAMA) [4] guides stronger attacks by function mapping of clean images. This paper posits that the discrepancy between model classification outputs and labels can be harnessed as prior knowledge to guide subsequent attacks. AutoPGD [11] adaptively schedules the step size according to the optimization progress and proposes scaling invariant difference of logits ratio (DLR) loss. Furthermore, Croce *et al.* [11] fuse AutoPGD, FAB, and black-box square attacks [12] to form a parameter-free integrated attack, AutoAttack. While Tashiro *et al.* [3] are the first to focus on the role of output information and propose ODI. Ma *et al.* [5] propose margin decomposition (MD) attack, which dynamically adjusts the loss function with the attack process. Recently, Liu *et al.* [13] introduce adaptive auto attack (A^3), a new state-of-the-art attack semble that uses an adaptive initialization strategy (ADI). ADI uses ODI to obtain initialization points, generates perturbation components of the label and random target classes with symbolic priors, and samples the remaining components from a uniform distribution. Many prior works have introduced the concept of multiple random restarts in their attack mechanisms. In this study, we select seven of these attack methods for subsequent experiments. The integrated attack AutoAttack falls outside the scope of comparison due to its inclusion of black-box methods.

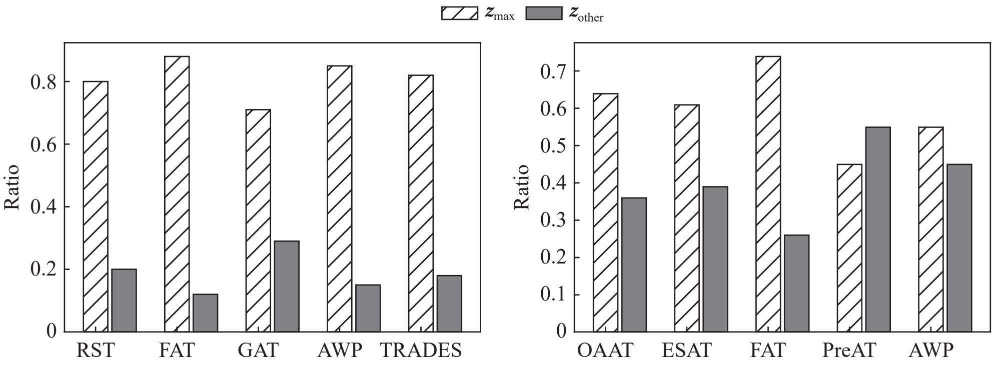
3. Adversarial defense

AT currently stands as the most effective defense strategy for enhancing model robustness, achieved by incorporating AEs in the training process. Zhang *et al.* [14] propose a feature scattering-based AT method called FeaScatter. AdvInterp [15] employs an adversarial interpolation scheme to produce AEs along with their respective adversarial labels. Robust self-training (RST) [16] utilizes unlabeled data to expand the self-training process. A notable work in this field, TRADES, devised by Zhang *et al.* [17], partitions the robustness error into two components, natural error and boundary error, thereby balancing the prediction accuracy of clean images and the model's robustness. This paper refers to the method of using pre-training to enhance robustness in [18] as PreAT, which will be used for subsequent comparisons. To encourage robust generalization, adversarial weight perturbation (AWP) [19] introduces a double-perturbation mechanism in AT. Rice *et al.* [20] demonstrate that early stopping is the most effective strategy to address robust overfitting, and we refer to AT process as ESAT in this paper. A method combining data augmentation and model weight averaging is proposed by Rebuff *et al.* [21], which we call FAT. Gowal *et al.* [22] focus on enhancing the original training set using generative models to reduce robustness accuracy gaps, referred to as GAT for ease of reference. Addepalli *et al.* [23] explore means of defense beyond perceptual limitations and propose oracle-aligned adversarial training (OAAT). Salman *et al.* [24] focus on the adversarial robustness of ImageNet and examine its integration with transfer learning, referred to as TAT. Wong *et al.* [25] find that AT with FGSM results in good defensive results, which we call FGAT. Several works, including [26]–[28], have investigated the significant impact of AEs on DNN-based few-shot classification models and have combined AT with meta-learning in few-shot scenarios. We will discuss and evaluate the above AT-based models under various attacks including our approach in Section IV.

III. Proposed Method

1. Motivation

In the standard untargeted PGD attack process, the naive PGD combines the original example x and random noise from the uniform distribution $U(-\epsilon, \epsilon)$ as the perturbation's starting point. We observe an intriguing phenomenon while implementing this process on various datasets and models. As depicted in Figure 1, we compute the proportion of target classes for successful AEs. The $\boldsymbol{z}_{\max}$ class with maximal logits is not dominant in the successful adversarial images. A significant number of images have the target classes as the original $\boldsymbol{z}_{\text{other}}$ classes. Even the weight from $\boldsymbol{z}_{\text{other}}$ classes exceeds that from the $\boldsymbol{z}_{\max}$ class on PreAT. That is, although optimizing in the direction of the $z_{\max}$ class increases the chances of obtaining an AE, the other classes can still guide the attack process to a certain degree. Thus, we think it is necessary to consider the contribution of the remaining classes. Moreover, the perturbation's effectiveness in the input space might lessen with the forward propagation of DNN. Tashiro *et al.* [3] propose introducing randomness in the output space to evade the negative effect of forward propagation on the perturbation. ODI uses a method where the random vector is multiplied by the model output to determine the initial loss. However, this approach has two drawbacks. Firstly, it does not differentiate between the attack potential of different classes. The model output reflects the attack difficulty of each class to some extent, and the use of randomized vectors confuses this information. Secondly, it is unclear whether the random noise introduced is beneficial for adversarial attacks. Random noise may be imposed on the logits of



(a) CIFAR-10 results

(b) CIFAR-100 results

Figure 1 Statistical results of classes on the adversarial examples.

the right classes, which could encourage the model to classify images correctly. In contrast, we want to backpropagate the relative relationships between logits of other classes besides the label class through the loss function. We hope this guide will enable the subsequent attack to find feasible directions as soon as possible and thus converge faster.

Developing an initialization method with perturbation diversity is crucial for both evaluating existing defenses and constructing AEs during AT. Faster converging untargeted attacks can refine AT's AE generation strategy, which is particularly beneficial for single-step defenses as they rely more heavily on the initial direction than multi-step methods.

2. Loss function

The surrogate losses optimized in existing white-box attacks are not ideally suited for direct use during the initialization phase. For instance, the MM loss is defined as follows:

$$\mathcal{L} = \left|\max_{i \neq y}(\boldsymbol{z}_i(x^{\text{adv}})) - \boldsymbol{z}_y(x^{\text{adv}})\right| \tag{2}$$

where y denotes the ground-truth label. As evident from (2), the MM loss optimizes solely for the class with the largest prediction vector $\boldsymbol{z}_i$, neglecting other classes.

According to the analysis in the preceding subsection, incorporating the contributions of other classes during the initialization phase is crucial for designing attack algorithms that converge faster. By acknowledging the potential of each class, we can pinpoint more effective initial points and directions. To address this need, we propose the one-hot MSE loss, which resembles the MSE:

$$\mathcal{L} = \frac{1}{C}\sum_{i=1}^{C}(\text{softmax}(\boldsymbol{z}_i(x^{\text{adv}})) - \text{one\_hot}(y))^2 \tag{3}$$

Here, $\text{softmax}(\cdot)$ normalizes the model output, $\text{one\_hot}(\cdot)$ represents the one-hot form of the label, and C stands for the total number of classes. The equation computes the mean distance between each class output of the DNN and the ground-truth label vector. We believe the one-hot MSE loss solves two shortcomings of ODI. Firstly, the discrepancy between the normalized prediction vector $\text{softmax}(\boldsymbol{z}_i(x^{\text{adv}}))$ and $\text{one\_hot}(y)$ retains the relative relationship between the initial logits of each class. Secondly, it compels the initial optimization to veer away from the label class. This idea of distinguishing label classes from other classes aligns with ADI. While our initialization method emphasizes maintaining the size relationship between logits, ADI generates multiplicative vectors based on prior knowledge.

The effectiveness of the MM loss has been proven in the C&W attack, which reveals the false robustness of the defensive distillation. We designate it as the loss function for the subsequent iterative process.

3. Double-stage attack

In this paper, we introduce a double-stage variant of PGD to check the enhancement of the designed one-hot MSE loss on the attack effect. This involves merging the one-hot MSE loss-guided initialization process with PGD. The iterative process of the subsequent PGD is based on the perturbations accrued in the initial phase. Simultaneously, we employ cosine annealing as a step size adjustment strategy to avoid oscillations between locally optimal solutions. The complete attack process is described in Algorithm 1.

Algorithm 1 Double-stage attack

```
Input: clean image $x$ with label $y$, classifier $f(\cdot)$, initial value of step size $\alpha$, total number of iterative steps $T$, initialization steps $T_{\text{init}}$, and restarts $R$;
Output: adversarial example $x^{\text{adv}}$;
for $R \in \{1, 2, ..., n\}$ do
    $x_0^{\text{adv}} \leftarrow x + \delta_{\text{random}}$;
    for $t \leftarrow 0$ to $T-1$ do
        $\alpha_t \leftarrow \text{consine\_annealing}(\alpha)$;
        if $t \leq T_{\text{init}}$ then
            $\mathcal{L} = \frac{1}{C}\sum_{i=1}^{C}(\text{softmax}(\boldsymbol{z}_i(x_t^{\text{adv}})) - \text{one\_hot}(y))^2$;
        else
            $\mathcal{L} = |\max_{i \neq y}(\boldsymbol{z}_i(x_t^{\text{adv}})) - \boldsymbol{z}_y(x_t^{\text{adv}})|$;
        end if
        $x_{t+1}^{\text{adv}} = \Pi_{B_\epsilon(x)}(x_t^{\text{adv}} + \alpha_t \cdot \text{sign}(\nabla_x \mathcal{L}(\boldsymbol{z}(x_t^{\text{adv}}), y)))$;
        if $f(x_{t+1}^{\text{adv}}) \neq y$ then
        return $x_{t+1}^{\text{adv}}$;
        end if
      end for
  end for
return $x_T^{\text{adv}}$.
```

IV. Evaluation Results and Analysis

1. Setup

Our collection of robust models is derived from various recent studies [16]–[25]. The weights of these models can be found in the open-source RobustBench [29]. We assess the performance of various baseline white-box attacks on CIFAR-10, CIFAR-100, and ImageNet datasets. In the subsequent discussion, we interchangeably use robust accuracy (RA) and attack success rate (equivalent to $1-\text{RA}$) to measure the potency of each attack. APGD, GAMA, MD, and FAB are executed using their official codes with relevant parameter settings maintained. Unless specified, the initial value of step size α is set at the standard value of $2/255$ for attacks that require manual adjustment of the step size, such as PGD, ODI-PGD, and our proposed attack. For our attack, we use the cosine annealing strategy to modify α during the attack. We set the perturbation threshold ϵ at $8/255$ for CIFAR-10 and CIFAR-100, $4/255$ for Imagenet, and the total number of iterative steps T to 100.

2. Validation of one-hot MSE loss effectiveness

We undertake two experiments: 1) We count the number of test images that present changes in their $\boldsymbol{z}_{\max}$ values on five adversarially trained models. 2) We select alternative loss functions for the initialization phase and compare their attack performance and computational overhead. For this purpose, the loss functions are replaced by CE and the two components of MM loss, which are the maximal softmax value of the model prediction $\boldsymbol{z}_{\max}$ and the softmax value corresponding to the label $\boldsymbol{z}_y$. For both experiments, the number of initialization steps T_{init} is fixed at 8, the initial value of step size α is set to 8/255, and we use the CIFAR-10 dataset.

Figure 2 shows the classification prediction changes after initialization using one-hot MSE loss and random noise, with more significant changes signifying a higher likelihood of successful attacks. The one-hot MSE initialization triggers more $\boldsymbol{z}_{\max}$ alterations for both the non-robust models (FeaScatter [14] and AdvInterp [15]) and the robust models (AWP [19], FAT [21], and GAT [22]). This visual validation experiment demonstrates the qualitative improvement of the perturbation starting point of one-hot MSE loss compared to random noise.

In the comparison experiments where the initialization loss function is altered, one back propagation (BP) corresponds to a single forward propagation. Consequently, the computational overhead, represented as Cost, can be measured by the number of BPs. RA signifies the model's classification accuracy when confronted with AEs. A lower RA indicates a more potent attack, whereas a lower Cost suggests that the algorithm is more efficient in finding AEs and consumes fewer computational resources. As seen from Table 1, the differences in the attack success rates of these four loss functions are minor, implying comparable initialization effects. However, regarding computational overhead, one-hot MSE can significantly reduce the number of BPs. In general, the initial points generated by the one-hot MSE loss exhibit greater attack potential. Hence, it is logical to consider the contribution of other classes.

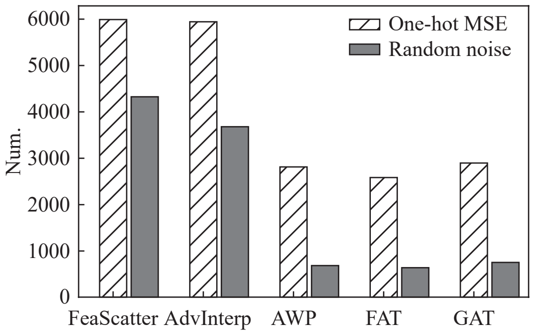


Figure 2 Variation of $\boldsymbol{z}_{\max}$.

Table 1 Results of different initial loss functions

Models	CE		$\boldsymbol{z}_{\max}$		$-\boldsymbol{z}_y$		One-hot MSE	
	RA (%)	Cost	RA (%)	Cost	RA (%)	Cost	RA (%)	Cost
FAT	56.83	773450	56.78	1038985	56.79	849810	**56.77**	**605203**
GAT	58.97	789090	58.76	1083835	**58.75**	912715	58.80	**622272**
RST	59.81	806800	59.72	1109020	**59.70**	873500	**59.70**	**633302**
AWP	56.26	765745	56.19	1059685	**56.18**	825775	56.25	**600091**
TRADES	52.77	720205	**52.63**	1053360	52.65	802660	52.77	**567478**

3. Attack performance comparison under limited computational budget constraints

A restart is one of the most demanding settings in terms of computational budget. To validate the feasibility of our proposed attack method (DSA), we focus on verifying its superior performance under a random restart. We select multiple defense models for three datasets to demonstrate that the results are not accidental. For clarity, we have abbreviated PGD and APGD to P and AP, respectively.

Table 2 presents the results of untargeted attacks on CIFAR-10 and CIFAR-100, where we restrict the number of restarts to demonstrate the effectiveness of the DSA method. In most cases, our approach achieves the highest attack success rate compared to other baseline attacks. Notably, DSA surpasses ODI-PGD by an additional 0.93%–2.16% on CIFAR-10 and 0.94%–2.22% on CIFAR-100. This demonstrates that using one-hot MSE to guide the generation of initial points is superior to the random noise initialization strategy in the output space. Moreover, DSA achieves better attack performance with only one restart compared to the existing state-of-the-art method A^3. A^3 uses ODI early in the attack and inevitably suffers from its drawbacks. Consistent results are also seen on the ImageNet dataset in Table 3.

In the meantime, we evaluate the time required to conduct the attack and the perturbation size. Figures 3–5 show that our method generates medium-amplitude perturbations with less time overhead. Regarding the l_∞ norm, the attack methods perform similarly, except for FAB, which is particularly dominant. However, the small perturbation of FAB comes at the cost of enormous time consumption. Specifically, for larger datasets such as ImageNet, the time overhead of FAB is an order of magnitude greater than that of other attacks. Therefore, FAB is not included in the ImageNet experiments.

In Figures 6–8, the original samples are shown in the first row, while the second row depicts the adversarial

Table 2 Adversarial robustness in $T = 100$ and $R = 1$ setting on CIFAR-10 and CIFAR-100

Dataset	Models	RA (%)										
		$\mathrm{P_{CE}}$	$\mathrm{P_{MM}}$	$\mathrm{P_{DLR}}$	$\mathrm{AP_{CE}}$	$\mathrm{AP_{DLR}}$	FAB	GAMA	MD	ODI-PGD	$\mathrm{A^3}$	DSA
CIFAR-10	FAT [21]	59.64	57.27	57.26	59.57	57.29	57.37	56.88	56.86	57.70	57.48	**56.77**
	GAT [22]	61.06	60.46	60.86	60.84	60.84	59.57	58.98	59.10	60.96	60.71	**58.80**
	RST [16]	62.06	60.63	60.90	61.83	60.91	62.04	59.80	59.79	61.00	61.42	**59.70**
	AWP [19]	58.84	56.75	56.82	58.80	56.90	56.83	56.32	56.29	57.49	57.24	**56.25**
	TRADES [17]	54.88	53.58	53.72	54.67	53.67	53.50	52.78	52.86	54.12	53.94	**52.77**
CIFAR-100	FAT [21]	31.89	29.25	29.29	31.78	29.17	28.99	28.72	**28.62**	29.63	29.27	28.69
	AWP [19]	33.54	30.98	31.44	33.27	31.31	29.32	29.36	29.32	31.53	31.16	**29.31**
	PreAT [18]	33.76	30.77	32.08	32.95	32.07	29.02	29.15	28.98	30.95	31.06	**28.96**
	ESAT [20]	20.73	20.28	20.51	20.60	20.43	20.21	19.26	19.42	20.96	20.58	**19.25**
	OAAT [23]	32.94	27.93	28.06	32.84	28.01	27.93	27.53	27.54	28.57	28.36	**27.52**

Table 3 Model robustness in $T = 100$ and $R = 1$ setting on ImageNet

Models	RA (%)									
	$\mathrm{P_{CE}}$	$\mathrm{P_{MM}}$	$\mathrm{P_{DLR}}$	$\mathrm{AP_{CE}}$	$\mathrm{AP_{DLR}}$	GAMA	MD	ODI-PGD	$\mathrm{A^3}$	DSA
RLAT [30]	31.90	31.41	32.00	31.11	31.11	**27.62**	28.22	31.31	37.68	**27.62**
FGAT [25]	25.72	24.03	25.22	25.12	24.33	22.13	23.33	25.02	37.66	**21.93**
TAT_R18 [24]	28.51	24.93	25.02	27.82	24.83	**23.63**	**23.63**	25.62	27.40	23.73
TAT_R50 [24]	37.79	34.60	35.00	36.69	34.90	**32.50**	**32.50**	35.00	37.68	**32.50**

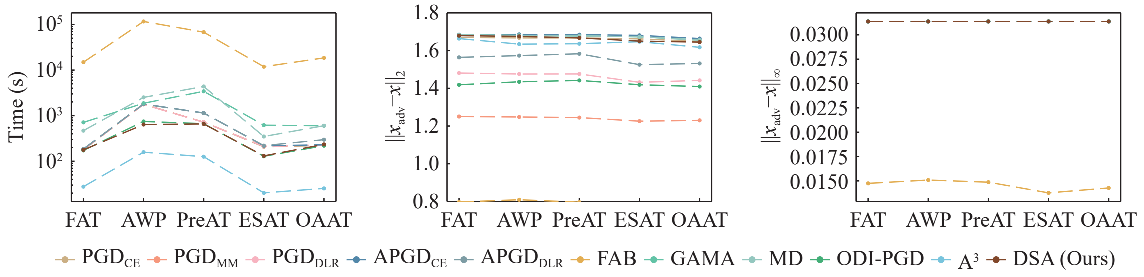


Figure 3 Time cost and perturbation on CIFAR-100.

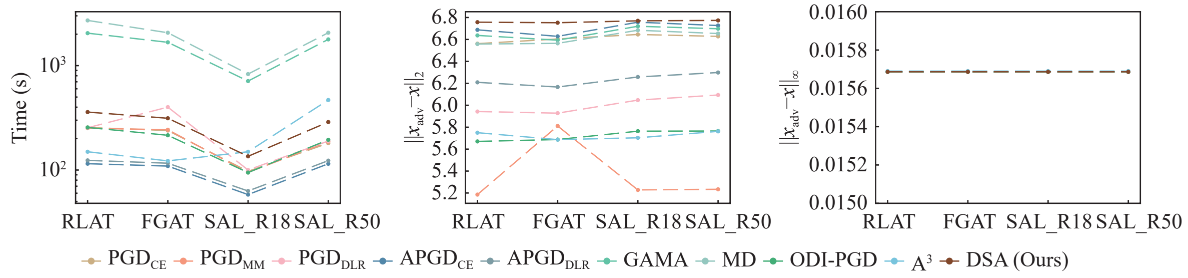


Figure 4 Time cost and perturbation on ImageNet.

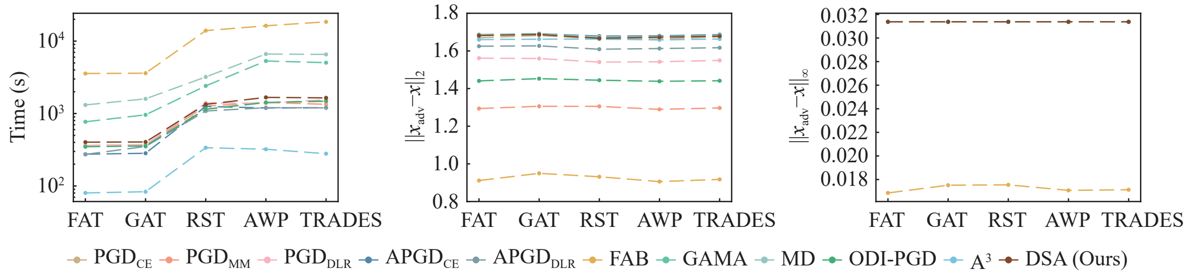


Figure 5 Time cost and perturbation on CIFAR-10.

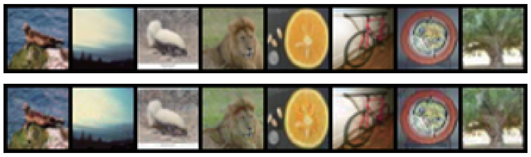


Figure 6 Perturbation scale's visualization results on CIFAR-100.

Figure 7 Perturbation scale's visualization results on ImageNet.

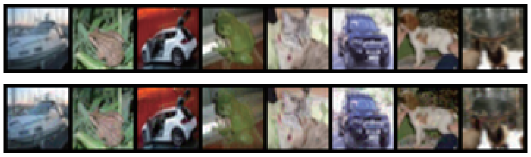


Figure 8 Perturbation scale's visualization results on CIFAR-10.

samples. It is difficult for humans to distinguish the adversarial samples from the original samples, indicating that the perturbations generated by our method are well hidden.

4. Ablation study: understanding the impact of individual components on overall performance

In this section, we dissect two key components of our proposed method: 1) the cosine annealing strategy and 2) one-hot MSE.

Cosine annealing strategy Tables 4 and 5 reveal that applying the cosine annealing strategy, which adjusts the attack step size, decreases the adversarial robustness. This emphasizes the cosine annealing strategy's vital role in implementing highly effective attacks.

One-hot MSE loss To ascertain the value of the one-hot MSE loss-guided initialization phase in white-box attacks, we evaluate two variants: one with and the other without one-hot MSE loss. The outcomes, as depicted in Tables 4 and 5, demonstrate that applying one-hot MSE loss in isolation can significantly reduce RA. Hence, this component is paramount for the success of an attack.

In conclusion, integrating both components leads to the lowest RA for the model. This highlights the substantial contribution each component provides to the efficacy of the attack.

5. Parameter analysis: exploring the impact of restarts on the attack

We delve further into the sensitivity of various attacks concerning the number of restarts. The results are illustrated in Figures 9–11. Overall, our proposed method delivers the most potent attacks. The detailed analysis is as follows:

1) When $R = 1$, our method tends to yield the lowest initial RA. Compared to ODI-PGD, which relies heavily on multiple random restarts, DSA shows rapid convergence as the number of restarts escalates. This suggests that DSA can effectively search for AE under given attack conditions (evaluation model, perturbation upper limit, and iteration count).

2) Intriguingly, the performance of several attacks with relatively high success rates, such as GAMA, MD, and DSA, does not markedly improve with the addition of restarts. This indicates that these attacks will likely find feasible AEs under specific parameter configurations during the initial or first few restarts. Among these attacks, DSA consistently maintains the highest or sec-

Table 4 Ablation experiments on CIFAR-10 and CIFAR-100

	One-hot MSE	Cosine annealing	FAT [21]	GAT [22]	RST [16]	AWP [19]	TRADES [23]
CIFAR-10			57.38	60.54	60.90	57.02	53.72
		√	57.31	60.51	60.67	56.84	53.58
	√		56.89	58.84	59.96	56.39	52.90
	√	√	**56.77**	**58.80**	**59.70**	**56.25**	**52.77**
	One-hot MSE	Cosine annealing	FAT [21]	AWP [19]	PreAT [18]	ESAT [20]	OAAT [23]
CIFAR-100			29.23	31.04	30.90	20.44	28.12
		√	29.22	30.93	30.82	20.29	27.94
	√		28.80	29.38	29.27	19.37	27.65
	√	√	**28.69**	**29.31**	**28.96**	**19.25**	**27.52**

Table 5 Ablation experiments on ImageNet

One-hot MSE	Cosine annealing	FGAT [25]	RLAT [30]	TAT_R18 [24]	TAT_R50 [24]
		24.83	31.41	25.22	34.70
	√	24.03	30.81	24.83	34.60
√		22.83	28.02	23.83	32.70
√	√	**21.93**	**27.62**	**23.73**	**32.50**

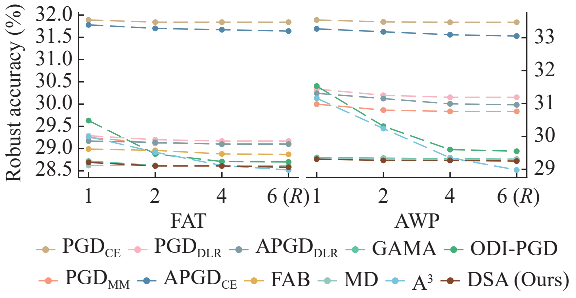


Figure 9 Results corresponding to different numbers of restarts on CIFAR-100.

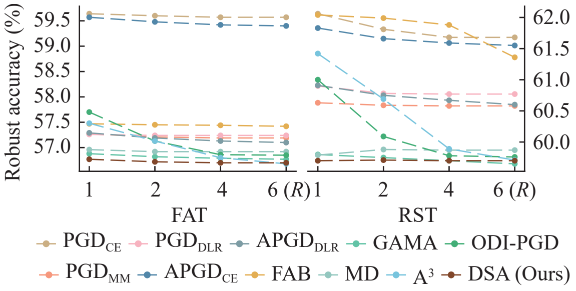


Figure 10 Results corresponding to varying numbers of restarts on CIFAR-10.

ond-highest success rate, highlighting the feasibility and superiority of the DSA algorithm.

3) In contrast to the adaptive A^3, our method employs a straightforward iteration strategy with lesser complexity, treating all images equally in each restart. Hence, if the number of restarts is considered a constraint, DSA outperforms at a minimal number of restarts. This occurs because A^3 adaptively modifies the iteration assignment and direction using loss function values, which is an attack strategy that requires time for accumulation. Instead, we utilize loss information to calibrate the initial direction of perturbation, yielding more prompt results. However, when the attack progresses to $R = 6$, A^3 occasionally outperforms, attributable to the initial random noise in multiple restarts and the online statistics-based discarding strategy (OSD).

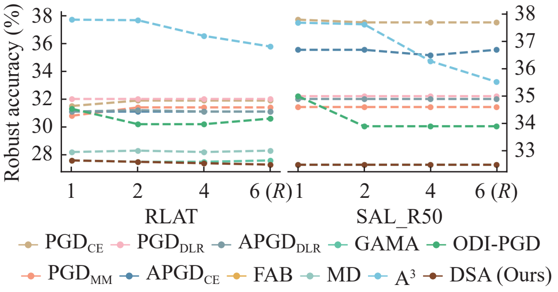


Figure 11 Results corresponding to varying numbers of restarts on ImageNet.

6. Extended application of one-hot MSE loss

Here, we aim to integrate one-hot MSE initialization into the AT process. Meta-learning models for few-shot scenarios belong to models that necessitate superior generalization. For the current defense strategy, we only tweak the loss function of the adversarial meta-learning inner process. We set the initialization step $T_{\text{init}} = 1$ and maintain the rest of the settings. The three adversarial meta-learning algorithms' default number of attack iterations is 1. The model's robustness is evaluated comprehensively using the harmonic mean accuracy (i.e., $H = \frac{2\times \text{SA}\times \text{RA}}{\text{SA}+\text{RA}}$). Table 6 respectively presents the adversarial robustness of the MiniImagenet and CIFAR-FS datasets. The model combined with one-hot MSE initialization consistently achieves higher robustness accuracy than merely increasing the number of iterations or adding ODI. This suggests that this initialization, which amplifies the adversarial strength of examples, enhances the model's defensive performance.

V. Discussion

In this paper, we dissect the shortcomings of exist-

Table 6 Standard accuracy (SA), robust accuracy (RA), and harmonic mean accuracy (H) on MiniImagenet and CIFAR-FS

Methods	MiniImagenet						CIFAR-FS					
	1-shot 5-way			5-shot 5-way			1-shot 5-way			5-shot 5-way		
	SA (%)	RA (%)	H (%)	SA (%)	RA (%)	H (%)	SA (%)	RA (%)	H (%)	SA (%)	RA (%)	H (%)
ADML [26]	31.96	24.38	27.66	55.13	36.33	43.80	39.45	26.44	31.66	57.20	31.57	40.69
2 step	31.54	25.00	27.89	54.80	37.00	44.17	39.04	26.86	31.82	55.86	32.52	41.11
+ ODI	31.00	24.41	27.31	**57.20**	35.30	43.66	**41.16**	24.74	30.90	**58.00**	29.76	39.34
+ one-hot MSE	**33.47**	**25.80**	**29.14**	55.37	**37.38**	**44.63**	39.38	**26.95**	**32.00**	56.25	**33.03**	**41.62**
AQ [27]	28.15	23.85	25.82	49.17	38.20	43.00	31.03	26.27	28.45	49.95	35.55	41.54
2 step	27.86	24.02	25.80	49.54	38.35	43.23	31.50	26.05	28.52	50.10	35.94	41.85
+ ODI	**28.54**	24.33	26.27	**50.70**	38.55	**43.80**	**34.03**	25.59	29.21	**51.56**	34.28	41.18
+ one-hot MSE	28.52	**24.85**	**26.56**	49.66	**39.18**	**43.80**	32.50	**27.03**	**29.51**	50.90	**36.43**	**42.47**
R-MAML [28]	37.87	23.41	28.93	56.20	34.57	42.81	**42.38**	22.33	29.25	57.57	31.05	40.34
2 step	36.62	23.68	28.76	56.20	36.90	44.55	40.50	23.58	29.81	55.60	32.08	40.69
+ ODI	**40.65**	20.69	27.42	**57.70**	33.03	42.01	41.36	20.15	27.10	**57.67**	28.66	38.29
+ one-hot MSE	38.77	**24.05**	**39.18**	56.70	**37.23**	**44.95**	41.14	**23.96**	**30.28**	55.90	**32.90**	**41.42**

ing white-box attacks and propose a unique loss function for the initialization stage. Building on this, we introduce the double-stage attack (DSA), an efficient generator of adversarial examples in a single-restart setting. Our experimental outcomes on various datasets demonstrate that the DSA approach surpasses various other white-box attack methods in terms of success rate. Furthermore, the DSA method is straightforward, offering a pragmatic solution to evaluate adversarial robustness. We have made preliminary attempts to incorporate our attack into the adversarial training process, aiming, in the future, to expose potential vulnerabilities in the defense algorithm and to guide the development of more robust defense strategies by generating a wider array of stealthier adversarial examples using DSA.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61972306), SongShan Laboratory (Grant No. YYJC012022005), and Zhejiang Laboratory (Grant No. 2021KD0AB03).

References

[1] S. D. Zhang, H. C. Gao, and Q. X. Rao, "Defense against adversarial attacks by reconstructing images," *IEEE Transactions on Image Processing*, vol. 30, pp. 6117–6129, 2021.

[2] A. Madry, A. Makelov, L. Schmidt, *et al.*, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, https://arxiv.org/pdf/1706.06083.pdf, 2018.

[3] Y. Tashiro, Y. Song, and S. Ermon, "Diversity can be transferred: Output diversification for white-and black-box attacks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, article no. 381, 2020.

[4] G. Sriramanan, S. Addepalli, A. Baburaj, *et al.*, "Guided adversarial attack for evaluating and enhancing adversarial defenses," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, article no. 1704, 2020.

[5] X. J. Ma, L. X. Jiang, H. X. Huang, *et al.*, "Imbalanced gradients: A subtle cause of overestimated adversarial robustness," *arXiv preprint*, arXiv: 2006.13726, 2020.

[6] N. Antoniou, E. Georgiou, and A. Potamianos, "Alternating objectives generates stronger PGD-based adversarial attacks," *arXiv preprint*, arXiv: 2212.07992, 2022.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, https://arxiv.org/pdf/1412.6572.pdf, 2015.

[8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint*, arXiv: 1607.02533, 2017.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of 2017 IEEE Symposium on Security and Privacy*, San Jose, CA, USA, pp. 39–57, 2017.

[10] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," https://openreview.net/forum?id=HJlzxgBtwH, 2020.

[11] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual Event, article no. 206, 2020.

[12] M. Andriushchenko, F. Croce, N. Flammarion, *et al.*, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp. 484–501, 2020.

[13] Y. Liu, Y. Cheng, L. Gao, *et al.*, "Practical evaluation of adversarial robustness via adaptive auto attack," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 15105–15114, 2022.

[14] H. C. Zhang and J. Y. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Proceedings of the 32nd Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 1829–1839, 2019.

[15] H. C. Zhang and W. Xu, "Adversarial interpolation training: A simple approach for improving model robustness," https://openreview.net/forum?id=Syejj0NYvr, 2020.

[16] Y. Carmon, A. Raghunathan, L. Schmidt, *et al.*, "Unlabeled data improves adversarial robustness," in *Proceedings of the 32nd Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 11190–11201, 2019.

[17] H. Y. Zhang, Y. D. Yu, J. T. Jiao, *et al.*, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, pp. 7472–7482, 2019.

[18] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, pp. 2712–2721, 2019.

[19] D. X. Wu, S. T. Xia, and Y. S. Wang, "Adversarial weight perturbation helps robust generalization," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, article no. 249, 2020.

[20] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual Event, article no. 749, 2020.

[21] S. A. Rebuff, S. Gowal, D. A. Calian, *et al.*, "Fixing data augmentation to improve adversarial robustness," *arXiv preprint*, arXiv: 2103.01946, 2021.

[22] S. Gowal, S. A. Rebuff, O. Wiles, *et al.*, "Improving robustness using generated data," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Virtual Event, pp. 4218–4233, 2021.

[23] S. Addepalli, S. Jain, G. Sriramanan, *et al.*, "Towards achieving adversarial robustness beyond perceptual limits," in *Proceedings of the ICLR 2022*, Vienna，Austria, https://ieeexplore.ieee.org/document/9157734/, 2022.

[24] H. Salman, A. Ilyas, L. Engstrom, *et al.*, "Do adversarially robust ImageNet models transfer better?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, article no. 208, 2020.

[25] E. Wong, L. Rice, and J. Z. Kolter, "*Fast is better than free: Revisiting adversarial training*," in *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, https://arxiv.org/abs/2001.03994, 2020.

[26] C. X. Yin, J. Tang, Z. Y. Xu, *et al.*, "Adversarial meta-learning," https://openreview.net/forum?id=Z_3x5eFk1l-, 2021.

[27] M. Goldblum, L. Fowl, and T. Goldstein, "Adversarially robust few-shot learning: A meta-learning approach," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, article no. 1501, 2020.

[28] R. Wang, K. D. Xu, S. J. Liu, *et al.*, "On fast adversarial robustness adaptation in model-agnostic meta-learning," in *Proceedings of the 9th International Conference on Learning Representations*, Vienna, Austria, https://arxiv.org/abs/2102.10454, 2021.

[29] F. Croce, M. Andriushchenko, V. Sehwag, *et al.*, "RobustBench: A standardized adversarial robustness benchmark," in

Proceedings of the 1st Neural Information Processing Systems Track on Datasets and Benchmarks, Virtual Event, https://arxiv.org/abs/2010.09670v1, 2021.

[30] L. Engstrom, A. Ilyas, H. Salman, *et al.*, "Robustness (python library)," Available at: *https://github.com/Madry-Lab/robustness*, 2019.

Yunyi ZHOU received the B.E. degree in software engineering from Xidian University, Xi'an, China, in 2021. She is currently an M.S. candidate in software engineering with Xidian University, Xi'an, China. Her current research interest is deep learning security.
(Email: zhouyy0206@foxmail.com)

Haichang GAO received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2006. Currently, he is a Professor with the School of Computer Science and Technology, Xidian University, Xi'an, China, and a member of the IEEE. He has published more than 30 papers. Now he is in charge of a project of the National Natural Science Foundation of China. His current research interests include CAPTCHA, computer security, and deep learning security.
(Email: hchgao@xidian.edu.cn)

Jianping HE received the B.E. degree in software engineering from Xidian University, Xi'an, Shaanxi, China, in 2022. He is currently a M.S. candidate in computer science and technology with Xidian University. His current research interest is deep learning security.
(Email: 981274517@qq.com)

Shudong ZHANG received the Ph.D. degree in software engineering from Xidian University, Xi'an, China, in 2022. His current research interest is artificial intelligence security.
(Email: sdong.zhang@outlook.com)

Zihui WU received the B.E. degree in automation from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019. He is currently a Ph.D. candidate in computer science and technology with Xidian University, Xi'an, China. His current research interest is deep learning security.
(Email: zihui@stu.xidian.edu.cn)