Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability

Yifeng Xiong[1]*, Jiadong Lin[1], Min Zhang[1], John E. Hopcroft[2], Kun He[1]†
[1]Department of Computer Science, Huazhong University of Science and Technology, Wuhan, China
[2]Department of Computer Science, Cornell University, Ithaca, NY, USA
{xiongyf,jdlin,m_zhang}@hust.edu.cn, jeh@cs.cornell.edu, brooklet60@hust.edu.cn

Abstract

The black-box adversarial attack has attracted impressive attention for its practical use in the field of deep learning security. Meanwhile, it is very challenging as there is no access to the network architecture or internal 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 the attack capability to other models, the ensemble-based adversarial attack methods are efficient and widely used for black-box attacks. However, ways of ensemble attack are rather less investigated, and existing ensemble attacks simply fuse the outputs of all the models evenly. In this work, we treat the iterative ensemble attack as a stochastic gradient descent optimization process, in which the variance of the gradients on different models may lead to poor local optima. To this end, we propose a novel attack method called the stochastic variance reduced ensemble (SVRE) attack, which could reduce the gradient variance of the ensemble models and take full advantage of the ensemble attack. Empirical results on the standard ImageNet dataset demonstrate that the proposed method could boost the adversarial transferability and outperforms existing ensemble attacks significantly. Code is available at https://github.com/JHL-HUST/SVRE.

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 [7, 23, 28]. Moreover, adversarial examples are known to be transferable that the examples crafted for one model can also mislead other black-box models [17, 20, 22]. Generating adversarial examples, *i.e.*, adversarial attack, has drawn enormous attention since it can help evaluate the robustness of different models [2, 29] and improve their robustness by adversarial training [7, 19].

Various adversarial attack methods have been proposed, including the optimization-based methods such as box-constrained L-BFGS [28] and Carlini & Wagner's method [2], the gradient-based methods such as Fast Gradient Sign Method [7] and its iterative variants [13, 19], *etc*. In general, these adversarial attack methods can achieve high attack success rates in the white-box setting [2], where the attacker can access the complete information of the target model, including the model architecture and gradient information. However, these methods often exhibit low attack success rates in the black-box setting [3], where the attacker can not access the information of the target model. In this case, the attacker either utilizes the transferability of adversarial examples to fool the black-box model, or attacks directly based on a small amount of queries on the output of the black-box model.

In recent years, a number of methods have been proposed to enhance the transferability of adversarial examples so as to improve the attack success rates in the black-box setting, including the gradient optimization attacks [3, 16, 31], input transformation attacks [4, 16, 35], and model ensemble attacks [3, 17]. Among these methods, the model ensemble attacks are efficient and have been broadly adopted in boosting the black-box attack performance [5, 16, 35]. However, as compared to the other two categories that have been explored in depth, the category of model ensemble attack is rather less investigated.

In this work, we observe that the existing model ensemble attack methods simply fuse the outputs of all models directly but ignore the variance of the gradients on different models, which may limit the potential capability of the model ensemble attacks. Due to the inherent difference of the model architectures, the optimization paths of the models may differ widely, indicating that there exists considerable difference on the variance of the gradient directions among the possible models. Such variance may cause the

*The first two authors contribute equally.
†Corresponding author.

optimization direction of the ensemble attack to be less accurate. As a result, the attack capability of the transferred adversarial examples is rather limited.

To address this issue, we propose a novel method called the stochastic variance reduced ensemble (SVRE) attack to enhance the adversarial transferability of ensemble attacks. Our method is inspired by the stochastic variance reduced gradient (SVRG) method [12] designed for stochastic optimization, which has an outer loop that maintains an average gradient on a batch of data and an inner loop that randomly draws an instance from the batch and calculates an unbiased estimate of gradient based on the variance reduction. In our method, we regard the ensemble models as the batch of data for the outer loop and randomly draw a model at each iteration of the inner loop. Taking the benign image as the initial adversarial example, the outer loop calculates the average gradient on the batch of models, and copies the current example to the inner loop, then the inner loop conducts multiple iterations of inner adversarial example updates. At each inner iteration, SVRE calculates the current gradient on a randomly picked model, tuned by the gradient bias of the outer adversarial example on this randomly picked model and on the ensemble model. At the end of the inner loop, the outer adversarial example is updated using the tuned gradient of the newest inner adversarial example.

In this way, SVRE can obtain a more accurate gradient update at the outer loop to escape from poor local optima such that the crafted adversarial example would not "overfit" the ensemble model. Hence, the crafted adversarial example is expected to have higher transferability to other unknown models. To our knowledge, this is the first work to investigate the limitation of existing ensemble attack through the lens of gradient variance on multiple models. Extensive experiments on the ImageNet dataset demonstrate that SVRE consistently outperforms the vanilla ensemble model attack in 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 for an adversarial example $x^{adv} \in \mathcal{B}_\epsilon(x)$ that maximizes 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 attack methods can be categorized into three groups, namely gradient optimization attacks [3, 7, 13, 16, 31], input transformation attacks [3, 16, 32, 35], and model ensemble attacks [3, 17].

Gradient optimization attacks. The most typical adversarial attack based on gradient is the Fast Gradient Sign Method (FGSM) [7], 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*. [13] propose the Basic Iterative Method (BIM) to run multiple iterations of FGSM with a small perturbation. Madry *et al*. [19] propose a noisy version of BIM, named the Projected Gradient Descent (PGD). Although PGD exhibits good attack performance in the white-box setting [1], it overfits the target model easily and yields weak transferability in the black-box setting. In order to improve the transferability of adversarial attacks, Dong *et al*. [3] propose to boost the adversarial attack with momentum. More recently, Lin *et al*. [16] introduce Nesterov accelerated gradient method into the gradient-based attack to look ahead effectively to avoid overfitting. Wang *et al*. [31] reduce the variance of the gradient at each iteration to stabilize the update direction.

Input transformation attacks. Another line of attacks focuses on adopting various input transformations to further improve the transferability of adversarial examples. Xie *et al*. [35] propose the Diverse Input Method (DIM) [35], which utilizes random resizing and padding to create diverse input patterns to generate adversarial examples. Dong *et al*. [4] propose the Translation-Invariant Method (TIM), which optimizes the perturbation over a set of translated images. Lin *et al*. [16] discover the scale-invariant property of deep learning models and propose the Scale-Invariant Method (SIM), which optimizes the adversarial perturbations over the scale copies of the input images. Wang *et al*. [32] propose the Admix, that calculates the gradient on the input image admixed with a small portion of each add-in image to craft more transferable adversaries.

Model ensemble attacks. Liu *et al*. [17] find that attacking multiple models simultaneously can also improve the attack transferability. They fuse the predictions of multiple models to get the loss of ensemble predictions and adopt existing adversarial attacks (*e.g*. FGSM and PGD) to generate adversarial examples using the loss. Dong *et al*. [3] propose two variants of the model ensemble attack, namely fusing the logits and fusing the losses, respectively. Compared with various explorations on gradient optimization or input transformation, the model ensemble attacks are far less investigated, and existing methods only simply fuse the output predictions, logits, or losses.

2.2. Adversarial Defenses

As the counterpart of adversarial attacks, various defense methods have also been proposed, including adversarial training based defenses [6, 19, 24, 25, 28, 30, 34, 37] and input transformation based defenses [8, 10, 11, 15, 18, 21, 33, 36].

Adversarial training based defenses. Adversarial training is considered one of the most efficacious defense

approaches which augments the training data by generating adversarial examples during the training process. Tramèr *et al*. [30] propose ensemble adversarial training, which augments the training data with perturbations transferred from other models. Madry *et al*. [19] propose PGD-Adversarial Training (PGD-AT), which augments the training data with adversarial examples crafted by PGD attack. Xie *et al*. [34] develop new network architectures that increase adversarial robustness by performing feature denoising.Adversarial training, while promising, is computationally expensive and hard to scale to large-scale datasets [14].

Input transformation based defenses. This line of defenses aims to diminish the adversarial perturbations from the input data. Guo *et al*. [8] and Xie *et al*. [33] conduct transformations on images to remove the adversarial perturbations. Liao *et al*. [15] use high-level representation guided denoiser (HGD) to purify the adversarial images. Xu *et al*. [36] propose two feature squeezing methods, *i.e*. bit reduction (Bit-R) and spatial smoothing to detect adversarial examples. Liu *et al*. [18] propose the feature distillation (FD), which adopts a JPEG-based defensive compression framework to diminish the adversarial perturbations. Jia *et al*. [11] utilizes an end-to-end image compression model named ComDefend to defend against adversarial examples. Jia *et al*. [10] leverage the randomized smoothing (RS) to train a certifiably robust ImageNet classifier. Naseer *et al*. [21] 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 adversarial transferability through the lens of reducing the gradient variance of the ensemble models used for crafting the adversarial example. Since our method is based on the model ensemble attack, we first introduce the existing ensemble attack methods, then present our motivation and elaborate the proposed SVRE in detail.

3.1. Ensemble Attack Methods

The ensemble attack [3, 17] is an effective strategy to enhance the adversarial transferability. Its basic idea is to generate the adversarial examples using multiple models.

Ensemble on predictions. Liu *et al*. [17] first propose to achieve an ensemble attack by averaging the predictions (predicted probability) of the models. For an ensemble of K models, the loss function of the ensemble model is:

$$J(\boldsymbol{x}, y) = -\mathbf{1}_y \cdot \log(\textstyle\sum_{k=1}^{K} w_k \boldsymbol{p}_k(\boldsymbol{x})), \tag{1}$$

where $\mathbf{1}_y$ is the one-hot encoding of the ground-truth label y of $\boldsymbol{x}$, $\boldsymbol{p}_k$ is the prediction of the k-th model, and $w_k \geq 0$ is the ensemble weight constrained by $\sum_{k=1}^{K} w_k = 1$.

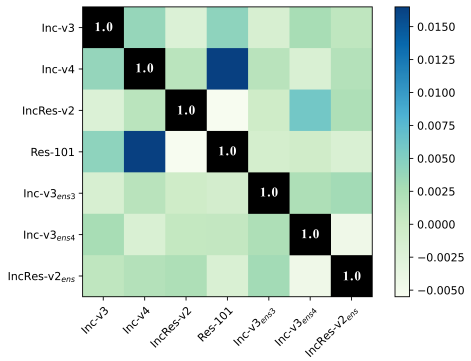


Figure 1. The cosine similarity between the gradients (processed by sign function) of a sampled image on different models.

Ensemble on logits. Dong *et al*. [3] propose to fuse the logits (output before softmax) of models. For the ensemble of K models, the loss function ensembled on logits is:

$$J(\boldsymbol{x}, y) = -\mathbf{1}_y \cdot \log(\mathrm{softmax}(\textstyle\sum_{k=1}^{K} w_k \boldsymbol{l}_k(\boldsymbol{x}))), \tag{2}$$

where $\boldsymbol{l}_k$ is the logits of the k-th model.

Ensemble on losses. Dong *et al*. [3] also introduce an alternative ensemble attack by averaging the loss of K models as follows:

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

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

For the weight parameters, the three methods simply choose the average weight in experiments, *i.e*. $w_k = 1/K$.

3.2. Rethinking the Ensemble Attack

The ensemble attack method has been broadly adopted in enhancing the performance of black-box attacks [3, 5, 16, 17, 31, 35]. However, to our knowledge, researchers only utilize the existing ensemble attack strategy as a plug-and-play module to enhance their own attack methods, but did not delve into the ensemble attack method itself.

Intuitively, existing ensemble attack methods [3, 17] are helpful in improving the adversarial transferability because attacking an ensemble model can help to find better local maxima and makes it easier to generalize to other black-box models. However, merely averaging the outputs (logits, predictions or loss) of the models to build an ensemble model for the adversarial attack may limit the attack performance, as the individual optimization path of different model may vary diversely, but the variance is not considered, leading to an overfit to the ensemble model.

As demonstrated in Figure 1, the cosine similarity between the update direction of a sampled image on different models is extremely low, indicating there exists a considerable gap in the optimization direction among these models (See model details in Section 4.1). We argue that simply

fusing the predictions/logits/losses of the models but ignoring the variance of the gradients on different models would lead to a suboptimal result, and limit the performance of ensemble attacks.

3.3. Stochastic Variance Reduced Ensemble Attack

In previous works, Lin *et al*. and Wang *et al*. [16, 31] analogize the process of the adversarial example generation to the process of neural network training, of which the white-box model is analogized to the training data and the black-box model is analogized to the test data. Hence, the iterative optimization on crafting the adversarial example using the input image can be regarded as the parameter update of neural networks, and the transferability of the adversarial example is analogized to the generalization of models.

In this work, we treat the iterative ensemble attack as a stochastic gradient descent optimization process, in which at each iteration, the attacker always chooses the batch of the ensemble models for update. During the course of the adversarial example generation, the gradient variance on different models may lead to poor local optima. Hence, we aim to reduce the gradient variance so as to stabilize the gradient update direction, making the induced gradient be generalized better to other possible models.

Inspired by the stochastic variance reduced gradient (SVRG) method [12] designed for stochastic optimization, we propose a stochastic variance reduced ensemble attack method to address the gradient variance of the models so as to take full advantage of the ensemble attack. The basic idea of SVRG is to reduce the inherent variance of Stochastic Gradient Descent (SGD) using predictive variance reduction, while we aim to reduce the inherent gradient variance of multiple models. The integration of SVRE with MI-FGSM [3], SVRE-MI-FGSM, is summarized in Algorithm 1.

Denote the traditional model ensemble attack method as Ens. The main difference of our method to Ens is that SVRE has an inner update loop, where SVRE obtains a variance reduced stochastic gradient by M updates. Specifically, we first obtain the gradient of the multiple models, $\boldsymbol{g}^{ens}$, by one pass over the models and maintain this value during M inner iterations. Then, we randomly pick a model from the ensemble models, obtain the stochastic inner gradient after variance reduction $\tilde{\boldsymbol{g}}_m$, and update the inner adversarial example using the accumulate gradients of $\tilde{\boldsymbol{g}}_m$. In the end, we update the outer gradient using the accumulate gradient of the last inner loop. As $\tilde{\boldsymbol{g}}_m$ is the unbiased estimate of the gradient of $\boldsymbol{g}_m^{ens}$, $(\nabla_{\boldsymbol{x}} J_k(\boldsymbol{x}_t^{adv}, y) - \boldsymbol{g}_t^{ens})$ helps to reduce the gradient on different models.

In a nutshell, the existing Ens method directly uses the average gradient of the ensemble models $\boldsymbol{g}^{ens}$ to update the adversarial example, while SVRE uses the stochastic variance reduced gradient $\tilde{\boldsymbol{g}}$ to update the adversarial example.

Algorithm 1 The SVRE-MI-FGSM attack algorithm

Input: A benign example $\mathbf{x}$ and its label y, a set of K surrogate models and the corresponding losses $\{J_1, \ldots, J_K\}$, an ensemble loss J chosen from $\{Eq.(1), Eq.(2), Eq.(3)\}$

Input: The perturbation bound ϵ, number of iterations T, internal update frequency M, internal step size β, decay factor μ_1, internal decay factor μ_2

Output: An adversarial example $\boldsymbol{x}^{adv}$ that fulfills $\|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_\infty \leq \epsilon$

$\alpha = \epsilon/T$; $\boldsymbol{G}_0 = 0$;
Initialize $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$;
for $t = 0$ to $T - 1$ **do**
  # Calculate the gradient of the ensemble model
  Get the loss of the ensemble model $J(\boldsymbol{x}_t^{adv}, y)$;
  Calculate the gradient of the ensemble model $\boldsymbol{g}_t^{ens}$:
    $\boldsymbol{g}_t^{ens} = \frac{1}{K} \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)$;
  # Stochastic variance reduction via M updates
  Initialize $\tilde{\boldsymbol{x}}_0 = \boldsymbol{x}_t^{adv}$;$\tilde{\boldsymbol{G}}_0 = 0$
  for $m = 0$ to $M - 1$ **do**
    Randomly pick a model index $k \in \{1, \ldots, K\}$
    Get the corresponding loss $J_k \in \{J_1, \ldots, J_K\}$
    $\tilde{\boldsymbol{g}}_m = \nabla_{\boldsymbol{x}} J_k(\tilde{\boldsymbol{x}}_m, y) - (\nabla_{\boldsymbol{x}} J_k(\boldsymbol{x}_t^{adv}, y) - \boldsymbol{g}_t^{ens})$
    # Update the inner gradient by momentum
    $\tilde{\boldsymbol{G}}_{m+1} = \mu_2 \cdot \tilde{\boldsymbol{G}}_m + \tilde{\boldsymbol{g}}_m$
    # Update the inner adversarial example
    Update $\tilde{\boldsymbol{x}}_{m+1} = \mathrm{Clip}_x^{\epsilon}\{\tilde{\boldsymbol{x}}_m + \beta \cdot \mathrm{sign}(\tilde{\boldsymbol{G}}_{m+1})\}$
  end for
  # Update the outer gradient by momentum
  $\boldsymbol{G}_{t+1} = \mu_1 \cdot \boldsymbol{G}_t + \tilde{\boldsymbol{G}}_M$
  # Update the outer adversarial example
  $\boldsymbol{x}_{t+1}^{adv} = \mathrm{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \mathrm{sign}(\boldsymbol{G}_{t+1})\}$
end for
return $\boldsymbol{x}^{adv} = \boldsymbol{x}_T^{adv}$

Theoretically, SVRE can be easily integrated with other iterative gradient-based attack methods. *E.g*. I-FGSM [7], MI [3], DI [4], TI [4], SI [16] can be integrated with SVRE using the same technique in inner loop and outer loop. But in SVRE-I-FGSM, we accumulate gradients in inner loop to have a better transferbility.

Compared with existing optimization-based methods of enhancing the attack transferability, our method is from a different perspective. Existing works mainly focus on the optimization along the iterative process. For instance, MI-FGSM [3] and NI-FGSM [16] aim to accelerate the convergence, while VT [31] aims to tune the current gradient using the variance of the gradient at the previous iteration for a single model. In contrast, our method seeks to reduce the variance of the gradient caused by various models in the ensemble attack.

4. Experiments

This section first introduces the experimental setup, then reports the attack success rate on normally trained models and defense models, showing that SVRE outperforms Ens significantly for black-box attacks. We further demonstrate that SVRE increases the average loss on black-box models by a large margin. In the end, we perform ablation studies to manifest the effectiveness of the key parameters in SVRE.

4.1. Experimental Setup

Dataset. We conduct experiments on an ImageNet-compatible dataset[1] which is comprised of 1,000 images and is widely used in recent FGSM-based attacks [4, 5].

Networks. We consider four normally trained networks, *i.e.*, Inception-v3 (Inc-v3) [27], Inception-v4 (Inc-v4), Resnet-v2-152 (Res-152) [26], and Inception-Resnet-v2 (IncRes-v2) [9]. For adversarially trained models, we consider Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$ [30]. Besides, 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 [15], R&P [33], NIPS-r3 [2] and six recently proposed defense methods: Bit-R [36], JPEG [8], FD [18], ComDefend [11], RS [10] and NRP [21].

Baselines. We compare the proposed SVRE with Ens based on the advanced gradient-based attacks, including I-FGSM [7], MI-FGSM [3], TIM [4], TI-DIM [4], and SI-TI-DIM [16]. For Ens, we adopt the ensemble method that fuses the logits of difference models [3], which is confirmed better than the ensemble on predictions or losses. In addition, we run the SVRE attack for 5 times with different random seeds and average the results to reduce the impact of randomness.

Hyper-parameters. To align with the previous works [3, 4, 16, 35], we set the maximum perturbation $\epsilon = 16/255$. For I-FGSM, the number of iterations is 10, and the step size is $\alpha = 1.6$. For MI-FGSM, we set the decay factor μ_1 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 SVRE, 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, the internal decay factor μ_2 is set to 1.0.

4.2. Attack 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. Specifically, we keep one model as the hold-out black-box model and generate adversarial examples on an ensemble of the other three models by Ens and SVRE integrated with various base methods.

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. For each model, the adversarial examples are crafted on an ensemble of the other three.

Base	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Average
I-FGSM	Ens	77.30	66.70	58.50	48.80	62.83
	SVRE	**89.24**	**83.64**	**77.60**	**65.58**	**79.02**
MI-FGSM	Ens	90.30	86.60	82.20	77.40	84.13
	SVRE	**96.84**	**95.30**	**92.80**	**89.40**	**93.59**
TIM	Ens	91.70	88.70	84.30	79.20	85.98
	SVRE	**96.10**	**93.66**	**90.18**	**85.36**	**91.33**
TI-DIM	Ens	95.70	94.10	93.20	90.10	93.28
	SVRE	**97.78**	**96.86**	**95.92**	**93.98**	**96.14**
SI-TI-DIM	Ens	97.60	97.60	97.20	95.90	97.08
	SVRE	**98.80**	**98.88**	**97.90**	**97.82**	**98.35**

Table 1 shows the attack performance on the hold out model. SVRE outperforms Ens across all the test models. The average improvement of SVRE over Ens on the base attack of I-FGSM is significant at 16.19%. Even on the advanced attack methods, MI-FGSM, TIM, DIM and SI-TI-DIM, the average improvements of SVRE over Ens are still considerable, which are 9.46%, 5.35%, 2.86% and 1.27%, respectively. The results demonstrate that SVRE can effectively improve the transferability of adversarial examples on normally trained models.

4.3. Attack Advanced Defense Models

To further validate the efficacy of our method in practice, we continue to evaluate SVRE on models with various advanced defenses. Specifically, we craft the adversarial examples on the ensemble of four normally trained models, *i.e.*, Inc-v3, Inc-v4, Res-15 and IncRes-v2, and test the transferability of the crafted adversaries on defense models.

We first evaluate the transferability of the adversaries on three adversarially trained models, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$. The results are shown in Table 2. We see that SVRE outperforms Ens on the black-box attack of each adversarially trained models by a large margin. Among the base methods that the ensemble attacks integrate with, SVRE exhibits the highest improvement on TIM, as SVRE-TIM yields a 17.30% higher average attack success rate than Ens-TIM. Besides, SVRE also performs well in the white-box setting, and can slightly improve the white-box attack performance in most cases.

In addition to the adversarially trained models, we also evaluate the crafted examples on nine models with advanced defense mechanisms. The results are shown in Table 3. SVRE outperforms Ens by a clear margin across all the comparisons. The strongest version of our method, SVRE

[1] https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset

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

Table 2. The black-box attack success rates (%) against three adversarially trained models. The adversarial examples are generated on the ensemble models, *i.e.*, Inc-v3, Inc-v4, IncRes-v2 and Res-101.

Base	Attack	White-box setting				Black-box setting			
		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	99.60	99.80	27.10	24.50	15.70	22.43
	SVRE	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
	SVRE	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
	SVRE	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
	SVRE	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	94.37
	SVRE	99.98	99.96	99.90	99.80	**98.56**	**97.78**	**95.80**	**97.38**

Table 3. The black-box attack success rates (%) against nine models with advanced defense mechanism.

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
	SVRE	**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
	SVRE	**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
	SVRE	**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
	SVRE	**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
	SVRE	**97.70**	**96.12**	**97.48**	**86.64**	**98.54**	**95.60**	**99.06**	**85.72**	**85.44**	**93.59**

integrated with SI-TI-DIM, 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 robustness of deep learning models.

4.4. Comparison on Loss

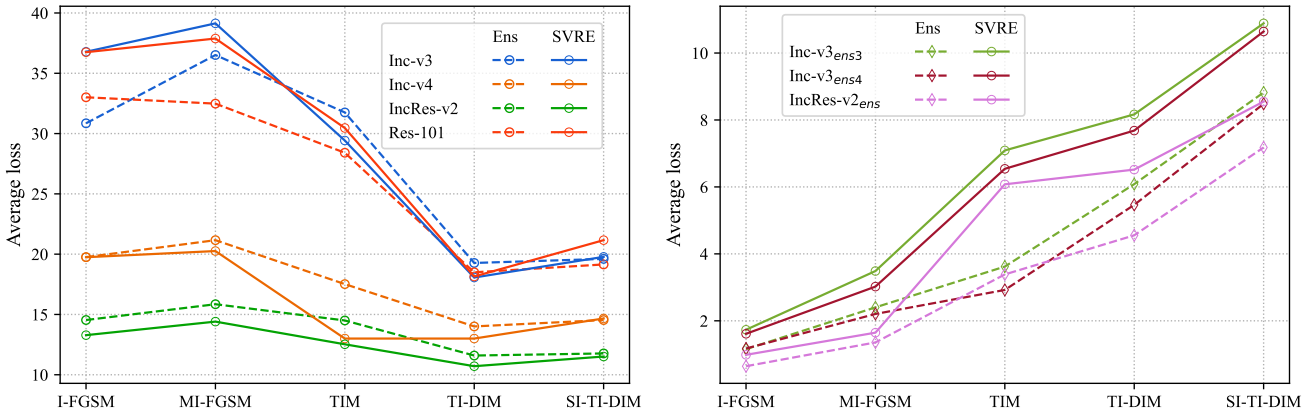
The above experiments have demonstrated that SVRE significantly improves the attack success rate of adversarial attacks. To provide intuitive evidence that SVRE can effectively boost the transferability of adversarial examples, we average the loss over the adversarial images generated in Section 4.3 on four white-box models and three black-box models respectively, and depict the improvement curve for the average loss in Figure 2. The loss can indirectly reflect the adversarial efficacy. A higher loss indicates a stronger adversarial intensity, and a higher loss on the black-box model indicates a stronger transferability.

We can see in Figure 2 (b) that SVRE increases the average loss over Ens on black-box models remarkably. In terms of the white-box setting in Figure 2 (a), SVRE and Ens are comparative, indicating that the improvement of SVRE in transferability is not based on the premise of sacrificing the performance of white-box attacks.

4.5. Ablation Study on Hyper-parameters

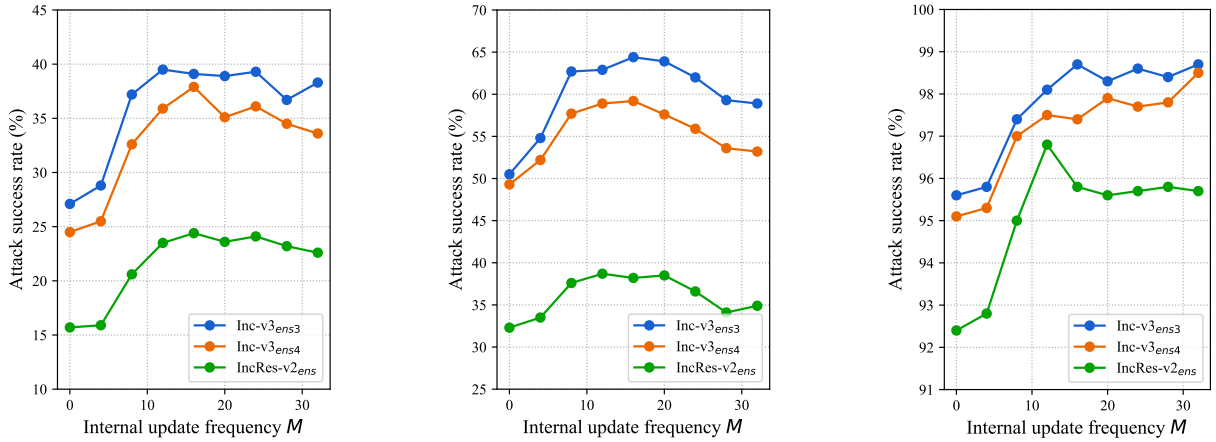
In this subsection, we conduct a series of ablation experiments to study the impact of the parameters in SVRE. Here we attack the ensemble of Inc-v3, Inc-v4, Res-152 and IncRes-v2 and test the transferability of the adversaries on the adversarially trained models Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and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

On the internal update frequency M. We first analyze the effectiveness of the internal update frequency M on the attack success rate of SVRE. We integrate I-FGSM, MI-FGSM and SI-MI-DIM attacks with SVRE, respectively, and range the internal update frequency M from 0 to 32 with a granularity of 4. Note that if $M = 0$, SVRE 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 the results for black-box attacks, as



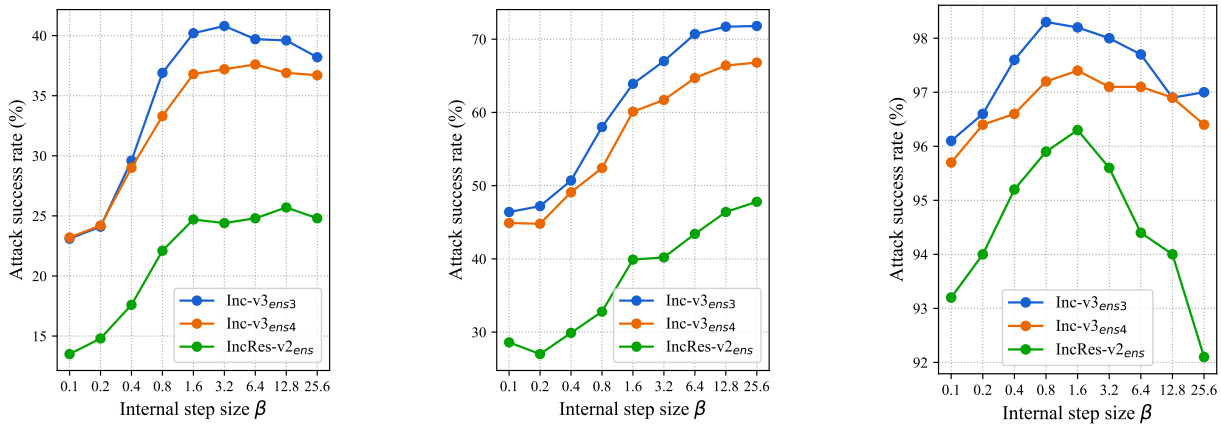
(a) White-box setting (b) Black-box setting

Figure 2. The average loss on seven models against Ens and SVRE integrated with five attacks, respectively.



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

Figure 3. The attack success rate (%) of SVRE integrated with I-FGSM, MI-FGSM and SI-TI-DIM. It degenerates to the integration with Ens when $M = 0$.



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

Figure 4. The attack success rate (%) of I-FGSM, MI-FGSM and SI-TI-DIM after integrated with SVRE on different internal step size β.

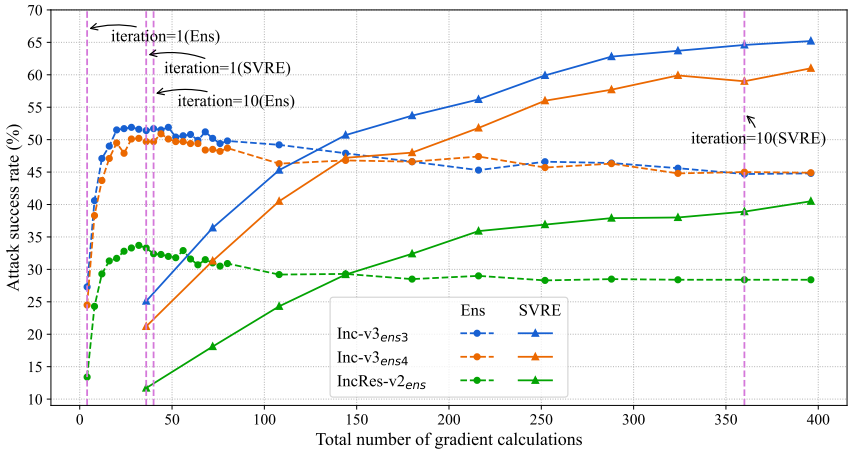


Figure 5. The attack success rate (%) of SVRE-MI-FGSM and Ens-MI-FGSM for different total number of gradient calculations.

shown in Figure 3. A first glance shows that our SVRE has achieved an impressive improvement over 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 many iterations or too few iterations may cause the adversarial examples to overfit the current model and harm the attack transferability.

On the internal step size β. The internal step size β plays a crucial role in improving the attack success rate, as it determines the extent of the data point update in each inner loop. Similarly, we perform SVRE integrated with I-FGSM, MI-FGSM and SI-MI-DIM respectively, fix $\alpha = 1.6$ and let β ranges from 0.1 doubled to 25.6. As shown in Figure 4, the performance of SVRE varies with the step size, and the best step size varies for different methods. For a fair comparison, we did not deliberately set different best parameters for each method but choose $\beta = 1.6$. For practical applications, the best step size can be adopted for a specific attack to obtain a higher performance.

On the number of iterations T. For the same number of iterations, SVRE has more gradient calculations due to its inner loop. To show that the gains of SVRE is not simply from increasing the number of gradient calculations, we perform additional analysis on the total 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 for SVRE, the inner loop requires $16 \times 2 = 32$ additional queries. The overall number of queries for SVRE is 9 times that of Ens. Then, what if we increase the number of iterations for other methods? One can observe from Figure 5 that the attack success rate of Ens-MI-FGSM against black-box models gradually decays with the increment on the total number of gradient calculations, and there is a big gap even when the total number reaches 360. This experiment demonstrates that simply increasing the number of iterations on Ens could not gain the high attack performance of SVRE.

5. Conclusion

In this work, we propose a novel method called the stochastic variance reduced ensemble (SVRE) attack to enhance the transferability of the crafted adversarial examples. Different from the existing model ensemble attacks that simply fuse the outputs of multiple models evenly, the proposed SVRE takes the gradient variance of different models into account and reduces the variance to stabilize the gradient update on ensemble attacks. In this way, SVRE can craft adversarial examples with higher transferability for other possible models. Extensive experiments demonstrate that SVRE outperforms the vanilla model ensemble attack in the black-box setting significantly, at the same time SVRE keeps roughly the same performance in the white-box setting.

Compared with broad investigations on the gradient optimization or input transformation attacks, the ensemble attacks are less explored in previous studies. Our work could shed light on the great potential of boosting the adversarial transferability through a better design on the ensemble methods. In future work, we wish our work could inspire more in-depth works in this direction for ensemble attacks.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62076105) and Hubei International Cooperation Foundation of China (2021EHB011).

References

[1] 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, pages 274–283, 2018. 2

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 1

[3] 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*, pages 9185–9193, 2018. 1, 2, 3, 4, 5

[4] 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*, pages 4312–4321, 2019. 1, 2, 4, 5

[5] 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*, pages 307–322, 2020. 1, 3, 5

[6] Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: Lightweight adversarial training with data augmentation improves neural network training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2474–2483, June 2021. 2

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations,ICLR*, 2015. 1, 2, 4, 5

[8] 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. 2, 3, 5

[9] 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*, pages 770–778, 2016. 5

[10] 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. 2, 3, 5

[11] 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*, pages 6084–6092, 2019. 2, 3, 5

[12] 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.*, pages 315–323, 2013. 2, 4

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5nd International Conference on Learning Representations,ICLR (Workshop)*, 2017. 1, 2

[14] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR*, 2017. 3

[15] 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*, pages 1778–1787, 2018. 2, 3, 5

[16] 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. 1, 2, 3, 4, 5

[17] 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. 1, 2, 3

[18] 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*, pages 860–868, 2019. 2, 3, 5

[19] 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. 1, 2, 3

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 86–94, 2017. 1

[21] 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*, pages 259–268, 2020. 2, 3, 5

[22] 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*, pages 506–519, 2017. 1

[23] 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)*, pages 372–387, 2016. 1

[24] 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. 2

[25] 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. 2

[26] 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*, pages 4278–4284, 2017. 5

[27] 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*, pages 2818–2826, 2016. 5

[28] 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. 1, 2

[29] 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. 1

[30] 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. 2, 3, 5

[31] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 4

[32] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. *CoRR*, abs/2102.00436, 2021. 2

[33] 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. 2, 3, 5

[34] 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*, pages 501–509, 2019. 2, 3

[35] 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*, pages 2730–2739, 2019. 1, 2, 3, 5

[36] 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. 2, 3, 5

[37] 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. 2