STOCHASTIC VARIANCE REDUCED ENSEMBLE ADVERSARIAL ATTACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box adversarial attack has attracted much attention for its practical use in deep learning applications, and it is very challenging as there is no access to the architecture and weights of the target model. Based on the hypothesis that if an example remains adversarial for multiple models, then it is more likely to transfer to other models, the ensemble-based attack methods are efficient and widely used in the black-box setting. Nevertheless, existing ensemble-based approaches simply aggregate the outputs of all models but ignore the variance of different models, leading to a rather poor local optimum. To address this issue, we propose a stochastic variance reduced ensemble attack method to boost the performance of black-box adversarial attacks. By integrating the stochastic variance reduced gradient technique into the model ensemble attack, our method can balance the gradient of different models and leads to better local maximum, resulting in highly transferable adversarial examples. Empirical results on the standard ImageNet dataset demonstrate that our method can boost the ensemble attack performance and significantly improve the transferability of the generated adversarial examples.

1 INTRODUCTION

Deep neural networks (DNNs) have shown impressive performance on various computer vision tasks. However, recent researches have shown that DNNs are strikingly vulnerable to adversarial examples crafted by adding human-imperceptible perturbations (Szegedy et al., 2014; Goodfellow et al., 2015; Papernot et al., 2016). Moreover, adversarial examples are known to be transferable that the examples crafted for one model can also mislead other unknown models (Papernot et al., 2017; Liu et al., 2017; Moosavi-Dezfooli et al., 2017). Generating adversarial examples (*i.e.*, adversarial attack) has drawn enormous attention since it can help evaluate the robustness of different models (Carlini & Wagner, 2017; Tramer et al., 2020) and then improve their robustness by adversarial training (Goodfellow et al., 2015; Madry et al., 2018)

Various adversarial attack methods have been proposed, including optimization-based methods such as box-constrained L-BFGS (Szegedy et al., 2014) and Carlini & Wagner's method (Carlini & Wagner, 2017), gradient-based methods such as fast gradient sign method (Goodfellow et al., 2015) and its iterative variants (Kurakin et al., 2017a; Madry et al., 2018). In general, these adversarial attack methods can achieve high success rates in the white-box setting (Carlini & Wagner, 2017), where the attacker can access the complete information of the target model, including the model structure and gradient information. However, these methods often exhibit low attack success rates in the black-box setting (Dong et al., 2018), where the adversary can not access the information of the target model but can only utilize the transferability of adversarial examples to fool the unknown models.

Recently, many methods have been proposed to enhance the transferability of adversarial examples so as to improve the attack success rates in the black-box setting. These methods include the gradient-optimization attacks (Dong et al., 2018; Lin et al., 2020; Wang & He, 2021), input-transformation attacks (Dong et al., 2019; Xie et al., 2019b; Lin et al., 2020), and model ensemble attacks (Liu et al., 2017; Dong et al., 2018). Among these methods, the model ensemble attacks are efficient and have been broadly adopted in boosting the black-box attack performance (Xie et al., 2019b; Lin et al., 2020; Gao et al., 2020). As compared with the first two categories that many

methods have been proposed, the model ensemble attack is rather less investigated, and existing ensemble methods are actually very straightforward.

In this work, we observe that the existing model ensemble attacks just fuse the outputs of all models directly but ignore the variance of different models, which may limit the potential performance of the model ensemble attacks. However, the optimization paths of different models may differ widely, indicating there exists considerable difference on the variance of optimization directions among models. Such variance causes the optimization direction of ensemble attack to be less accurate. As a result, the attack performance of the transferred adversarial examples decays considerably.

To address the above issue, we propose a stochastic variance reduced ensemble (SVR-Ens) method to reduce the variance of different models during the adversarial attack, so as to improve the transferability of adversarial examples. Technically, at each iteration of the adversarial attack, we adopt the idea of stochastic variance reduced gradient method (Johnson & Zhang, 2013) to obtain a more accurate gradient, rather than the gradient from the ensemble models that simply fuses the outputs. In this way, the proposed SVR-Ens method can surpass the existing model ensemble attacks in both the white-box setting and the black-box setting. To the best of our knowledge, we are the first to investigate the limitation of existing ensemble attack through the lens of variance on multiple models. By adopting the stochastic variance reduced gradient method, our proposed method reduces the gradient variance to have a more stable gradient direction and avoid overfitting to the models being attacked, leading to a better local optimum, which helps improve the attack performance and transferability. Extensive experiments on the ImageNet dataset demonstrate that our proposed method consistently outperforms the vanilla ensemble model attack in both the white-box setting and the black-box setting.

2 Related Works

Let x and y be a benign image and the corresponding true label, respectively. Let $J(x, y)$ be the loss function of the classifier and $\mathcal{B}_\epsilon(x) = \{x' : \|x - x'\|_p \leq \epsilon\}$ be the L_p-norm ball centered at x with radius ϵ. The goal of non-targeted adversarial attacks is to search an adversarial example $x^{adv} \in \mathcal{B}_\epsilon(x)$ that maximize the loss $J(x^{adv}, y)$. To align with previous works, we focus on L_∞-norm non-targeted adversarial attacks.

2.1 Adversarial Attacks

Existing adversarial attacks for crafting adversarial examples can be categorized into three groups, namely gradient-optimization attacks (Goodfellow et al., 2015; Kurakin et al., 2017a; Dong et al., 2018; Lin et al., 2020), input transformation attacks (Dong et al., 2018; Xie et al., 2019b; Lin et al., 2020), and model ensemble attacks (Liu et al., 2017; Dong et al., 2018).

Gradient-optimization attacks. The most typical adversarial attack based on the gradient is the Fast Gradient Sign Method (Goodfellow et al., 2015), which uses the gradient direction of the loss function with respect to the input image to generate a fixed amount of perturbation. Kurakin et al. (2017a) propose the Basic Iterative Method (BIM) to run multiple iterations of FGSM with a small perturbation instead of a single step. Madry et al. (2018) propose the Projected Gradient Descent (PGD), which is a noisy version of BIM. Although PGD is effective in the white-box attack setting (Athalye et al., 2018), it overfits the target model easily and yields weaker transferability in the black-box attack setting. In order to improve the transferability of adversarial attacks, Dong et al. (2018) propose to boost the adversarial attack with momentum. More recently, Lin et al. (2020) introduce Nesterov accelerated gradient method into the gradient-based attack to look ahead effectively to avoid overfitting.

Input transformation attacks. Another line of adversarial attacks focus on adopting various input transformations to improve the transferability of adversarial examples. Xie et al. (2019b) propose the Diverse Input Method (DIM), which applies random resizing and padding to the input image before feeding the example to the classifier to improve the transferability. Dong et al. (2019) propose the Translation-Invariant Method (TIM), which evades the defense models by calculating the gradients over a set of translated images. To reduce the gradient calculation, Dong et al. (2019) also develops an efficient algorithm to calculate the gradients by convolving the gradient at untranslated images. Lin et al. (2020) propose the Scale-Invariant Method (SIM), which introduces the scale-invariant

property of deep learning models and calculates the gradient over a set of scaled images to avoid overfitting and improve the transferability.

Model ensemble attacks. The model ensemble attacks are first introduced by Liu et al. (2017), in which the predictions of multiple models are fused to get the loss of ensemble predictions, and they apply existing adversarial attacks (*e.g*. FGSM and PGD) to generate adversarial examples. Dong et al. (2018) propose two variants of model ensemble attacks, namely fusing the logits and fusing the losses, respectively. Compared with various explorations on the gradient optimization or input transformation attacks, the model ensemble attacks are far less investigated, and existing methods only simply fuse the output predictions/logits/losses. In this work, motivated by the fact that there is great variance of different models, we propose a stochastic variance reduced ensemble attack to boost the adversarial attacks.

2.2 Adversarial Defenses

As the counterpart of adversarial attacks, numerous methods have been proposed to defend against adversarial examples. One intuitive approach is the *adversarial training* (Szegedy et al., 2014; Tramèr et al., 2018; Madry et al., 2018; Song et al., 2019; Xie et al., 2019a; Zhai et al., 2019; Song et al., 2020), which augments the training data by generating adversarial examples during the training process. Tramèr et al. (2018) propose ensemble adversarial training, which augments the training data with perturbations transferred from other models, in order to further improve the robustness against black-box attacks. Madry et al. (2018) propose PGD-Adversarial Training (PGD-AT), which augments the training data with adversarial examples crafted by PGD attack. Xie et al. (2019a) develop new network architectures that increase adversarial robustness by performing feature denoising. By combining PGD-AT with feature denoising networks, they further improve the adversarial robustness. Although adversarial training is promising, it is computationally expensive and is hard to scale to large-scale datasets (Kurakin et al., 2017b).

Another line of defense aims to diminish the adversarial perturbations from the input data. Guo et al. (2018) find that many image transformations, such as JPEG compression, have the potential to remove adversarial perturbations while preserving the visual information of the images. Xie et al. (2018) mitigate adversarial effects through randomized transformations, including resizing and padding (R&P). Liao et al. (2018) use high-level representation guided denoiser (HGD) to purify the adversarial images. Xu et al. (2018) propose two feature squeezing methods: bit reduction (Bit-Red) and spatial smoothing to detect adversarial examples. Liu et al. (2019) propose the feature distillation (FD), which adopts a JPEG-based defensive compression framework to diminish adversarial perturbations. Jia et al. (2019) propose the ComDefend, which utilizes an end-to-end image compression model to defend against adversarial examples. Jia et al. (2020) leverage randomized smoothing (RS) to train a certifiably robust ImageNet classifier. Naseer et al. (2020) develop a neural representation purifier (NRP) model, which learns to purify the adversarially perturbed images through automatically derived supervision.

3 Methodology

We focus on addressing the transferability through the lens of reducing the variance of ensemble model attack, and propose a stochastic variance reduced model ensemble attack method, which can be integrated with any existing gradient-based attack method to boost the attack transferability. Since our proposed method is based on the model ensemble attack, we first introduce the existing ensemble attack methods, then present our motivation and elaborate our method in detail.

3.1 Ensemble Attack Method

Ensemble attack method, which is an effective strategy to improve the transferability of adversarial examples, is first studied by Liu et al. (2017) and has been extended by Dong et al. (2018). The basic idea of ensemble attack is to generate adversarial examples for the ensemble models.

Ensemble on Predictions. Liu et al. (2017) propose to achieve an ensemble attack by averaging the predictions (softmax outputs of logits) of K models as: $\boldsymbol{p}(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \boldsymbol{p}_k(\boldsymbol{x})$, where $\boldsymbol{p}_k$ is the prediction of the k-th model, and $w_k \geq 0$ is the ensemble weight constrained by $\sum_{k=1}^{K} w_k = 1$.

Then, the loss function of an ensemble model is defined as:

$$J(\boldsymbol{x}, y) = -\mathbf{1}_y \cdot \log(\boldsymbol{p}(\boldsymbol{x})), \tag{1}$$

where $\mathbf{1}_y$ is the one-hot encoding of the ground-truth label y of $\boldsymbol{x}$.

Ensemble on Logits. Dong et al. (2018) propose to fuse the logits (the output before the softmax) for an ensemble of K models as: $\boldsymbol{l}(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \boldsymbol{l}_k(\boldsymbol{x})$, where $\boldsymbol{l}_k$ is the logits information of the k-th model, Then, the loss function of an ensemble model is defined as:

$$J(\boldsymbol{x}, y) = -\mathbf{1}_y \cdot \log(\mathrm{softmax}(\boldsymbol{l}(\boldsymbol{x}))), \tag{2}$$

which is the same as in Eq. 1.

Ensemble on Loss. Dong et al. (2018) introduce an alternative ensemble attack of (Liu et al., 2017). Specifically, they achieve the ensemble attack by averaging the loss of K models as:

$$J(\boldsymbol{x}, y) = \sum_{k=1}^{K} w_k J_k(\boldsymbol{x}, y), \tag{3}$$

where J_k is the loss of the k-th model .

3.2 Rethinking of the Ensemble Attack

The ensemble attack method has been broadly adopted in enhancing the performance of black-box attacks (Liu et al., 2017; Dong et al., 2018; Xie et al., 2019b; Lin et al., 2020; Gao et al., 2020). To the best of our knowledge, however, most of the existing researches only utilize the ensemble attack strategy as a plug-and-play to enhance their proposed attack methods, but did not delve into the ensemble attack method itself.

Intuitively, the existing ensemble attack methods are useful for improving the adversarial transferability because attacking an ensemble model can help to find a better local maxima and makes it easy to generalize to other black-box models. However, merely averaging the outputs (logits, predictions or loss) of all models to build an ensemble model for adversarial attack may limit the attack performance, as the variance of different models is ignored, which may lead to a rather poor local optimum. The optimization path of different models may vary significantly, indicating there exists a considerable gap in the variance of optimization directions among models. Simply fusing the outputs of the models but ignore the variance of different models would lead to a suboptimal result, which limits the performance on both white-box and black-box ensemble attacks. Moreover, existing ensemble attack methods compute the gradient information at each iteration, in which previously computed gradient information is not utilized. After all, the current iteration point, *i.e*. adversarial example here, is not too far from the previous point, and thus the gradient information from previous adversarial examples may still be useful.

3.3 Stochastic Variance Reduced Ensemble Attack

Based on the aforementioned observation on existing model ensemble attacks, we propose a stochastic variance reduced ensemble algorithm to take full advantage of the ensemble models. The stochastic variance reduced gradient (SVRG) method (Johnson & Zhang, 2013) is a promising approach for gradient descend for the classic continuous optimization problems. The basic idea of SVRG is to reduce the inherent variance of Stochastic Gradient Descent (SGD) using predictive variance reduction. In this work, we adopt the idea of SVRG to design a new ensemble adversarial attack so as to reduce the inherent gradient variance of multiple models.

We denote the traditional model ensemble algorithm as **Ens**, and our proposed stochastic variance reduced ensemble algorithm as **SVR-Ens**. SVR-Ens can be integrated with any existing gradient-based attack methods. The integration of SVR-Ens with I-FGSM, denoted by I-FGSM-SVR-Ens, is summarized in Algorithm 1.

The biggest difference between SVR-Ens and Ens is in the inner update loop of SVR-Ens, where SVR-Ens obtains a stochastic variance reduced gradient via M updates. Specifically, we first obtain the gradient of the models, $\boldsymbol{g}_{ens}$, by one pass over the models and maintain the value during M inner iterations. Then, we randomly pick a model from the ensemble models, obtain the stochastic variant reduced gradient $\boldsymbol{g}_{svrg}$, and update the inner adversarial example using $\boldsymbol{g}_{svrg}$. Finally, we update the

outer adversarial example using the last $\boldsymbol{g}_{svrg}$ of the inner update loop. Note that, for SVR-Ens, we keep the gradient calculation and adversarial perturbation update of the inner loop the same as that of the outer loop. In this way, SVR-Ens can easily combine with any existing attacks, just like the integration of SVR-Ens with I-FGSM. In summary, Ens directly uses the gradient of the ensemble models $\boldsymbol{g}_{ens}$ to update the adversarial example, while SVR-Ens uses the stochastic variance reduced gradient $\boldsymbol{g}_{svrg}$ to update the adversarial example.

Algorithm 1 The I-FGSM-SVR-Ens attack algorithm

Require: a benign example $\mathbf{x}$ and its label y, a set of K surrogate models and corresponding losses $\{J_1, \ldots, J_K\}$, an ensemble loss J chosen from $\{Eq.(1), Eq.(2), Eq.(3)\}$
Require: the perturbation bound ϵ, number of iterations T, internal update frequency of SVR-Ens M, internal step size β.
Ensure: An adversarial example $\boldsymbol{x}^{adv}$ that fulfills $\|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_\infty \leq \epsilon$

$\alpha = \epsilon/T$;
$\boldsymbol{x}_0^{adv} = \boldsymbol{x}$;
for $t = 0$ to $T - 1$ **do**
    # Calculate the gradient of the ensemble model
    Input $\boldsymbol{x}_t^{adv}$ to have the loss of the ensemble model $J(\boldsymbol{x}_t^{adv}, y)$;
    Obtain the gradient of the ensemble model by $\boldsymbol{g}_{ens} = \frac{1}{m}\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)$;
    # Stochastic variance reduction via M updates
    $\boldsymbol{x}_0 = \boldsymbol{x}_t^{adv}$;
    for $m = 0$ to $M - 1$ **do**
        Randomly pick a model index $k \in \{1, \ldots, K\}$
        Obtain the corresponding loss $J_k \in \{J_1, \ldots, J_K\}$
        Obtain $\boldsymbol{g}_{svrg}$ by $\boldsymbol{g}_{svrg} = \nabla_{\boldsymbol{x}} J_k(\boldsymbol{x}_m, y) - \nabla_{\boldsymbol{x}} J_k(\boldsymbol{x}_0, y) + \boldsymbol{g}_{ens}$
        # Update the inner adversarial example
        Update $\boldsymbol{x}_{m+1}$ by $\boldsymbol{x}_{m+1} = \text{Clip}_x^{\epsilon}\{\boldsymbol{x}_m + \beta \cdot \text{sign}(\boldsymbol{g}_{svrg})\}$
    # Update the outer adversarial example
    $\boldsymbol{x}_{t+1}^{adv} = \text{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{svrg})\}$
return $\boldsymbol{x}^{adv} = \boldsymbol{x}_T^{adv}$

4 Experiments

For experiments, we first introduce the experimental setup, then measure the attack success rate on normally trained models and defense models, and show that SVR-Ens outperform Ens significantly on both cases for black-box attacks. We continue to compare from the perspective of loss to show that SVR-Ens improves the average loss on black-box models by a large margin. In the end, we perform ablation studies to show the effectiveness of the key parameters in SVR-Ens.

4.1 Experimental Setup

Dataset. We conduct experiments on an ImageNet-compatible dataset [1] which comprises of 1,000 images and is widely used in recent FGSM-based attacks (Dong et al., 2019; Gao et al., 2020).

Networks. We consider four normally trained networks, *i.e.*, Inception-v3 (Inc-v3) (Szegedy et al., 2016), Inception-v4 (Inc-v4), Resnet-v2-152 (Res-152) (Szegedy et al., 2017), and Inception-Resnet-v2 (IncRes-v2) (He et al., 2016). For adversarially trained models, we consider Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$ (Tramèr et al., 2018).

Furthermore, we consider nine defense models which are shown to be robust against black-box attacks, including the top-3 defense methods in the NIPS competition: HGD (Liao et al., 2018), R&P (Xie et al., 2018), NIPS-r3 [2] and six recently proposed defense methods: Bit-R (Xu et al., 2018), JPEG (Guo et al., 2018), FD (Liu et al., 2019), ComDefend (Jia et al., 2019), RS (Jia et al., 2020) and NRP (Naseer et al., 2020).

[1] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/dataset

[2] https://github.com/anlthms/nips-2017/tree/master/mmd

Table 1: The attack success rates (%) of adversarial examples against the hold-out model. We study four normal models: Inc-v3, Inc-v4, IncRes-v2 and Res-101. The adversarial examples are crafted via an ensemble of the other three. We run the SVR-Ens attack for 5 times with different random seeds to reduce the randomness.

Base	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Average
I-FGSM	Ens	77.30	66.70	58.50	48.80	62.83
	SVR-Ens	**89.24**	**83.64**	**77.60**	**65.58**	**79.02**
MI-FGSM	Ens	90.30	86.60	82.20	77.40	84.13
	SVR-Ens	**96.84**	**95.30**	**92.80**	**89.40**	**91.59**
TIM	Ens	91.70	88.70	84.30	79.20	85.98
	SVR-Ens	**96.10**	**93.66**	**90.18**	**85.36**	**91.33**
TI-DIM	Ens	95.70	94.10	93.20	90.10	93.28
	SVR-Ens	**97.78**	**96.86**	**95.92**	**93.98**	**96.14**
SI-TI-DIM	Ens	97.60	97.60	97.20	95.90	97.08
	SVR-Ens	**98.80**	**98.88**	**97.90**	**97.82**	**98.35**

Baselines. We compare the proposed SVR-Ens with Ens based on the advanced gradient-based attacks, including I-FGSM (Goodfellow et al., 2015), MI-FGSM (Dong et al., 2018), TIM (Dong et al., 2019), TI-DIM (Dong et al., 2019), and SI-TI-DIM (Lin et al., 2020).

Hyper-parameters. To align with the previous works (Dong et al., 2018; Xie et al., 2019b; Dong et al., 2019), we set the maximum perturbation $\epsilon = 16/255$, the number of iterations is 10, and the step size is $\alpha = 1.6$. For MI-FGSM, we set the decay factor μ to 1.0. For TIM, we adopt the Gaussian kernel with size 7×7. For TI-DIM, the transformation probability p is set to 0.5. For SI-TI-DIM, we set the number of copies m to 5. For SVR-Ens, we set the internal update frequency M to four times the number of ensemble models and the internal step size β is set the same as α.

4.2 Attacking Normally Trained Models

We first compare the performance of our method on the normally trained models, including Inc-v3, Inc-v4, Res-152 and IncRes-v2. We keep one model as the hold-out black-box model and attack an ensemble of the other three models by various attacks with or without the SVR-Ens strategy. As shown in Table 1, SVR-Ens improves the attack success rate across all experiments over the normal ensemble strategy. For instance, under the base attack of I-FGSM, SVR-Ens increases the attack success rate on Res-101 from 48.80% to 65.58%. Under the SI-TI-DIM attack, which is integrated with multiple techniques, SVR-Ens can still effectively increase the attack success rate on Inc-v4 from 97.60% to 98.88% even though the previous value has already been very high. On average, SVR-Ens increases the attack success rate on I-FGSM, MI-FGSM, TIM, DIM, and SI-TI-DIM by 16.19%, 7.46%, 5.35%, 2.86% and 1.27%, respectively. The results demonstrate that SVR-Ens can effectively improve the transferability of adversarial examples on normally trained models.

4.3 Attacking Advanced Defense Models

To further demonstrate the efficacy of the proposed SVR-Ens in practice, we continue to evaluate our method on various defense models. Specifically, we attack the ensemble of the four normally trained models introduced in Section 4.2, and test the transferability of the crafted adversaries on several defense models.

Adversarially trained defense models are shown to be resistant to adversarial examples. We first test the transferability of the adversaries on three adversarially trained models, Inc-v3$_{\text{ens3}}$, Inc-v3$_{\text{ens4}}$ and IncRes-v2$_{\text{ens}}$. The results are shown in Table 2. We can observe that SVR-Ens improves the black-box attack success rate of the three adversarially trained models by a large margin over Ens for all the base attack methods that the two ensemble methods integrated with. Among various attack base methods, SVR-Ens exhibits the highest improvement on TIM, as TIM-SVR-Ens yields a 17.30% higher average attack success rate than TIM-Ens. In addition, for the SI-TI-DIM base attack, SVR-Ens outperforms Ens by 3.04% on average even though the average value of Ens is already very high at 94.34% on the three models. Besides, SVR-Ens performs well in the white-box setting, and can slightly improve the white-box attack performance in most cases.

Table 2: The attack success rates (%) of adversarial examples against seven models in the multi-model setting. We run the SVR-Ens attack for 5 times with different random seeds to reduce the randomness.

Base	Attack	White-Box				Black-Box			
		Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Average
I-FGSM	Ens	100.00	100.00	99.06	99.80	27.10	24.50	15.70	22.43
	SVR-Ens	99.80	99.60	99.38	99.58	**40.08**	**37.30**	**24.76**	**34.05**
MI-FGSM	Ens	99.90	99.90	99.70	99.50	50.50	49.30	32.30	44.03
	SVR-Ens	99.96	99.96	99.86	99.82	**64.54**	**59.02**	**39.08**	**54.21**
TIM	Ens	99.80	99.70	99.40	99.20	73.50	68.10	59.70	67.10
	SVR-Ens	99.84	99.90	99.80	99.70	**87.88**	**85.62**	**79.70**	**84.40**
TI-DIM	Ens	99.50	99.40	99.00	98.70	87.40	84.30	77.60	83.10
	SVR-Ens	99.86	99.80	99.68	99.34	**95.32**	**93.66**	**90.08**	**93.02**
SI-TI-DIM	Ens	99.70	99.40	99.30	99.40	95.60	95.10	92.40	94.34
	SVR-Ens	99.98	99.96	99.90	99.80	**98.56**	**97.78**	**95.80**	**97.38**

Table 3: The attack success rates (%) of adversarial examples against nine models with advanced defense mechanism. We run the SVR-Ens attack for 5 times with different random seeds to reduce the randomness.

Base	Attack	HGD	R&P	NIPS-r3	Bit-R	JPEG	FD	ComDefend	RS	NRP	Average
I-FGSM	Ens	27.00	15.20	18.90	26.00	41.80	37.10	56.00	25.20	17.30	29.39
	SVR-Ens	**45.48**	**25.02**	**34.10**	**30.96**	**62.06**	**50.42**	**66.98**	**26.98**	**21.60**	**40.40**
MI-FGSM	Ens	41.30	33.00	44.60	39.70	75.90	62.80	77.50	36.90	27.30	48.78
	SVR-Ens	**44.06**	**40.72**	**59.54**	**43.42**	**89.06**	**73.28**	**86.60**	**39.12**	**28.46**	**56.03**
TIM	Ens	72.50	60.50	67.20	49.30	82.60	74.80	85.10	47.80	37.60	64.16
	SVR-Ens	**87.10**	**80.16**	**83.84**	**62.26**	**91.96**	**83.96**	**92.22**	**62.46**	**52.24**	**77.36**
TI-DIM	Ens	87.40	81.20	85.70	63.00	91.70	84.30	91.90	57.90	49.80	76.99
	SVR-Ens	**94.86**	**91.92**	**93.22**	**72.88**	**96.48**	**90.76**	**95.98**	**73.60**	**65.38**	**86.12**
SI-TI-DIM	Ens	95.70	93.20	94.10	82.70	96.70	93.30	97.90	78.00	76.80	89.82
	SVR-Ens	**97.70**	**96.12**	**97.48**	**86.64**	**98.54**	**95.60**	**99.06**	**85.72**	**85.44**	**93.59**

In addition to the adversarially trained models, we also evaluate our methods on another nine models with advanced defense methods as noted in Section 4.1. In Table 3, we show the results of ensemble attacks against the nine defense models. The proposed method also improves the attack success rates across all experiments over the baseline attacks. We can observe that the SI-TI-DIM integrated with SVR-Ens can achieve an average attack success rate of 93.59% on these defense models in the black-box setting, which raises a new security issue for the robust deep neural networks.

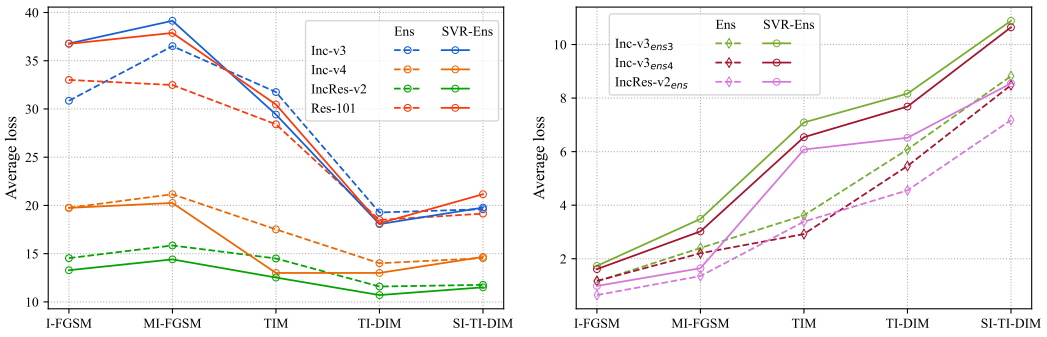
4.4 Comparison on Loss

The above experiments have demonstrated that SVR-Ens has a significant impact in improving the attack success rate of adversarial attacks. To provide more intuitive evidence to show that SVR-Ens can effectively boost the transferability of adversarial examples, we average the loss over the adversarial images generated in Section 4.3 on four white-box models and three black-box models and depict the improvement curve for the average loss. Loss can indirectly reflect the adversarial effect: a higher loss indicates a stronger adversarial effect, and a higher loss on the black-box model indicates a stronger transferability.

The results are shown in Figure 1. Compared with Ens, SVR-Ens improves the average loss on black-box models by a large margin. Specially, the loss of the adversarial examples generated by TIM-SVR-Ens is more than twice of the TIM-Ens in the black-box setting. In terms of the white-box setting, SVR-Ens and Ens are comparable, showing that the improvement of SVR-Ens in transferability is not based on the premise of sacrificing the performance of white-box attack.

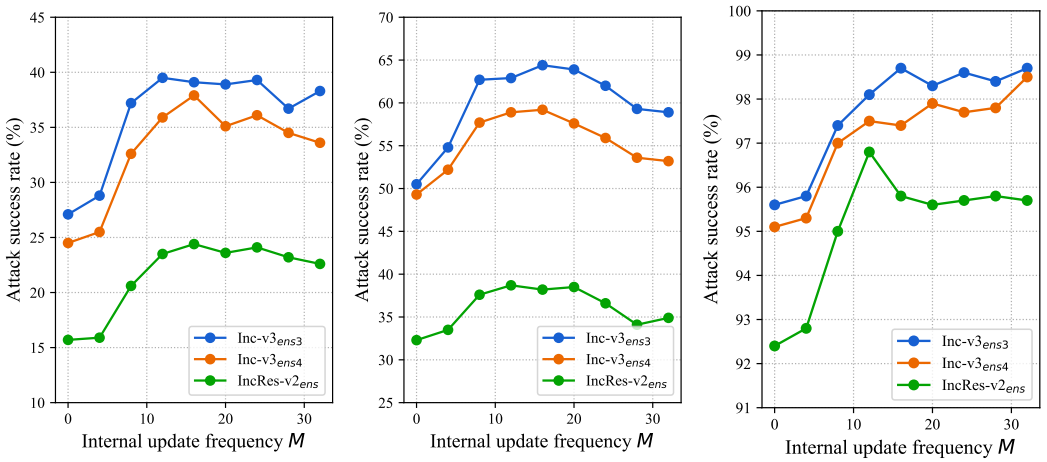
4.5 Ablation Study

In this subsection, we conduct a series of ablation experiments to study the impact of the parameters in SVR-Ens. We attack the ensemble of Inc-v3, Inc-v4, Res-152 and IncRes-v2 and test the transferability of the adversaries on Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$, as the setting in 4.2.



(a) white-box (b) black-box

Figure 1: The average loss on seven models against five attacks integrated with Ens and SVR-Ens, respectively.



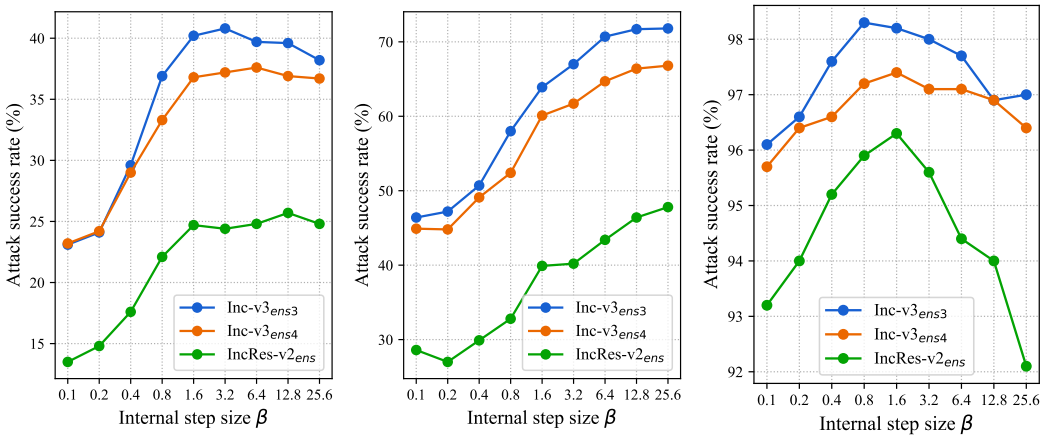
(a) I-FGSM (b) MI-FGSM (c) SI-TI-DIM

Figure 2: The attack success rate (%) of the I-FGSM, MI-FGSM and SI-TI-DIM base attacks after integrated with SVR-Ens. It degenerates to the integration with Ens when $M = 0$.

On the internal update frequency M. We first analyze the effectiveness of the internal update frequency M on the attack success rate of SVR-Ens. We integrate I-FGSM, MI-FGSM and SI-MI-DIM attacks with SVR-Ens respectively and range the internal update frequency M from 0 to 32 with granularity 4. Note that if $M = 0$, SVR-Ens trivially degenerates to the normal ensemble method of Ens. Since the attack success rate in the white-box setting is close to 100%, we only show results for black-box attacks. A first glance shows that our SVR-Ens has achieved an impressive improvement than Ens ($M = 0$). As the number of iterations increases, the attack success rate increases and reaches the peak at about $M = 16$. We also observe from the convex curve that either too high or too low number of iterations may cause the adversarial examples overfit to the current model and reduce the attack transferability.

On the internal step size β. The internal step size β plays a key role in improving the attack success rate, as it determines the extent of the data point update of each inner loop. Similarly, we perform I-FGSM, MI-FGSM and SI-MI-DIM attacks integrated with SVR-Ens with β ranging from 0.1 doubled to 25.6. As shown in Figure 3, the performance of SVR-Ens varies with the step size, and the best step length varies for different methods. In the above experiments for comparison, we did not deliberately set different best parameters for each method. So for practical applications, we can adopt the best step size for a specific attack to obtain higher performance.

On the number of iterations T. For the same number of iterations, SVR-Ens has more gradient calculations due to its internal update process. To show that the improvement of SVR-Ens is not



(a) I-FGSM (b) MI-FGSM (c) SI-TI-DIM

Figure 3: The attack success rate (%) of I-FGSM, MI-FGSM and SI-TI-DIM after integrated with SVR-Ens on different internal step size β of SVR-Ens.

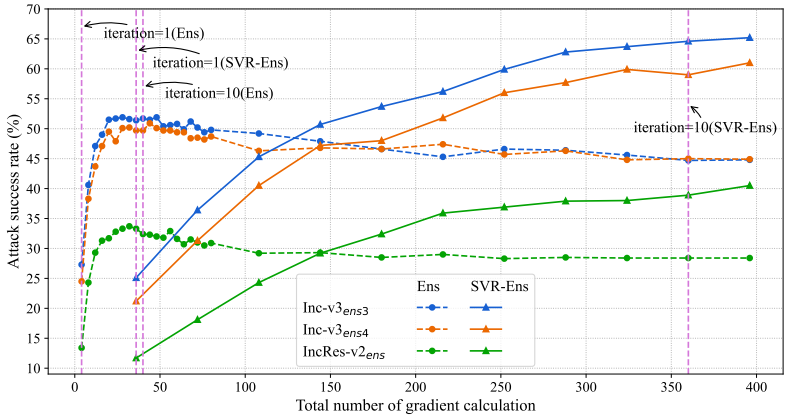


Figure 4: The attack success rate (%) of the MI-FGSM base attack after integrated with Ens or SVR-Ens.

simply caused by increasing the number of gradient calculations, we perform additional analysis on the number of iterations. Taking the internal update frequency $M = 16$ and the number of ensemble models $K = 4$ as an example, each iteration requires 4 queries of the model in Ens, while integrated SVR-Ens, the additional inner loop requires $16 \times 2 = 32$ additional queries. The overall number of queries for SVR-Ens is 9 times that of Ens. Then, what if we increase the number of iterations for other methods? It can be observed that the attack success rate of Ens against black-box model gradually decays with the increment on the number of iterations, and there is a big gap even when their iterations reach 360. This experiment shows that simply increasing the number of iterations on Ens could not gain the high attack performance of SVR-Ens.

5 Conclusion

In this work, we study the model ensemble attacks and propose a novel stochastic variance reduced ensemble attack (SVR-Ens) method. Different from the existing model ensemble attacks, which simply fuse the outputs of multiple models, the proposed SVR-Ens takes the variance of different models into account and leverages the stochastic variance reduce method to address this issue. In this way, the SVR-Ens can generate adversarial examples with larger loss and better transferability. Extensive experiments demonstrate that our method surpasses the vanilla model ensemble attack in both the loss and attack transferability.

REPRODUCIBILITY STATEMENT

In Section 4.1, we provide a complete description of the setup of our experiments. All the dataset and models are open source, the dataset comes from the NIPS 2017 Adversarial Attacks and Defenses Competition, which was widely used in the liteature, and all the models have the corresponding references or download links. And the implementation details for the attack methods are introduced in the Hyper-parameters part in Section 4.1. Furthermore, we repeat the experiments in Table 1, Table 2, Table 3 for five times to reduce the variance. We also provide the code for integrating SVR-Ens on I-FGSM as an example. We promise to provide the complete source code for the final version.

REFERENCES

Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80, pp. 274–283, 2018.

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

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4312–4321, 2019.

Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision, ECCV*, pp. 307–322, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations,ICLR*, 2015.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR*, 2018.

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

Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *8th International Conference on Learning Representations, ICLR*, 2020.

Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. ComDefend: An efficient image compression model to defend adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 6084–6092, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, pp. 315–323, 2013.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5nd International Conference on Learning Representations,ICLR (Workshop)*, 2017a.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR*, 2017b.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1778–1787, 2018.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *8th International Conference on Learning Representations, ICLR*, 2020.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR*, 2017.

Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented JPEG compression against adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 860–868, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 86–94, 2017.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 259–268, 2020.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387, 2016.

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

Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *7th International Conference on Learning Representations, ICLR*, 2019.

Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E. Hopcroft. Robust local features for improving the generalization of adversarial training. In *8th International Conference on Learning Representations, ICLR*, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-Resnet and the impact of Residual Connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278–4284, 2017.

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.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems, NIPS*, 33, 2020.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1924–1933, 2021.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR*, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 501–509, 2019a.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2730–2739, 2019b.

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

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John E. Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *CoRR*, abs/1906.00555, 2019.