Chinese Journal of Electronics
Vol.32, No.1, Jan. 2023

Towards Evaluating the Robustness of Adversarial Attacks Against Image Scaling Transformation

ZHENG Jiamin[1], ZHANG Yaoyuan[2], LI Yuanzhang[2], WU Shangbo[2], and YU Xiao[3]

(1. *School of E-Business and Logistics, Beijing Technology and Business University, Beijing 100048, China*)

(2. *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*)

(3. *School of Computer Science and Technology, Shandong University of Technology, Shandong 255000, China*)

Abstract — The robustness of adversarial examples to image scaling transformation is usually ignored when most existing adversarial attacks are proposed. In contrast, image scaling is often the first step of the model to transfer various sizes of input images into fixed ones. We evaluate the impact of image scaling on the robustness of adversarial examples applied to image classification tasks. We set up an image scaling system to provide a basis for robustness evaluation and conduct experiments in different situations to explore the relationship between image scaling and the robustness of adversarial examples. Experiment results show that various scaling algorithms have a similar impact on the robustness of adversarial examples, but the scaling ratio significantly impacts it.

Key words — Adversarial examples, Image scaling, Image classification, Deep learning.

I. Introduction

Machine learning has many applications in malware detection [1], Internet of things security [2], [3], cyber security [4], computing servers [5], and mobile security [6]. However, recent work [7] have shown that deep neural networks (DNNs) are vulnerable to adversarial examples [8].

An increasing number of researches on generating adversarial examples has accelerated the risks for applying the deep learning models in the safety-critical fields [9], [10]. However, most of these researches focus on attacking the target models to expose vulnerability; even though the studies on defenses [11] concentrate on improving the robustness of deep learning models, the vulnerability of adversarial examples has not been studied much. Xie *et al.* [12] mention that adding a random image transformation [13] during the iterative processes of adversarial attacks increase the transferability of adversarial examples [14]. This phenomenon inspires us to study whether the adversarial examples can remain adversarial after the image geometric transformation or whether the adversarial examples are robust to the image geometric transformation.

In this paper, we focus on the relationship between image scaling and adversarial examples. Dong *et al.* [15] establish a comprehensive benchmark to evaluate the robustness of defense models against adversarial attacks on the image classification tasks. Based on their work, we define the various metrics involved in the image scaling process and design multiple sets of experiments to quantitatively evaluate the impact of image scaling on the robustness of adversarial examples. We conduct experiments for robustness evaluation from three aspects: models, perturbation budgets and image scaling processes.

Experiment results show that adversarial examples often cannot maintain the adversarial after image scaling, which leads to the failure of the adversarial attacks. We hope our research can provide a reference for designing adversarial attack algorithms that generate more robust adversarial examples and also provide an idea for designing better methods to defend against the adversarial attacks [16].

The remaining of this paper is organized as follows: We review the background and related work in Section Ⅱ. We design the evaluation methodology in Section Ⅲ.

Manuscript Received Aug. 26, 2021; Accepted Nov. 30, 2021. This work was supported by the National Natural Science Foundation of China (61876019, U1936218, 62072037).

DOI:10.23919/cje.2021.00.309

The experiment results and analysis are presented in Section Ⅳ. We draw the conclusion in Section Ⅴ.

II. Background and Related Work

1. Threat model

Adversarial attacks can be classified into targeted attacks and non-targeted attacks. Non-targeted attacks only need to construct adversarial samples that misclassify the classifier without paying attention to the specific category. Targeted attacks misclassify the classifier as a particular class. We choose the threat models defined in [15] as the foundation for robustness evaluation and focus on non-targeted attacks. For a classifier $\mathcal{C}(\cdot)$, a non-targeted attack aims to construct an adversarial sample x_{adv} to satisfy (1):

$$\mathcal{C}(x_{\mathrm{adv}}) \neq C(x) \tag{1}$$

where $C(x)$ denotes the ground-truth label of the original image x.

Adversaries often have two strategies to construct tiny perturbations: a fixed ϵ attack based on model gradients and a minimized perturbation attack based on optimization.

Gradient-based fixed ϵ attacks This strategy searches for the adversarial examples x_{adv} to satisfy $||x_{\mathrm{adv}} - x||_p \leq \epsilon$ within a specific perturbation budget ϵ. A Gradient-based fixed ϵ attack solves the problem described in (2) by maximizing the loss function $\mathcal{J}$ to get a non-targeted adversarial example:

$$x_{\mathrm{adv}} = \mathrm{argmax}_{x':||x'-x||_p \leq \epsilon}\, \mathcal{J}(x', y^{\mathrm{true}}) \tag{2}$$

Specifically, FGSM, BIM, MI-FGSM mentioned in Section II.2 belong to this strategy. GenAttack uses of genetic algorithms to generate adversarial examples. Although it also specifies ϵ, GenAttack does not belong to this attack strategy in the strict sense.

Optimized-based minimized perturbation attacks This strategy solves the problem described in (3) by an optimizer to find the minimum perturbation to meet the conditions.

$$x_{\mathrm{adv}} = \mathrm{argmin}_{x':x'\ \mathrm{is\ adversarial}}\, ||x' - x||_p \tag{3}$$

DeepFool, C&W, and HopSkipJumpAttack [17] mentioned in Section II.2 belong to this attack strategy.

2. Adversarial attack algorithm

Adversarial attack algorithms can be divided into white-box attacks and black-box attacks. FGSM (fast gradient sign method), the most classic white-box attack algorithm, generates adversarial examples by the fixed ϵ via linearizing the loss function in the input space. BIM (basic iterative method), based on FGSM, takes the smaller gradient steps iteratively to get the more accurate attack. PGD (projected gradient descent) extends BIM by starting with a random point and searching for the most suitable gradient direction to attack after multiple iterations. MI-FGSM (momentum iterative method) integrates momentum into BIM to stabilize update gradient directions and jump out of the local optimum.

DeepFool is an optimization-based white-box attack method. The adversarial samples obtained by this attack are often the optimal adversarial samples. C&W (Carlini&Wagner's method) is a more powerful white-box attack in the form of Lagrange with the Adam optimizer to generate adversarial examples.

Boundary attack is the first decision-based black-box attack by random access at the decision boundary of the target model. HopSkipJumpAttack [17] upgrades boundary attack for fewer queries. GenAttack uses a genetic algorithm (Genetic search) to search for adversarial examples with a fixed ϵ.

This paper selects HopSkipJumpAttack and GenAttack algorithms for our subsequent experiments.

III. Evaluation Methodology

We describe the concrete process of robustness evaluation into two steps. One is the attack process: the adversarial attack algorithm generates a corresponding adversarial example after multiple iterations starting from an original image. The other is the evaluation progress: The scaling algorithm scales the generated adversarial examples and inputs the scaled adversarial example to the same classifier. The classifier often classifies adversarial examples and the scaled adversarial examples into different categories.

Fig.1 simply shows the vulnerability of adversarial samples to image scaling.

1. Evaluation metrics

1) Adversarial attack

We adopt the definitions of adversarial attacks in [15].

Accuracy. Given an adversarial attack $\mathcal{A}_{\epsilon,p}$, which generates adversarial examples with perturbation budget ϵ under $\mathcal{L}_p$-norm, when $\mathcal{A}_{\epsilon,p}$ attacks classifier $\mathcal{C}$ on the dataset $< x_i, y_i^{\mathrm{true}} >$ $(i \in [1, N])$, the accuracy of the classifier is defined as follows.

$$ACC(\mathcal{C}, \mathcal{A}_{\epsilon,p}) = \frac{1}{N}\sum_{i=1}^{N} 1(\mathcal{C}(\mathcal{A}_{\epsilon,p}(x_i)) = y_i^{\mathrm{true}}) \tag{4}$$

where $1(\cdot)$ is the indicator function.

Fooling rate. Given a non-targeted adversarial attack $\mathcal{A}_{\epsilon,p}$, the fooling rate of attacking classifier $\mathcal{C}$ is calculated as follows.

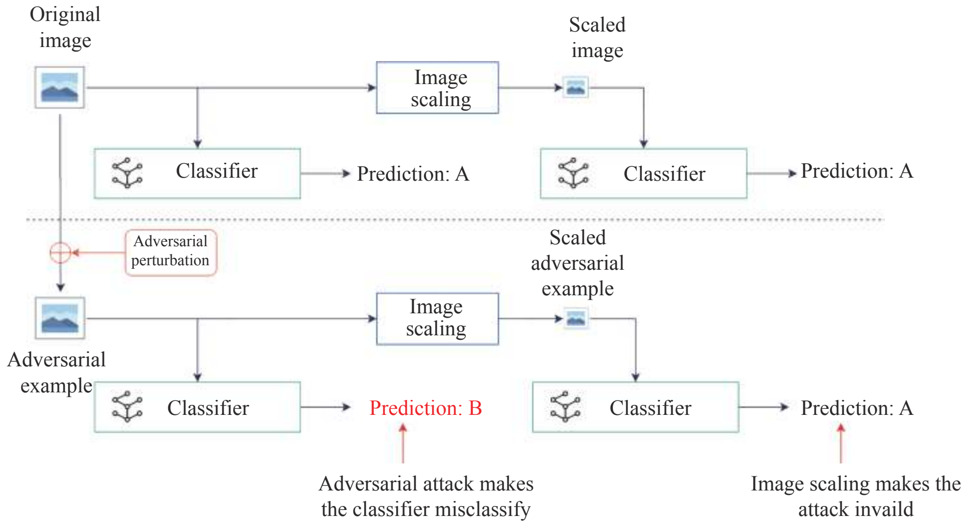


Fig. 1. Vulnerability of adversarial examples to image scaling.

$$FR(\mathcal{A}_{\epsilon,p},\mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} 1(\mathcal{C}(x_i) = y_i^{\text{true}}) \wedge (\mathcal{C}(\mathcal{A}_{\epsilon,p}(x_i)) \neq y_i^{\text{true}})) \qquad (5)$$

while the fooling rate of the targeted attack is defined as follows:

$$FR(\mathcal{A}_{\epsilon,p},\mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} 1(\mathcal{C}(\mathcal{A}_{\epsilon,p}(x_i)) = y*_i) \qquad (6)$$

where $y*$ is the target class.

For DeepFool, C&W, and HopSkipJumpAttack used in this paper, we remove the adversarial samples generated by such attacks with L_p distances exceeding the ϵ limit, which is:

$$\begin{aligned} &x_{\text{adv}} = A_{\epsilon,p}(x) \\ &\text{if } ||x_{\text{adv}} - x||_p > \ \epsilon, \text{ then let } x_{\text{adv}} = x \end{aligned} \qquad (7)$$

2) Image scaling definition

We define the five standard scaling algorithms and their image scaling ratios.

Image scaling algorithms. Let $\sigma_i (i \in [1, 2, 3,4, 5])$ denote the Nearest neighbour algorithm, Bilinear algorithm, Bicubic algorithm, Area algorithm, and Lanczos algorithm, respectively.

Image scaling ratios. Let $\zeta_j = 2^j (j \in [-1, 0, 1])$ denote the image scaling ratio.

According to the above definition, given an image x, an image scaling process $\mathcal{S}$ is described as follows:

$$\mathcal{S}_{\sigma_i,\zeta_j}(x) = \sigma_i(x \cdot \zeta_j) = \sigma_i(x \cdot 2^j) \qquad (8)$$

which denotes the process of scaling original image x by image scaling algorithm σ_i by ζ_j times to obtain the scaled image. In this paper, the original image targeted by the image scaling process is the adversarial sample x_{adv}.

We define the fooling rate of the scaled adversarial examples attacking target classifier to standardize the problem of decline in fooling rate of attacks brought by the image scaling. The fooling rate of $A_{\epsilon,p}$ under $\mathcal{S}_{\sigma_i,\zeta_j}$ is as described in (9) (non-targeted attack):

$$R(\mathcal{A}_{\epsilon,p},\mathcal{S}_{\sigma_i,\zeta_j},\mathcal{C}) = \frac{1}{N}\sum_{k=1}^{N} 1(\mathcal{C}(\mathcal{S}_{\sigma_i,\zeta_j}(x_k)) = y_k^{\text{true}}) \wedge (\mathcal{C}(\mathcal{S}_{\sigma_i,\zeta_j}(\mathcal{A}_{\epsilon,p}(x_k))) \neq y_k^{\text{true}})) \qquad (9)$$

2. Model and algorithms

1) Attack algorithms

ImageNette[*1], the subset of ImageNet [18], created by fast.ai[*2] is selected as the image dataset. We use this dataset according to the scheme described below:

Model training. We use the training set of ImageNette for training models and use the validation set to evaluate the accuracy of the trained models.

Adversarial attack and image scaling. We choose the validation set of ImageNette and select 100 pictures from these ten categories for evaluating in the subsequent experiments.

Four different architectures models (ResNet-18 [19], VGG-11, MobileNet-v2, and Inception-v3) are selected as the advanced image classification models. Through transfer training, we retrain the last layer (full connection layer) of these four DNNs by the dataset mentioned above. The output is ten categories to match

[*1] ImageNette is open source at https://github.com/fastai/imagenette

[*2] fast.ai is the first deep learning library to provide a unified interface for all the most commonly used deep learning applications for vision, text, tabular data, time series, and collaborative filtering. Official website: https://www.fast.ai/

the experiments.

Based on the threat models established in Section Ⅱ, we select seven attack scenarios shown in Table 1. Three attack algorithms of FGSM, BIM and MI-FGSM are regarded as one type: white-box attacks under the L_∞ norm. Two attack algorithms of Deepfool and C&W are regarded as one type: white-box attacks under the L_2 norm. Two attack algorithms of HopSkipJumpAttack and GenAttack are regarded as one type: black-box attacks under the L_∞ norm.

Table 1. The attack methods implemented in our experiments

Attack method	Knowledge	Goals	Capability	Distance
FGSM	White-box	Non-targeted	ϵ constrained	L_∞
BIM	White-box	Non-targeted	ϵ constrained	L_∞
MI-FGSM	White-box	Non-targeted	ϵ constrained	L_∞
DeepFool	White-box	Non-targeted	Optimized-based	L_2
C&W	White-box	Non-targeted	Optimized-based	L_2
HopSkipJumpAttack	Decision-based	Non-targeted	Optimized-based	L_∞
GenAttack	Decision-based	Targeted	ϵ constrained	L_∞

The gradient-based fixed ϵ attacks can directly set the adversarial sample perturbation distance through the setting of ϵ, while the optimization-based minimization perturbation attacks cannot directly set. To meet the perturbation budget setting, we make minor adjustments to the three algorithms, DeepFool, C&W, and HopSkipJumpAttack.

2) Image scaling process

The image scaling process as described in Algorithm 1. Among them, the algorithm's input is the image dataset $< x_k, y_k^{\text{true}} >$, which is the set of adversarial examples obtained by the adversarial attack $\mathcal{A}_{\epsilon,p}$, and the output is the adversarial examples set $< x_{k(\text{scaled})}^{\text{adv}} >$ after image scaling.

Algorithm 1: Using different image scaling algorithms to scale adversarial examples at different scaling ratios

```
Input: Evaluation dataset: $\langle x_k, y_k^{\text{true}} \rangle$
Output: Scaled adversarial examples set: $\langle x_{k(\text{scaled})}^{\text{adv}} \rangle$
Adversarial attack: $\mathcal{A}_{\epsilon,\ p}$. Image scaling: $\mathcal{S}_{\sigma_i,\ \varsigma_j}$;
Parameters: $i \in [1, 2, \dots, 5],\ j \in [-1, 0, 1],\ k \in [1, N]$;
for $k \leftarrow 1$ to $N$ do
    $x_k^{\text{adv}} = \mathcal{A}_{\epsilon,\ p}(x_k)$ //Using $\mathcal{A}_{\epsilon,p}$ to generate adversarial examples
    for $i \leftarrow 1$ to 5 do
        for $j$ in $[-1, 0, 1]$ do
            if $j == 0$ then
                $x_{k(\text{scaled})}^{\text{adv}} = x_k^{\text{adv}}$ //Control group
            else
                $x_{k(\text{scaled})}^{\text{adv}} = \mathcal{S}_{\sigma_i,\ \varsigma_j}(x_k^{\text{adv}})$
                    // Using $\mathcal{S}_{\sigma_i,\varsigma_j}$ to scale image
                    image
            end
        end
    end
end
```

As mentioned in Section III, we use 11 image scaling processes $\mathcal{S}_{\sigma_i,\varsigma_j}(i \in [1, ..., 5], j \in [-1, 0, 1])$ to scale 112 adversarial examples generated by seven adversarial attack algorithms that attack four different image classification models under four levels of perturbation budgets and then send the scaled adversarial examples to corresponding image classification models again for classification.

IV. Evaluation Results and Analysis

1. Experimental setup

We run our experiments on the GPU Server equipped with Intel E5-2678×2, NVIDIA RTX 2080Ti×3, and 32 GB of memory. Before we implement the experiments, we describe the set of data corresponding to the experiments.

In this experiment, each model was trained iteratively for 10 epochs. After 10 iterations, the accuracy of ResNet-18, VGG-11, MobileNet-v2 and Inception-v3 is 96.67%, 97.18%, 97.22% and 98.04%, respectively.

Perturbation budget setting: We implement perturbation budgets on the above three types of adversarial attacks according to the four levels of perturbation budgets shown in Table 2.

Hyperparameters setting: We set the hyperparameters of adversarial attacks based on the original papers shown in Table 3.

In this experiment, we all obtained the fooling rates of 1232 group adversarial attacks.

2. Results on adversarial attacks

Before the image scaling process, we conduct experiments on the fooling rates of various adversarial attacks. The adversarial attacks attack four different image classification models sequentially. After seven adversarial attacks under four levels of perturbation budgets, the accuracies of the four classifiers have decreased significantly. The accuracy of image classifica-

Table 2. Four levels of perturbation budgets for three different types of adversarial attacks

Knowledge attack methods		Perturbation budget				
		L_p norm	Level1	Level2	Level3	Level4
White-box	FGSM,MI-FGSM,BIM	L_∞	4/255	8/255	16/255	32/255
	DeepFool,CW	L_2	2	4	6	8
Black-box	HSJA,GA	L_∞	64/255	72/255	80/255	88/255

Table 3. Hyperparameters set in adversarial attacks

Attack method	Hyperparameters	Setting value
FGSM,BIM,MI-FGSM	default	default
DeepFool	steps	100
C&W	initial_const	1000
	learning_rate	0.07
	max_iterations	1000
HSJA	iterations	64
	initial_num_evals	10
	max_num_evals	1000
	gamma	0.1
GenAttack	generations	1000

tion models under four levels of perturbation budgets is shown in Table 4.

Under the different perturbation budgets, the unscaled adversarial examples successfully attack the image classification models. Models' accuracies decrease to varying degrees, most of which decrease from more than 90% to 10%. Partial white-box attacks like MI-FGSM and BIM even drop the model accuracy to 0%. Additionally, we find that all models' accuracy decreases with the increase of perturbation budget.

Among five white-box attacks, the accuracy of models after adversarial attacks reduces significantly under the Level1 perturbation budget due to their mature attack algorithms and accessible attack environments. The accuracy of the target model thus doesn't change sharply. For the remaining two black-box attacks, the accuracy of the target model reflects a more apparent downward trend.

Table 4. The accuracy of different classifiers under the different perturbation budgets

(a) The accuracy of ResNet-18

ResNet-18 (base: 98%)

Attack	**Level1**	**Level2**	**Level3**	**Level4**
FGSM	12%	11%	11%	9%
BIM	2%	0%	0%	0%
MIM	0%	0%	0%	0%
DF	3%	4%	4%	4%
C&W	31%	8%	0%	0%
HSJA	17%	11%	10%	7%
GA	11%	5%	5%	0%

(b) The accuracy of VGG-11

VGG-11 (base: 97%)

Attack	**Level1**	**Level2**	**Level3**	**Level4**
FGSM	20%	17%	16%	14%
BIM	2%	1%	0%	0%
MIM	2%	2%	0%	0%
DF	4%	4%	6%	12%
C&W	34%	9%	1%	0%
HSJA	9%	9%	7%	3%
GA	61%	52%	45%	41%

(c) The accuracy of Inception-v3

Inception-v3 (base: 93%)

Attack	**Level1**	**Level2**	**Level3**	**Level4**
FGSM	53%	48%	36%	16%
BIM	8%	4%	0%	0%
MIM	0%	1%	0%	0%
DF	10%	5%	6%	5%
C&W	62%	40%	27%	11%
HSJA	11%	6%	4%	3%
GA	17%	11%	7%	3%

(d) The accuracy of MobileNet-v2

MobileNet-v2 (base: 96%)

Attack	**Level1**	**Level2**	**Level3**	**Level4**
FGSM	17%	17%	12%	7%
BIM	2%	0%	0%	0%
MIM	0%	0%	0%	0%
DF	2%	2%	3%	4%
C&W	40%	13%	5%	2%
HSJA	11%	13%	5%	2%
GA	14%	12%	8%	10%

3. Evaluation results and analysis on perturbation budgets

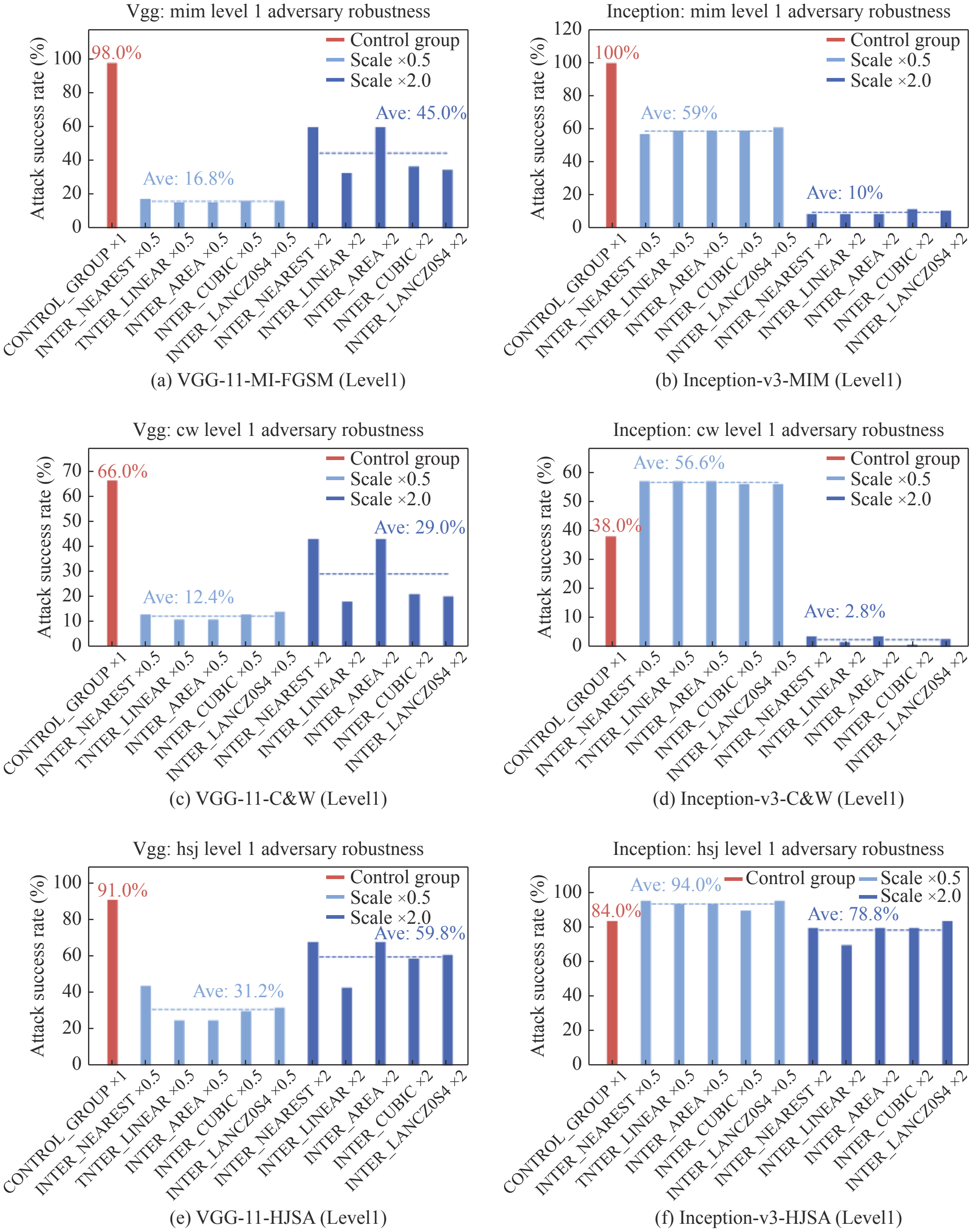
We study the effect of image scaling on the robustness of adversarial examples generated by different adversarial attacks under multi-level perturbation budgets attacking the same image classification model. The ResNet-18 shows relatively stable performance and standard data rules throughout the experiment. Thus in this evaluation, we focus on the analysis related to ResNet-18.

The experiment obtain seven robustness evaluation data of the adversarial examples generated by adversarial attacks under four levels perturbation budgets attacking ResNet-18. The fooling rate of most adversarial attacks is significantly improved with the increase of the perturbation budget. No matter it is a black-box attack or a white-box attack, no matter what kind of image scaling process the adversarial example is subjected to, the fooling rate increases with the increase of perturbation budget. The data relating to the other image classification models also show similar trends. Therefore, we think that we can construct adversarial examples that are more robust against various image scaling processes by increasing the level of the perturbation budget.

4. Evaluation results and analysis on image scaling processes

We research how the different image scaling algorithms σ_i , and various image scaling ratios ζ_j affect the robustness of adversarial examples generated by the same adversarial attack under the same perturbation budget attacking the same classifier. In this evaluation, we chose to analyze the data related to VGG-11 and Inception-v3, according to the analysis in Section 4. Image scaling processes of MI-FGSM, C&W, and HopSkipJumpAttack under the Level1 perturbation budget has the most significant possible impact on the fooling rate of adversarial examples. As shown on the left of Fig.2 , we obtain some findings (MobileNet-v2 and ResNet-18 are similar to VGG-11).

For the image scaling algorithm σ_i, when the scaling ratio is ζ_{-1}, different σ_i have a similar effect on the fooling rate, but when the scaling ratio is ζ_1, σ_1 and σ_3, compared to the other three algorithms, can not have an approximate level of impact on the fooling rate (σ_1



(a) VGG-11-MI-FGSM (Level1)

(b) Inception-v3-MIM (Level1)

(c) VGG-11-C&W (Level1)

(d) Inception-v3-C&W (Level1)

(e) VGG-11-HJSA (Level1)

(f) Inception-v3-HJSA (Level1)

Fig. 2. Different image scaling algorithms and image scaling ratios affect the robustness of adversarial examples for VGG-11 (left: (a), (c), and (e)) and Inception-v3 (right: (b), (d), and (f)).

and σ_3 causes the fooling rate to decrease to a less extent than σ_2, σ_4,σ_5).

For the image scaling ratio ζ_j, ζ_{-1} (reduced to 0.5 times) generally has a greater impact on the fooling rate of adversarial examples than that of ζ_1 (enlarged to 2 times). Under the same conditions, $\mathcal{S}_{\sigma_i,\zeta_{-1}}$ causes the decline degree of the fooling rate to be 20%–30% higher than that of $\mathcal{S}_{\sigma_i,\zeta_1}$ on average.

While as shown in the right of Fig.2, we have some different findings. For Inception-v3, the impact of image scaling algorithm σ_i on the fooling rate of adversarial examples is not much different. Under the same conditions, $\mathcal{S}_{\sigma_i,\zeta_{-1}}$ leads to a decreasing degree of the fooling rate of adversarial examples 30%–60% lower than that of $\mathcal{S}_{\sigma_i,\zeta_1}$.

Based on the above analysis, we find that the effect of the image scaling algorithm on the robustness of adversarial examples is not very obvious. The five scaling algorithms can effectively reduce the fooling rate, but the impact of the image scaling ratio on the robustness of adversarial examples is quite different. For the three models of VGG-11, ResNet-18, and MobileNet-v2, the fooling rates show a similar law: "Reducing the size of an adversarial example to 0.5 times" makes the fooling rate drop more generally than "magnifying the adversarial example size to 2 times". However, Inception-v3 is the opposite.

V. Conclusions

In this paper, we designed groups of experiments in different situations, including other models, different adversarial attack algorithms, different perturbation budgets, and various image scaling processes to systematically evaluate the impact of image scaling on the robustness of adversarial examples. Based on the experiment results, we got some interesting observations. Our work will be beneficial to design more robust adversarial attack algorithms against image scaling. It can also provide a solid foundation for developing more efficient methods to defend against adversarial attacks.

References

[1] W. Li, W. Meng, Z. Tan, and Y. Xiang, "Design of multi-view based email classification for IOT systems via semi-supervised learning," *Journal of Network and Computer Applications*, vol.128, pp.56–63, 2019.

[2] F. Ullah, H. Naeem, S. Jabbar, S. Khalid, M. A. Latif, *et al.*, "Cyber security threats detection in Internet of things using deep learning approach," *IEEE Access* , vol.7, pp.124379–124389, 2019.

[3] F. Al-Turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning," *IEEE Access*, vol.7, pp.115749–115759, 2019.

[4] Z. Lv, W. Mazurczyk, S. Wendzel, and H. Song, "Guest editorial: Recent advances in cyber-physical security in industrial environments," *IEEE Transactions on Industrial Informatics*, vol.15, no.12, pp.6468–6471, 2019.

[5] M. Daraghmeh, I. Al Ridhawi, M. Aloqaily, Y. Jararweh, and A. Agarwal. " A power management approach to reduce energy consumption for edge computing servers," in *Proceedings of 2019 Forth International Conference on Fog and Mobile Edge Computing, Rome*, Italy, pp.259–264, 2019.

[6] Shuming Qiu, Ding Wang, Guoai Xu, and Saru Kumari, "Practical and provably secure three-factor authentication protocol based on extended chaotic-maps for mobile lightweight devices," *IEEE Transactions on Dependable and Secure Computing*, vol.19, no.2, pp.1338–1351, 2020.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, *et al.*, "Intriguing properties of neural networks," *arXiv preprint*, arXiv:1312.6199, 2013.

[8] Z. Gu, Y. Xie, W. Hu, L. Yin, Y. Han, and Z. Tian, "Marginal attacks of generating adversarial examples for spam filtering," *Chinese Journal of Electronics* , vol.30, no.4, pp.595–602, 2021.

[9] C. Wang, D. Wang, Y. Tu, *et al.*, "Understanding node capture attacks in user authentication schemes for wireless sensor networks," *IEEE Transactions on Dependable and Secure Computing*, vol.19, no.1, pp.507–523, 2022.

[10] H. Lyu, Y. Tan, Y. Xue, Y. Wang, and J. Xue, "A CMA-ES-based adversarial attack against black-box object detectors," *Chinese Journal of Electronics* , vol.30, no.3, pp.406–412, 2021.

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

[12] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, " Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.2725–2734, 2019.

[13] D. Wang, X. Zhang, Z. Zhang, and P. Wang, "Understanding security failures of multi-factor authentication schemes for multi-server environments," *Computers & Security*, vol.88, article no.101619, 2020.

[14] T. Yu, S. Wang, C. Zhang, Z. Wang, Y. Li, and X.Yu, "Targeted adversarial examples generating method based on cVAE in black box settings," *Chinese Journal of Electronics*, vol.30, no.5, pp.866–875, 2021.

[15] Y. Dong, Q. Fu, X. Yang, T. Pang, *et al.*, "Benchmarking adversarial robustness on image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.318–328, 2020.

[16] C. Wang, D. Wang, G. Xu, and D. He, "Efficient privacy-preserving user authentication scheme with forward secrecy for industry 4.0," *Science China Information Sciences*, vol.65, no.1, article no.112301, 2022.

[17] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *Proceedings of 2020 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, pp.1277–1294, 2020.

[18] J. Deng, W. Dong, R. Socher, L. Li, K. Li and F. Li, "Im-

agenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp.248–255, 2009.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.770–778, 2016.

ZHENG Jiamin is a Ph.D. and Associate Professor in School of E-Business and Logistics at Beijing Technology and Business University. His main research interests include information security, and artificial intelligence.
(Email: zhengjm@btbu.edu.cn)

ZHANG Yaoyuan received the B.E. degree in computer science of technology from Beijing Institute of Technology, in 2017, where she is currently pursuing the Ph.D. degree. Her research interests include artificial intelligence security. Recently, her research focus has been in the area of computer vision adversarial attack.
(Email: yaoyuan@bit.edu.cn)

LI Yuanzhang received the B.S., M.S., and Ph.D. degrees in software and theory of computer from Beijing Institute of Technology in 2001, 2004, and 2015, respectively. He has been an Associate Professor with Beijing Institute of Technology. His research interests include mobile computing and information security.
(Email: popular@bit.edu.cn)

WU Shangbo graduated from the School of Computer Science and Technology, Beijing Institute of Technology, in 2020. He received the M.S. degree from University of Glasgow in 2022. His main research interest lies in the areas of semantic black-box adversarial attacks for both classifiers and object detectors. (Email: wu@bit.edu.cn)

YU Xiao (corresponding author) is a Ph.D., Associate Professor and Master Supervisor in Department of Computer Science and Technology, Shandong University of Technology. His current research interests include artificial intelligence security and embedded system.
(Email: yuxiao8907118@163.com)