An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability

Bin Chen[1] Jiali Yin[1] Shukai Chen[1] Bohao Chen[2] Ximeng Liu[1]

[1]Fuzhou University, Fujian, China [2]Yuan Ze University, Taipei, Taiwan

c_chenbin@foxmail.com, jlyin@fzu.edu.cn, chenshukai770@163.com, {hd840207, snbnix}@gmail.com

Abstract

While the transferability property of adversarial examples allows the adversary to perform black-box attacks (i.e., the attacker has no knowledge about the target model), the transfer-based adversarial attacks have gained great attention. Previous works mostly study gradient variation or image transformations to amplify the distortion on critical parts of inputs. These methods can work on transferring across models with limited differences, i.e., from CNNs to CNNs, but always fail in transferring across models with wide differences, such as from CNNs to ViTs. Alternatively, model ensemble adversarial attacks are proposed to fuse outputs from surrogate models with diverse architectures to get an ensemble loss, making the generated adversarial example more likely to transfer to other models as it can fool multiple models concurrently. However, existing ensemble attacks simply fuse the outputs of the surrogate models evenly, thus are not efficacious to capture and amplify the intrinsic transfer information of adversarial examples. In this paper, we propose an adaptive ensemble attack, dubbed AdaEA, to adaptively control the fusion of the outputs from each model, via monitoring the discrepancy ratio of their contributions towards the adversarial objective. Furthermore, an extra disparity-reduced filter is introduced to further synchronize the update direction. As a result, we achieve considerable improvement over the existing ensemble attacks on various datasets, and the proposed AdaEA can also boost existing transfer-based attacks, which further demonstrates its efficacy and versatility. The source code: https://github.com/CHENBIN99/AdaEA

1. Introduction

Deep neural networks (DNNs), including convolutional neural networks (CNNs) [10, 36, 15] and vision transformers (ViTs) [6, 26, 19], have brought impressive advances to the state-of-the-art across various machine-learning tasks. At the moment, however, they are found to be vulnerable to
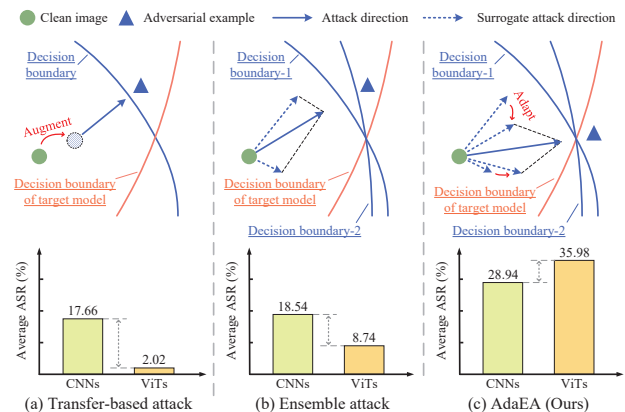

Figure 1. Overview of different attack schemes and performance. (a) Transfer-based methods strengthen the critical parts in images to improve the attack transferability, but fail to transfer across DNNs with wide differences due to the limited adversarial information. (b) Model ensemble attacks integrate multiple surrogate models for finding the more transferable attack, but existing works generally neglect the individual characteristics of each model, leading to under-optimal results. (c) Our AdaEA performs adaptive ensemble by amplifying the transferable information in each surrogate model and achieves remarkable improvements.

adversarial examples [25], *i.e.*, adding imperceptible hand-crafted perturbations to the original inputs can lead to wrong prediction behavior of DNNs. This discovery arises severe security hazards in the deployment of DNNs. More importantly, some well-designed adversarial examples can transfer across models. That is, an adversarial example crafted from a surrogate model can also disturb other models. This property of adversarial examples, known as *tranferability*, allows the adversary to attack a target model without knowing its interior, thus poses a more realistic threat to *black-box* applications (*i.e.*, the architectures and parameters are inaccessible to users).

To set up the first step for improving model robustness and prevent potential threats from black-box attacks, the research on improving the transferability of adversarial examples has attracted wide attention in recent years. The attack

transfer success rates vary depending on the difference between the surrogate and target models, the more similar the surrogate and target models are, the higher transfer success rate can be achieved. Thus a bunch of works have been proposed to improve the transferability of adversarial examples by maximizing the perturbation on critical parts that are shared among DNNs. The mainstream strategies include maximizing information from important neurons [37, 30], increasing input diversity [32, 1], and incorporating momentum [4, 29] into iterative-based attack. Despite their effectiveness, these methods always fail in transferring across models with wide architecture differences (*i.e.*, CNNs and ViTs), as shown in Figure 1 (a).

Similar to traditional ensemble methods which draw on the wisdom of multiple weak learners with diverse predictions to improve the overall accuracy, a line of research proposes to utilize an ensemble of surrogate models to generate adversarial examples that can successfully attack all the surrogate models. Intuitively, the approach can improve the transferability of adversarial examples as it can potentially capture intrinsic transferable adversarial information since the adversary can fool several models with wide differences concurrently. Moreover, such an ensemble could also be easily incorporated with existing transfer-based adversarial attacks without conflict. Several model ensemble based methods have been explored [18, 11], however, most of them only equally fuse the outputs (*i.e.*, logits or losses) of all models to get an ensemble loss for applying gradient-based attack, which may limit the potential capability of the model ensemble attacks, as shown in Figure 1 (b). Although a recent work [33] noticed the gradient variances among the surrogate models, the ensemble is still under-optimal due to the ignorance of individual characteristics of each model.

In this paper, we focus on the model ensemble adversarial attack for improving the transferability of adversarial examples. We first observe that simply averaging the outputs of ensemble models ignore the advantages of each model, where the transferable information captured from one model can be smoothed by another model during the fusion process, thus leading to the under-optimized results. To cope with this problem, we propose to adptively ensemble the outputs of each model via the adaptive gradient modulation (AGM) strategy. Specifically, we define the *adversarial ratio* to evaluate the contribution discrepancy among the surrogate models to the overall adversarial objective, which is then exploited to adaptively modulate the gradient fusion, offering more efforts on the amplification of transferable information in the generated adversarial examples. Moreover, the ensemble gradient may greatly differ or even oppose with the individual gradient of surrogate models, which has been proven to have a correlation with the overfitting problem in ensemble [33]. Hence, we further introduce a disparity-reduced filter (DRF) where a

disparity map is computed to reduce the variances among surrogate models and synchronize the update direction. Finally, the adversarial transferability could be enhanced by applying the above two mechanisms, as demonstrated in Figure 1. We term the proposed method as adaptive esnemble attack (AdaEA), and perform extensive experiments on diverse datasets to validate that our AdaEA can consistently outperform the existing methods. To sum up, the key contributions of this work are three-fold:

- We propose an adaptive ensemble adversarial attack, dubbed AdaEA, which offers a more comprehensive ensemble attack for a broad class of models with wide architecture differences, such as CNNs and ViTs.

- Our AdaEA views the ensemble attack from the gradient optimization perspective, and controls the optimization process via AGM strategy as well as reducing the disparity by DRF to synchronize the optimization direction.

- The proposed AdaEA can not only largely enhance the ensemble effectiveness compared to existing ensemble methods, but also consistently improve the attack performance when incorporated with the existing transfer-based gradient attacks.

2. Related Works

2.1. Adversarial Attacks

Since Szegedy *et al.* [25] first reported the existence of adversarial examples, extensive efforts have been devoted to highlighting the vulnerability of DNNs. An adversarial attack usually produces adversarial examples by adding a perturbation δ to an original input image x with the objective that can make the model discriminative loss $\mathcal{L}$ maximized, *i.e.*, $\arg\max_{x+\delta} \mathcal{L}(f(x + \delta), y)$. To make the perturbation imperceptible, the perturbation δ is subject to a constraint $\mathcal{S}$, which is defined as $\mathcal{S} = \{\|\delta\|_p \leq \epsilon\}$ by the given ℓ_p-norm distance and the maximum strength ϵ.

Gradient-based adversarial attacks. To optimize the attack objective, the gradient information are usually used to maximize the model loss. Goodfellow *et al.* [8] designed a Fast Gradient Sign Method (FGSM) to produce strong adversarial examples based on the investigation of CNN linear nature. Wang *et al.* [28] and Madry *et al.* [21] further broke the one-step generation of perturbation in FGSM into iterative generation and proposed I-FGSM and Projected Gradient Descent (PGD) attack. While these attacks can exhibit high attack success rate on the white-box models, they usually reveal low transfer rate to black-box models since the gradients information is hard to approximate.

Transfer-based adversarial attacks. To improve the transferability, existing works try to maximize the distortion on the critical parts of inputs. Wang *et al.* [30]
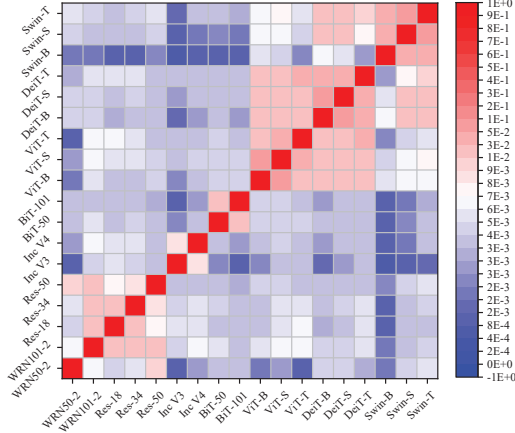
Figure 2. Visualization of the cosine similarity between the gradients produced from different models. Note that the gradients are closer when the model architectures are more similar.

and Zhang *et al*. [37] investigated the distortion on features based on the importance of neural in the DNNs. Xie *et al*. [32] and Dong *et al*. [5] incorporated the FGSM with either input diversity or translation-invariant strategies to produce diverse input patterns for generation of adversarial examples. Gao *et al*. [7] proposed the PI-FGSM which generates patch-wise perturbation rather than pixel-wise, that is beneficial for black-box attack. Although these attacks can achieve transferability improvements over the primordial gradient-based attacks, they can hardly transfer to the new architecture of DNNs, *i.e.*, the ViT family.

Model ensemble attacks. Ensemble attack methods usually craft adversarial examples by performing a weighted linear sum of the multiple white-box attacks in parallel. Liu *et al*. [18] directly averaged the predictions of multiple modes to get an ensemble loss for applying gradient-based attack. Dong *et al*. [5] further fused the logits and losses of ensemble models. Xiong *et al*. [33] noticed the variance among the ensemble models and proposed a stochastic variance-reduced ensemble (SVRE) attack to improve the attack generalization. While improvements being achieved, the ensemble is still under-optimal due to the less investigation in the individual advantages of each model.

2.2. Adversarial Defenses

As the counterpart of adversarial attack, enormous efforts have been proposed to defend against adversarial examples, which generally fall into two categories. The first is referred to as adversarial training [21, 2, 38, 27, 23, 14, 35], which is regarded as the most reliable and effective method. Its key idea is to leverage the online generated adversarial examples into the training dataset so that the model can prefer more robust features during learning [21]. To improve the defense efficiency, state-of-the-art methods further propose to incorporate curriculum attack generation [2], early stopping [38], and ensemble

schemes [27, 23, 14, 35]. The second line of adversarial defense is input transformation-based methods, which aim to eliminate the adversarial information from adversarial examples by preprocessing. Many state-of-the-art defense methods for defending against adversarial examples have been proposed, including denoising images with high-level representation [17], randomly resizing [31] and smoothing [12], compressing input image [13, 9, 34, 20] and purifying the input images using neural network [22]. In this paper, we employ these state-of-the-art defenses to evaluate the effectiveness of our attack method.

3. Methodology

3.1. General Overview

Improving the transferability of adversarial examples aims to make an adversarial example generated from a white-box surrogate model stay adversarial to holdout black-box models. Typically, using a gradient-based method to iteratively find the optimal perturbation for a white-box model can be given by:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \operatorname{sign}(\nabla_{x_t^{adv}} \mathcal{L}(f(x_t^{adv}), y)), \quad (1)$$

where $\operatorname{sign}(\cdot)$ is the sign function, α is the step size, and $\nabla_{x_t^{adv}} \mathcal{L}$ denotes the gradient of the loss function $\mathcal{L}$ w.r.t. x_t^{adv}. Note that x_1^{adv} is set to be x, and the final adversarial example is obtained by x_T^{adv}, T is the iteration number. Intuitively, it can achieve high attack successful rate under the white-box setting, where $\nabla_{x_t^{adv}} \mathcal{L}$ is known. However, when transferred to a black-box model in which the $\nabla_{x_t^{adv}} \mathcal{L}$ is unknown, the attack successful rate would be dropped since the gradients are diverse in different models, as shown in Figure 2. In particular, when the model architectures significantly differ, such as ViTs and CNNs, the gradients are extremely different, leading to a lower transfer attack rate.

To make the generated adversarial examples adversarial to a broad class of models, the ensemble attack is an effective strategy to enhance the attack transferability. The basic idea is to utilize the outputs of multiple white-box models to obtain the averaged model loss, and then the gradient-based attack is applied to generate the adversarial example. It transforms Eq. (1) into the following representation:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \operatorname{sign}(\nabla_{x_t^{adv}} \mathcal{L}(\sum_{k=1}^{K} w_k f_k(x_t^{adv}), y)), \quad (2)$$

where w_k is the ensemble weights for k-th surrogate model f_k, $\forall w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$; and K is the number of surrogate models. Existing ensemble methods generally average the logits [18], predicted probabilities [5], or losses [5] of surrogate models to obtain the ensemble loss for generating gradient information. However, such simple ensemble ignores the individual variance across the surrogate models, thus significantly limits the overall attack
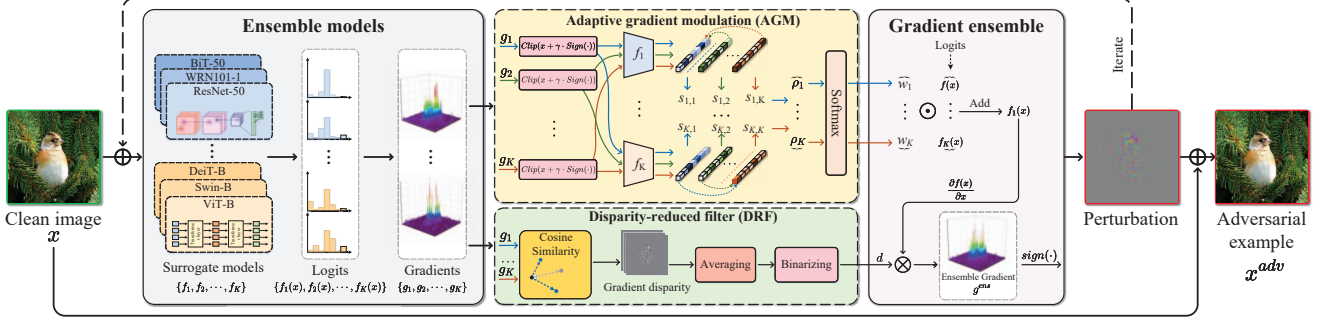
Figure 3. An overview of our AdaEA. The gradients obtained from CNNs and ViTs are feed into the AGM and DRF to get the ensemble gradient for generating adversarial examples with gradient-based attack.

performance. Let us take Figure 2 as an example again, as the gradients vary across different models, directly equally merging the outputs of models would lead to under-optimal results since the adversarial information captured by each model is not evaluated and amplified.

3.2. Adaptive Ensemble Adversarial Attack

In this work, we focus on the model ensemble methods following Eq. (2). Instead of directly averaging the outputs of surrogate models as the previous works, we propose AdaEA equipped with AGM and DRF mechanisms to amend the gradient optimization process for boosting the transferable information in the generated adversarial examples. Specifically, AGM first modulates the gradient of each ensemble model by the defined *adversarial ratio* which identifies the contribution discrepancy of each surrogate model to the overall adversarial object, and then the DRF further synchronizes the gradient update direction by filtering out the disparity part of ensemble gradients. An overview of AdaEA is shown in Figure 3.

Adaptive gradient modulation. After obtaining the outputs $f_i(x)$ and gradient information g_i from each surrogate model f_i by feeding the input image, *i.e.*, $g_i = \nabla_{x_t^{adv}} \mathcal{L}(f_i(x_t^{adv}), y)$, we propose to adaptively modulate the model ensemble via monitoring the discrepancy of their contributions to the adversarial attack objective. Specifically, for the i-th ensemble model f_i, we evaluate the potential adversarial transferability in the g_i by testing the attack performance of adversarial examples generated from g_i on other models, which we define as *adversarial ratio*, and then adjust the ensemble weight based on the adversarial ratio of each model. Here we first conduct the testing process by computing:

$$s_{k,i} = -\mathbf{1}_y \cdot \log\left(\text{softmax}\left(\mathbf{p}_k[x_t^{adv} + \alpha \, \text{sign}(g_i)]\right)\right) , \quad (3)$$

where $\mathbf{p}_k(\cdot)$ denotes the logits output from f_k, and $\mathbf{1}_y$ is the ground truth logits. $s_{k,i}$ can be considered as the k-th model loss on the adversarial example generated by using

the gradient from i-th model. We then define the adversarial ratio ρ_i as:

$$\rho_i = \frac{\beta}{K-1} \sum\nolimits_{k=1, k\neq i}^{K} \frac{s_{k,i}}{s_{k,k}} , \quad (4)$$

where β denotes a hyperparameter that controls the effect of ensemble weighting, which is further discussed in Sec. 4.3. Note that a higher value of ρ_i denotes a better transfer attack of adversarial example generated from g_i, implying that g_i contains more transferable adversarial information. By doing so, we can figure out which model can provide more generic adversarial information and adaptively assign a higher ensemble weight. Thus, according to the adversarial ratio of each model, we use a softmax function to normalize the ensemble weight of each model by:

$$w_1^*, w_2^*, ..., w_K^* = \text{softmax}(\rho_1, \rho_2, ..., \rho_K). \quad (5)$$

With the obtained w_i^*, the output of each surrogate model with more potential adversarial transferability information is amplified in the ensemble gradient of Eq. (2), thus leading to a higher transfer attack success rate on the hold-out black-box models.

Disparity-reduced filter. As discussed, the gradient optimization direction of surrogate models vary tremendously in a big range, sometimes the gradients walk towards direction against each other and the result leads to an overfit to the ensemble model [33]. To solve the problem and synchronize the update direction, we introduce an extra disparity-reduced filter to reduce the gradient variances among surrogate models. We first apply the cosine similarity to evaluate the deviation of gradients in surrogate models, and compute the disparity map d_i by averaging the similarity score with gradients of other models, which can be described as follows:

$$d_i^{(p,q)} = \frac{1}{K-1} \sum\nolimits_{k=1, k\neq i}^{K} \cos\left(\overrightarrow{g}_i^{(p,q)}, \overrightarrow{g}_k^{(p,q)}\right) , \quad (6)$$

where $\cos(\cdot)$ denotes cosine similarity function, $\overrightarrow{g}_i^{(p,q)}$ and $\overrightarrow{g}_k^{(p,q)}$ denote the vector extracted from the position (p, q)

through channels of gradient g_i and g_k, respectively. The final disparity map d for ensemble gradients is obtained by averaging all the d_i. We then clean the disparity part in the ensemble gradient by using a filter $\mathbf{B}$ as:

$$\mathbf{B}(p, q) = \begin{cases} 0, & \text{if } d_i^{(p,q)} \leq \eta \\ 1, & \text{otherwise} \end{cases}, \quad (7)$$

where η is the tolerance threshold for the disparity filtering. By filtering out the disparity part of the ensemble gradients, the gradient optimization direction can be synchronized. To this end, the ensemble gradient can be obtained by rewriting Eq. (2) as:

$$g_{t+1} = \nabla_{x_t^{adv}} \mathcal{L}(\sum_{k=1}^{K} w_k^* f_k(x_t^{adv}), y) \otimes \mathbf{B}, \quad (8)$$

where $\otimes$ denotes the element-wise multiplication. Hence, the disparity among the surrogate models can be suppressed. We provide more discussions about DRF in terms of both qualitative and quantitative analysis in the supplementary material. The overall AdaEA procedure is shown in Algorithm 1.

4. Experiments

4.1. Experimental Setting

Datasets. We conduct experiments on CIFAR-10, CIFAR-100 and ImageNet datasets [16, 3] which are widely used in both classification and adversarial attack tasks [33, 18].

Networks. We choose target models from both branches of CNNs and ViTs for the black-box attack task, including ResNet-50 (Res-50) [10], WideResNet-50 (WRN-50) [36], BiT-M-R50×1 (BiT-50) [15] and BiT-M-R101 (BiT-101) [15] in CNN branch; and ViT-Base (ViT-B) [6], DeiT-Base (DeiT-B) [26], Swin-Base (Swin-B) [19] and Swin-Small (Swin-S) [19] in ViT branch. As for surrogate models, we choose ResNet-18 (Res-18) [10], Inception v3 (Inc-v3) [24], ViT-Tiny (ViT-T) [6] and DeiT-Tiny (DeiT-T) [26] in the later experiments by default.

Comapred methods. Two pioneering ensemble attack methods, *i.e.*, Ens [18] and SVRE [33], are employed as baselines to compare with our AdaEA. All the ensemble methods follow the same ensemble settings in experiments.

Implementation details. For the baselines and our AdaEA, we use the I-FGSM with 20 iterations under l_∞ constraint as the basic attack method, and set $\epsilon = 8/255$ and $\alpha = 2/255$ during the adversarial example generation. As for hyperparameter, we set $\eta = -0.3$ in DRF and $\beta = 10$ in AGM. The inner update time in SVRE is set to be 4 following its default setting. All the experiments were implemented using Pytorch on an Intel Xeon Sliver and a NVIDIA A6000 GPU with 48GB graph memory.

Algorithm 1 The AdaEA algorithm

Input: Input (x, y), a list of K surrogate models. Maximum range of perturbation ϵ, the step size of iteration attack α, and the number of iterations in the inner gradient-based attack T.

Output: Adversarial example x^{adv}.

1   $x_1^{adv} \leftarrow x$
2   **for** $t \leftarrow 1$ **to** T **do**
3   # Calculating the gradients of all K models
4      $g_k \leftarrow \nabla_{x_t^{adv}} \mathcal{L}(f_k(x_t^{adv}), y)$
5   # Performing adaptive gradient modulation
6   Compute the adversarial ratio ρ_i of each model using Eqs. (3)-(4)
7   Compute the weight w for each model using Eq. (5)
8   # Performing disparity-reduced filter
9   Compute the disparity map d using Eq. (6)
10   # Ensemble the gradient
11   Compute the gradient g_{t+1}^{ens} using Eqs. (7)-(8).
12   # Updating adversarial example
13      $x_{t+1}^{adv} \leftarrow \text{Clip}_x^\epsilon \{x_t^{adv} + \alpha \, \text{sign}(g_{t+1}^{ens})\}$
14   **end for**
15   $x^{adv} \leftarrow x_T^{adv}$

4.2. Main Results

General attack performance. We first compare the general attack performance of AdaEA with existing ensemble methods on the naturally trained models under the black-box setting on CIFAR-10/100 and ImageNet. Table 1 reports the attack results on a broad class of black-box models, including both CNNs and ViTs. As we can see, SVRE can slightly improve the attack performance by reducing gradient variance across models compared to the baseline Ens. The improvements in terms of average success rates are around 3% on CIFAR datasets. In contrast, our AdaEA can improve the attack transfer rate by a large margin, where we achieve more than 15% averaging improvements over SVRE on CIFAR-10, demonstrating the effectiveness of our AdaEA in finding and amplifying the intrinsic adversarial information of inputs via the AGM-DRF strategies.

Combinations with transfer-based attacks. We then attempt to test the integration of the existing transfer-based attacks in our AdaEA. We additionally use FGSM, MI-FGSM [4], and DI²-FGSM [32] as the base attacks for ensemble, and summarize the results in Table 2. The results show that the attack success rate significantly improves combined with our AdaEA regardless of base attacks. Specifically, for FGSM and I-FGSM, using our AdaEA improves the average transfer success rate around 20%. For MI-FGSM and DI²-FGSM attacks, our method also achieves consistently improvements over the existing ensemble attacks by a large margin, further indicating the

Table 1. The black-box attack success rate (%) against eight naturally trained models. The bolded numbers indicate the best results and Δ represents the improvements over the baseline.

Dataset	Attack	Res-50	WRN101-2	BiT-50	BiT-101	ViT-B	DeiT-B	Swin-B	Swin-S	Average (Δ)
	Ens	50.42	26.85	21.83	17.61	11.59	26.15	22.61	35.42	26.56
CIFAR-10	SVRE	54.08	28.47	23.28	19.06	13.83	31.00	25.17	40.17	29.38 (+2.82)
	AdaEA	**61.54**	**38.07**	**33.36**	**28.99**	**31.77**	**59.72**	**45.90**	**61.38**	**45.09 (+18.53)**
	Ens	80.13	67.89	60.79	44.78	45.46	69.50	64.40	77.14	63.76
CIFAR-100	SVRE	82.06	68.68	62.59	46.30	48.11	73.63	67.94	80.49	66.23 (+2.47)
	AdaEA	**82.19**	**70.02**	**65.28**	**48.63**	**60.20**	**82.83**	**75.21**	**84.41**	**71.10 (+7.34)**
	Ens	52.90	58.10	56.86	48.27	39.94	51.38	25.95	37.66	46.38
ImageNet	SVRE	**53.10**	57.84	56.90	48.38	40.03	52.06	25.54	37.26	46.39 (+0.01)
	AdaEA	**53.10**	**58.33**	**58.57**	**50.06**	**46.13**	**58.05**	**29.37**	**41.30**	**49.36 (+2.98)**

Table 2. The attack success rate (%) of adversarial examples generated by ensemble attacks based on different attack methods on CIFAR-10. The bolded numbers indicate the best results and Δ represents the improvements over the baseline.

Base	Attack	Res-50	WRN101-2	BiT-50	BiT-101	ViT-B	DeiT-B	Swin-B	Swin-T	Average (Δ)
	Ens	21.32	16.22	12.58	10.69	7.17	10.68	8.30	15.23	12.77
FGSM [8]	SVRE	26.05	20.61	21.26	18.87	17.84	22.23	17.66	25.99	21.31 (+8.54)
	AdaEA	**32.96**	**31.41**	**34.35**	**32.57**	**38.40**	**45.83**	**35.82**	**43.78**	**36.89 (+24.12)**
	Ens	50.42	26.85	21.83	17.61	11.59	26.15	22.61	46.93	28.00
I-FGSM [28]	SVRE	51.92	27.50	22.90	18.29	13.30	30.74	24.84	51.01	30.06 (+2.06)
	AdaEA	**61.54**	**38.07**	**33.36**	**28.99**	**31.77**	**59.72**	**45.90**	**70.77**	**46.27 (+18.27)**
	Ens	55.10	33.89	29.68	25.28	20.96	42.12	31.30	58.20	37.07
MI-FGSM [4]	SVRE	31.46	21.37	18.53	16.21	15.53	26.86	20.70	33.69	23.04 (-14.03)
	AdaEA	**66.58**	**44.45**	**41.90**	**37.23**	**45.96**	**70.78**	**53.61**	**78.00**	**54.81 (+17.74)**
	Ens	90.28	67.34	63.06	57.65	51.19	82.44	76.31	91.26	72.44
DI^2-FGSM [32]	SVRE	39.30	32.12	29.78	27.41	26.82	36.99	35.35	40.20	33.49 (-38.95)
	AdaEA	**91.49**	**74.08**	**72.26**	**68.83**	**66.96**	**89.23**	**84.48**	**95.20**	**80.32 (+7.88)**

promising versatility of our proposed AdaEA.

Attack advanced defense models. We also evaluate AdaEA on attacking models with various advanced defenses, including adversarial training defenses and input transformation-based defenses. The results are summarized in Table 3. For adversarial training defense, we use adversarial trained Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and Inc-v2$_{ens}$ networks as the target model following previous works [33, 27]. But unlike they set the surrogate model as the same architecture as the model used in ensemble training, we set the experiments under a more challenging scenario where we use totally different architectures as surrogate models (*i.e.*, our default settings). As we can see from Table 3, despite the challenge to attack an adversarially trained black-box model, our AdaEA exhibits the strongest attack performance among the compared methods. For the input transformation-based defenses, we adopt six popular input transformation-based defenses to test the attack performance of each method. From the results in columns seven to thirteen of Table 3, AdaEA achieves the best results where it surpasses the baseline by 7.9, 8.27 and 4.93 on the base I-FGSM, MI-FGSM, and DI^2-FGSM attack, respectively.

Visualization of attack performance. To intuitively show the attack performance, we visualize the heatmaps of clean image and adversarial examples generated by different ensemble methods in both white-box and black-box models in Figure 4. As can be observed in the Figure 4 (b) and (c), the attention of the white-box models changes on all the generated adversarial images compared with the clean image, which indicates that the generated adversarial examples can effectively trigger the wrong prediction of these models. However, when transferred to black-box models, the Ens and SVRE methods fail to mislead the model attention where the heatmaps are similar to the clean image, as shown in the second to third rows of Figure 4 (d)-(e). In contrast, thanks to the amplification of potential intrinsic adversarial information via AGM-DRF schemes in AdaEA, the generated adversarial example can still fool the attention of black-box models where the attention is dramatically changed in Figure 4 (d)-(e).

4.3. Ablation Studies

In this subsection, we conduct a series of ablation experiments to study the effects of key components and hyperparameters in our AdaEA.

Table 3. The robust accuracy (%) against three adversarial training models and six advanced defense methods on CIFAR-10. The results of input transformation-based defenses are the average results of all target models. The bolded numbers indicate the best results.

| Base	Attack	Adversarial training defense				Input transformation-based defenses						
		Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Inc-v2$_{ens}$	Avg.	R&P	Bit-R	JPEG	ComDefend	RS	NRP	Avg.
I-FGSM	Ens	0.54	0.67	0.55	0.59	18.98	32.75	23.58	83.82	57.44	13.07	38.27
	SVRE	0.64	0.79	0.65	0.69	20.56	35.94	26.35	83.77	57.77	12.86	39.54
	AdaEA	**0.79**	**0.98**	**0.79**	**0.85**	**26.93**	**49.67**	**40.20**	**84.06**	**59.65**	**16.51**	**46.17**
MI-FGSM	Ens	0.73	0.99	0.75	0.82	26.38	43.51	36.10	83.94	58.56	5.11	42.27
	SVRE	0.55	0.65	0.66	0.62	16.41	25.39	23.08	83.74	56.67	3.91	34.87
	AdaEA	**1.14**	**1.38**	**1.21**	**1.24**	**37.31**	**60.90**	**53.74**	**84.21**	**61.64**	**5.41**	**50.54**
DI^2-FGSM	Ens	1.47	1.72	1.79	1.66	62.92	76.80	72.54	84.16	60.96	5.30	60.44
	SVRE	0.85	1.02	1.01	0.96	30.77	34.46	33.79	83.77	57.75	4.28	40.80
	AdaEA	**2.27**	**2.49**	**2.50**	**2.42**	**71.83**	**82.24**	**79.99**	**84.37**	**64.90**	**8.92**	**65.37**
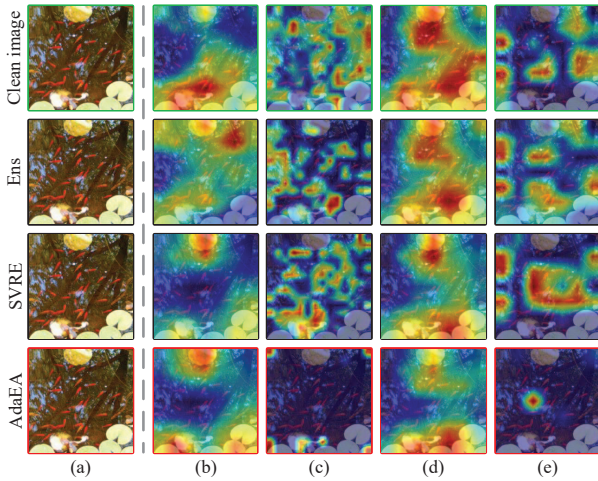
Figure 4. Heatmaps of different inputs in the surrogate models and black-box models. (a) input images, including clean image and adversarial examples generated by each attack method. (b)-(e) are the heatmaps on the surrogate models (Res-18, ViT-T) and black-box models (WRN50-2, Swin-T), respectively.

On the components of AdaEA. We first examine the effectiveness of AGM and DRF mechanisms in our AdaEA. Specifically, we perform four ensemble methods: the naive ensemble attack, ensemble with AGM, ensemble with DRF, and our AdaEA involving both AGM and DRF on black-box attacks. The results are reported in Table 4. As can be seen, using AGM can effectively enhance the attack transferability with 12.50% averaging improvements, indicating its effectiveness in amplification of adversarial information during gradient ensemble. It is interesting to see that adding DRF into baseline brings significant improvements on the transferability to ViTs, *i.e.*, 23.94% → 47.57%. This is due to the wide differences across CNNs and ViTs, reducing the gradient disparity among the CNNs and ViTs can provide more stable and better attack performance. In general, AGM together with DRF can provide the best transferability with a large improvements over the baseline, *i.e.*, 27.29% → 44.78% in average.

Table 4. Experimental results of average attack success rate (%) on the component ablations in AdaEA.

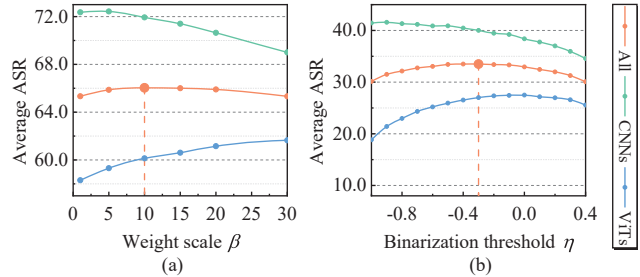| Ens models	Method	CNNs	ViTs	All (Δ)
Res-18,	Ens	29.96	23.94	27.29
Inc-v3,	+AGM	39.48	40.18	39.79 (+12.5)
ViT-T,	+DRF	38.31	47.57	42.42 (+15.13)
DeiT-T	AdaEA	**40.85**	**49.69**	**44.78 (+22.4)**

Figure 5. Ablation study on (a) weighting scale β in AGM and (b) binarization threshold η in DRF.

On hyper-parameter sensitivity. We study the sensitivity of our AdaEA to the weighting scale β in Eq. (4) and the binarization threshold η in Eq. (7). We use Res-18 and ViT-T as the surrogate models for ensemble, and show the curves of averaging success rate on black-box CNNs, ViTs, and all the models in Figure 5. As we can see in Figure 5 (a), a larger value of β leads to better trasferability to ViTs but lower transferability to CNNs. This suggests that the gradients of ViTs play a critical role in AGM process, a larger β can amplify the focus on ViTs. We set $\beta = 10$ as the average attack success rate on all the target models reaches the peak at $\beta = 10$. For the binarization threshold η in Figure 5 (b), the transferability to ViTs gains large improvements by reducing the disparity as η increases, but the transferability to CNNs shows a bit drop. The average performance on all the models increases and reaches the peak at $\eta = -0.3$.

4.4. Further Analysis

Since our work is among the first grups to study the adversarial transfer across both CNNs and ViTs, we further

Table 5. Comparison of average attack success rate (%) between ensemble attack and our AdaEA under different ensemble models on CIFAR-10. Bolded numbers signify better results.

Ensemble models	# CNNs	# ViTs	Attack	CNNs				ViTs			
				Res-50	WRN101-2	BiT-101	Average	ViT-B	DeiT-B	Swin-S	Average
Inc-v3, DeiT-T	1	1	Ens	14.52	8.09	5.73	9.45	3.76	7.18	8.79	6.58
			AdaEA	**29.02**	**18.53**	**19.25**	**22.27**	**20.52**	**36.75**	**33.09**	**30.12**
Res-18, Inc-v3, ViT-T	2	1	Ens	43.39	22.83	13.23	26.48	5.12	10.58	21.68	12.46
			AdaEA	**49.30**	**27.03**	**16.51**	**30.95**	**8.47**	**17.71**	**29.61**	**18.60**
Inc-v3, ViT-T, Swin-T	1	2	Ens	24.19	12.70	9.34	15.41	6.19	13.07	71.43	30.23
			AdaEA	**39.07**	**20.86**	**16.95**	**25.63**	**15.02**	**33.33**	**95.66**	**48.00**
Res-18, Inc-v3 BiT-50	3	0	Ens	52.86	31.69	68.21	50.92	4.15	7.08	21.01	10.75
			AdaEA	**60.27**	**37.90**	**72.20**	**56.79**	**5.28**	**9.49**	**25.97**	**13.58**
ViT-T, DeiT-T, Swin-T	0	3	Ens	**52.70**	29.41	27.90	**36.67**	38.76	71.60	**99.00**	69.79
			AdaEA	50.14	**30.05**	**29.41**	36.53	**45.92**	**75.25**	97.05	**72.74**
Res-18, Inc-v3, ViT-T, DeiT-T	2	2	Ens	50.42	26.85	17.61	31.63	11.59	26.15	35.42	24.39
			AdaEA	**61.54**	**38.07**	**28.99**	**42.87**	**31.77**	**59.72**	**61.38**	**50.96**
Res-18, ViT-T, DeiT-T, Swin-T	1	3	Ens	66.79	38.00	26.49	43.76	21.20	47.75	94.53	54.49
			AdaEA	**71.39**	**42.88**	**34.70**	**49.66**	**44.45**	**76.05**	**98.00**	**72.83**
Res-18, Inc-v3, BiT-50, ViT-T	3	1	Ens	61.66	37.43	72.86	57.32	9.64	18.64	39.08	22.45
			AdaEA	**69.91**	**45.16**	**76.39**	**63.82**	**14.64**	**27.88**	**49.15**	**30.56**

analyze the transferability of adversarial examples from the perspective of surrogate models used during the ensemble by considering the following questions.

What effect does the number of surrogate models have on the transferability? We first test the effect of different numbers of surrogate models on the ensemble attack performance. From Table 5, we can see that as the number of surrogate model increases, the overall attack success rate improves from the first row to the bottom row. The ensemble using four surrogate models improves the average success rate by around 20% on both CNNs and ViTs over using two surrogate models, as can been seen in the second and seventh rows of Table 5. Intuitively, using more surrogate models can lead to better transferability since more adversarial information can be captured. More importantly, our AdaEA consistently improves the ensemble attack performance regardless the number of ensemble models.

How does different proportions of CNNs to ViTs in surrogate models affect the overall transferability? As CNNs and ViTs are two main branches in the family of DNNs, we investigate the effects of the proportions of CNNs to ViTs in the surrogate models on the overall transferability. By observing the second, third, and ninth rows of Table 5, as the number of CNNs increases in the surrogate models, the attack rate on CNNs obviously improves. But in contrast, the attack success rate on ViTs is not going higher. This indicates that the ensemble gradient focuses more on the gradients of CNNs when the CNNs dominate in the surrogate models. When only CNNs are used as surrogate models in the fifth row of Table 5, the attack has high success rates on CNNs but reveals a low transfer rate on ViTs. But interestingly, when the proportion of CNNs to

ViTs becomes $0:3$ in the sixth row of Table 5, where only ViTs are used, the ensemble attack still exhibits a high transfer rate to CNNs. The same results can be seen in the fourth and eighth rows of Table 5 when the ViTs dominate the surrogate models, the transfer to CNNs can still maintain a high attack success rate. This phenomenon indicates that *it is easier to transfer attacks from ViTs to CNNs compared with transferring from CNNs to ViTs*. We attribute this to the more complex architecture and global modeling ability of ViTs, which makes ViTs capable of extracting more generic adversarial information.

5. Conclusion

In this work we propose AdaEA, an adaptive ensemble adversarial attack that merges the gradients of surrogate models via monitoring on the contribution of each model to the overall adversarial objective, for boosting the transferability of adversarial examples. We show that AdaEA can effectively enhance the adversarial transferability across models with a large margin over the existing ensemble methods under various settings, even those with wide architecture differences, *e.g.*, CNNs and ViTs, which demonstrates the effectiveness of our method in capturing intrinsic adversarial information of inputs.

6. Acknowledgements

References

[1] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 15244–15253, June 2022.

[2] Qi-Zhi Cai, Chang Liu, and Dawn Song. Curriculum adversarial training. In *Proc. Int'l Joint Conf. Artif. Intell.*, pages 3740–3747, Stockholm, Sweden, 2018.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, June 2009.

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.

[5] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int'l Conf. Learn. Repres.*, 2021.

[7] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Proc. Euro. Conf. Comput. Vis.*, pages 307–322, 2020.

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. Int'l Conf. Learn. Repres.*, 2015.

[9] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, jun 2016.

[11] Ziwen He, Wei Wang, Xinsheng Xuan, Jing Dong, and Tieniu Tan. A new ensemble method for concessively targeted multi-model attack. *arXiv preprint arXiv:1912.10833*, 2019.

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

[13] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6085, 2019.

[14] Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv: 1901.09981*, 2019.

[15] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Proc. Euro. Conf. Comput. Vis.*, pages 491–507, 2020.

[16] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009.

[17] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 1778–1787, 2018.

[18] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. Int'l Conf. Learn. Repres.*, 2017.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int'l Conf. Comput. Vis.*, pages 9992–10002, 2021.

[20] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2019.

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. Int'l Conf. Learn. Repres.*, 2018.

[22] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 259–268, 2020.

[23] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *Proc. Int'l Conf. Machine Learn.*, pages 8759–8771, 2019.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 1–9, June 2015.

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & amp; distillation through attention. In *Proc. Int'l Conf. Machine Learn.*, volume 139, pages 10347–10357, 18–24 Jul 2021.

[27] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *Proc. Int'l Conf. Learn. Repres.*, 2018.

[28] Jiakai Wang. Adversarial examples in physical world. In *Proc. Int'l Joint Conf. Artif. Intell.*, pages 4925–4926, 8 2021.

[29] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. In *Proc. British Conf. Machine Vis.*, 2021.

[30] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren. Feature importance-aware transferable adversarial attacks. In *Proc. IEEE Int'l Conf. Comput. Vis.*, pages 7619–7628, 2021.

[31] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[32] 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 *Proc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[33] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 14983–14992, June 2022.

[34] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[35] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: Diversifying vulnerabilities for enhanced robust generation of ensembles. In *Proc. Adv. Neural Inform. Process. Syst.*, 2020.

[36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proc. British Conf. Machine Vis.*, pages 87.1–87.12, September 2016.

[37] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 14993–15002, June 2022.

[38] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *Proc. Int'l Conf. Machine Learn.*, 2020.