Asia Pacific Academy of Science Pte. Ltd.

Metaverse 2024, 5(2), 2764.
https://doi.org/10.54517/m.v5i2.2764

Article

Efficient transfer attacks via enhancing perturbation robustness

Chengzhi Zhong[1,†], Jipeng Hu[1,†], Mengda Xie[1], Meie Fang[2,*]

[1] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510700, China
[2] School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China
*** Corresponding author:** Meie Fang, fme@gzhu.edu.cn
[†] The two authors contributed equally.

CITATION

Zhong C, Hu J, Xie M, Fang M. Efficient transfer attacks via enhancing perturbation robustness. Metaverse. 2024; 5(2): 2764. https://doi.org/10.54517/m.v5i2.2764

ARTICLE INFO

Received: 5 June 2024
Accepted: 5 July 2024
Available online: 25 October 2024

Abstract: With the rapid development of deep learning technology, artificial intelligence (AI) has found wide applications in diverse domains such as image classification, text processing, and autonomous driving. However, the increasing prevalence of security issues cannot be ignored. Studies have shown that deep neural network models face security risks due to adversarial sample attacks. These attacks involve adding imperceptible perturbations to deceive the model's classification results, exposing vulnerabilities in deep learning model applications. While transfer attack methods offer practicality in real-world scenarios, their current performance in black-box attacks is limited. In this study, we propose a method that combines an attention mechanism and a frequency domain transformation to enhance the robustness of adversarial perturbations, thereby improving the performance of transfer attacks in black-box attack scenarios of deep learning models. Specifically, we introduce the CBAM-ResNet50 enhancement model based on attention mechanisms into transfer attacks, enhancing the model's ability to identify important image regions. By adding perturbations to these attention-concentrated regions, adversary perturbation robustness is improved. Furthermore, we introduce a method for randomly transforming image enhancement in the frequency domain, which increases the diversity and robustness of adversarial perturbation by distributing perturbations across edges and textures. Experimental results demonstrate that our proposed method, considering both human perceptibility and computational cost, achieves a maximum black-box transfer attack success rate of 60.05%, surpassing the 49.65% success rate achieved by the NI-FGSM method across three models. The average success rate of the five methods exceeds an improvement of 6 percentage points in black-box attacks.

Keywords: deep learning; adversarial samples; transferability; black-box attacks; attention mechanisms; frequency domain transformations

1. Introduction

As deep learning technology continues to gain traction across different fields, the revelation of adversarial samples [1] which are maliciously designed to attack deep learning models raises wide concern on potential practical applications of them. Among different attacks, black-box attacks [2] are more realistic and challenging than white-box attacks, where attackers have full knowledge of the victim model. In black-box attacks, attackers lack access to the victim model and can only observe its classification outputs, sometimes without confidence scores. Therefore, black-box attacks are more relevant and practical for real-world scenarios.

Black-box attacks can be traced back to late 1990s, when researchers started investigating how to attack classifiers based on traditional machine learning algorithms [3]. In this attack scenario, the attackers lack access to the internal model details, limiting them to perform the attack solely by utilizing the outputs of the model

after several queries. Hence, this attack is referred to as a black-box attack. The initial adversarial attack methods targeted linear classifiers such as support vector machines [4] and logistic regression [5]. These attacks used mathematical techniques to generate misleading samples for the classifier, causing it to output incorrect results. Over time, researchers delved into the exploration of attacking increasingly intricate models, including deep neural networks.

Regarding adversarial sample generation, researchers have proposed several improved algorithms, including those based on model uncertainty [6] and reinforcement learning [7]. In terms of attack evaluation, some metrics have been applied to assess the effectiveness of black-box adversarial attacks, such as success rate, failure rate, and sustained attack count. To defend such adversarial attacks, researchers have proposed various defense methods, such as adversarial training [8], randomization defense [9], and ensemble defense [10], to enhance model robustness.

Transfer attacks [11] are a variant of black-box adversarial attacks, where attackers exploit successful adversarial samples generated against one target model to attack another target model. Research on transfer attacks primarily focuses on the transferability of adversarial samples and the transferability of attacks. Adversarial sample transferability examines whether adversarial samples generated for one target model can also remain adversarial on another. Attack transferability investigates whether attackers can use a set of adversarial samples to attack multiple target models. Scholars have proposed improved algorithms to enhance the transferability of adversarial samples in transfer attacks, including transfer learning-based attack algorithms [12–14], feature alignment-based attack algorithms [15,16], and adaptive attack algorithms [17–19].

In conclusion, black-box attacks and transfer attacks are prominent research directions in the field of deep learning. As deep learning continues to advance in practical applications, these attack algorithms will undergo further improvements and optimizations to enhance model robustness.

In this paper, we focus on image classification tasks in deep neural network models. The proposed attack method combines an attention mechanism and a spatial-to-frequency domain transformation to enhance the robustness of perturbations and improve the performance of transfer attacks.

The main contributions of this paper can be summarized as follows:

1) We introduced an attention mechanism into adversarial sample transferability attacks. Specifically, we build a CBAM-ResNet50-based model on top of ResNet50 as the victim model. This model enhances the attention to the robust regions in the image, thereby reducing the overfitting problem of the generated adversarial perturbations in the victim model.
2) We proposed a method introduces random transformations in the frequency domain for image enhancement during the generation of adversarial samples. Frequency domain random transformations enhance the diversity of adversarial perturbations and forces attack algorithms to seek more robust perturbations, thereby improving their transferability.
3) By combining the CBAM attention mechanism with frequency domain image transformation, empirical results show that the two approaches can enhance the transferability of adversarial samples. Furthermore, to optimize the performance

of the proposed adversarial algorithm combined with the CBAM attention mechanism and frequency domain image transformation, this paper explores the specific impact of changes in factors related to adversarial sample transferability, such as the maximum perturbation value, iteration times, and frequency domain transformation probability P.

2. Background knowledge

Adversarial attacks can be divided in to two types, white-box attack and black-box attack. White-box attack means the attackers have full knowledge of the victim model, and black-box attack refers to the scenario where the victim model seems like a black box to the attackers, and only the output is available in common settings.

Black-box attacks can be roughly categorized into three types: 1) surrogate model attacks [20–22], where a comparable neural network model is trained using the same training data as the target model, and then white-box attacks are conducted on the surrogate model to indirectly target the original model. 2) attack based on decision boundary [22,23], where the parameters and architecture of the neural network are not accessible, but the model's output can be queried repeatedly to iteratively generate adversarial samples. The main idea is to initialize the sample as an image of the target class and iteratively approach the original image near the decision boundary based on the query results. By continuously iterating the adversarial samples while keeping the classification result as the target class, the goal is to get as close as possible to the original sample, and output final adversarial samples when the predefined stopping criteria is met. 3) attack based on adversarial sample transferability [24,25], where research has shown that adversarial samples generated for one model have a certain probability of causing misclassification when applied to another model. Therefore, powerful black-box attacks can be achieved by enhancing the transferability of adversarial samples. Black-box attacks typically involve making numerous queries to the neural network model in the first two types. However, in real-world situations, models can easily identify these unusual and frequent queries. Therefore, exploring black-box attacks that rely on the transferability of adversarial samples holds more practical promise for application.

FGSM (Fast Gradient Sign Method) class methods are classic gradient based methods for generating adversarial samples. This article is also based on several FGSM class methods, so this section introduces several FGSM class methods. Define x as the original sample, x^{adv} as the generated adversarial sample, and ε as the perturbation size that controls the attack intensity of the adversarial sample. $J(\theta, x, y)$ is the loss function of the target model, $\nabla_x J(\theta, x, y)$ is the gradient of the loss function to the input sample, reflecting the sensitivity of the model to the sample. $sign$ represents a sign function that takes a gradient.

The FGSM [26] attack is based on the concept of introducing a perturbation to the input image in order to create an adversary sample. This perturbation is computed by maximizing the gradient of the loss function with respect to the input data. By doing so, the attackers can create a significant perturbation that causes the neural network to generate different outputs for the original image and the adversary sample. The calculation Equation of FGSM attack algorithm is shown in Equation (1):

$$x^{adv} = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, y)). \quad (1)$$

I-FGSM (Iterative Fast Gradient Sign Method) [27] is an iterative attack algorithm based on FGSM. The following is the specific Equation expression of I-FGSM attack:

$$X_0^{adv} = X \quad (2)$$

$$X_{N+1}^{adv} = Clip_{X,\varepsilon}\{X_N^{adv} + \alpha sign(\nabla_X J(\theta, X_N^{adv}, y_{true}))\} \quad (3)$$

In Equation (2), X is the original picture, and in Equation (3), X_N^{adv} is the adversary sample processed by FGSM algorithm n times, $\mathrm{Clip}_{X,\varepsilon}(A_{i,j})$ cuts each element $A_{i,j}$ in the input vector to $[X_{i,j-\varepsilon}, X_{i,j+\varepsilon}]$. sign $(\nabla_X \mathrm{J}\ (\theta, \mathrm{X}_N^{\mathrm{adv}}, \mathrm{y_{true}})$ and the corresponding calculation amount in the original FGSM are the same, α represents the amplitude of image pixel update in each iteration.

The MI-FGSM (Momentum Iterative Fast Gradient Sign Method) [28] attack method is an improved method based on the BIM attack method. The momentum term is introduced to memorize the gradient of each iteration, so as to increase the directional stability of gradient update. Specifically, the MI-FGSM attack method makes a weighted average of the previous gradient direction and the current gradient direction at each iteration, that is:

$$g_t = \alpha g_{t-1} + \frac{\nabla_x J(\theta, x, y)}{\| \nabla_x J(\theta, x, y) \|_1} \quad (4)$$

In Equation (4), g_t is the gradient of iteration t, α is the momentum factor, $\nabla_x J(\theta, x, y)$ represents the gradient of the loss function of the model with respect to input x under the current input x and label y. In each iteration, the MI-FGSM attack method uses the accumulated gradient direction to update the current adversary sample x, as shown in Equation (5):

$$x_{t+1}^{'} = x_t^{'} + \epsilon sign\ (g_t) \quad (5)$$

NI-FGSM (non-local iterative fast gradient sign method) [29] is an algorithm for generating adversary samples based on iterative optimization. Its main idea is to use a better optimization algorithm to improve the transferability. Specifically, NAG-Nesterov accelerated gradient is used to optimize gradient based iterative attacks. NAG can be regarded as an improved version of momentum-based optimization. The specific Equation is shown as follows:

$$\begin{aligned} v_{t+1} &= \mu \cdot v_t + \nabla_{\theta_t} \mathcal{J}(\theta_t - \alpha \cdot \mu \cdot v_t) \\ \theta_{t+1} &= \theta_t - \alpha \cdot v_{t+1} \end{aligned} \quad (6)$$

NI-FGSM can escape from local optima more quickly, thus greatly enhancing attack performance and portability, specifically realizing:

$$\begin{aligned} x_t^{nes} &= x_t^{adv} + \alpha \cdot \mu \cdot g_t, \\ g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x_t^{nes}, y^{true})}{\| \nabla_x J(x_t^{nes}, y^{true}) \|_1} \\ x_{t+1}^{adv} &= Clim_x^{\epsilon}\{x_t^{adv} + \alpha \cdot sign\ (g_{t+1})\}, \end{aligned} \quad (7)$$

In Equation (7), g_t represents the cumulative gradient at iteration t, μ represents attenuation coefficient of g_t.

PGD (Projected Gradient descent) [30] is an iterative attack method. Compared with FGSM and FGM, PGD carries out multiple iterations. Each iteration takes small steps to project the perturbation to the specified range to achieve the purpose of attack.

$$g_t = \nabla X_t\big(L(f_\theta(X_t), y)\big) \tag{8}$$

In Equation (8), g_t represents the gradient of the loss at iteration t with respect to the input at iteration.

$$X_{t+1} = \prod_{X+S} (X_t + \varepsilon(\frac{g_t}{||\ g_t\ ||})) \tag{9}$$

In Equation (9), the input of the $t+1$ time is obtained through the input of the t time and its corresponding gradient information. $\prod_{X+S}$ is to map the input back to the specified range *S* when the perturbation exceeds the specified range.

With regard to the researches on attack transferability, there are roughly 2 types of transferability-enhancing methods: 1) optimization-based methods [31–34] that directly optimize for the adversarial perturbations based on one or more surrogate models at inference time, without introducing additional generative models, and 2) generation-based methods [35,36] that introduce generative models dedicated for adversary synthesis. The latter methods take an alternative approach by directly synthesizing the adversarial example (or the adversarial perturbation) with generative models. Generation-based attacks comprise two stages: Training and the attack stages. These works have gained a lot success and shed light on new researches regarding attack transferability.

3. Method

We introduce a novel approach to address the issue of limited transferability in black-box transfer attacks. It focuses on enhancing the performance of transfer attack by employing an attention mechanism and frequency domain transformation. The proposed method operates at both the model and image preprocessing levels within the black-box attack scenario of deep learning models. The following describes the methods to enhance the performance of transfer attack by introducing attention mechanism and random transformation in frequency domain into the white-box model.

3.1. Introducing CBAM-ResNet50 model into transferability attack

The attention mechanism [31] has gained widespread usage in deep learning models in recent years, significantly enhancing both accuracy and model robustness. This robustness improvement can be attributed to the attention mechanism's ability to focus on crucial image features. Consequently, the transferability of adversarial samples, referring to their generalization across different models, is typically poor. This poor transferability stems from the inability of added perturbations to effectively target the robustness characteristics of the image, leading to overfitting in individual models. Building upon this notion, this paper introduces a novel approach that leverages the attention mechanism to enhance the performance of transfer attack.

Using attention mechanism to mine local features in the deep neural network and combine them with global features is the principle of attention mechanism to make the model pay more attention to local robust features. This paper applies this principle to the lifting adversary sample transferability method based on attention mechanism to guide the attention to image robust features when adding perturbations, Thus, it focuses on the perturbation of robust features and reduces the perturbation of non-robust features, so that the generated adversary samples have better generalization

performance for different models, and enhance the transferability of adversary samples.

The CBAM (Convolutional Block Attention Module) [32], introduced by Woo et al. In 2018, is a widely adopted module in computer vision. It incorporates a spatial attention mechanism to enhance model performance and robustness. It consists of two components: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). By integrating the CBAM module into the ResNet50 model, we obtain the CBAM-ResNet50 architecture, as illustrated in **Figure 1**.

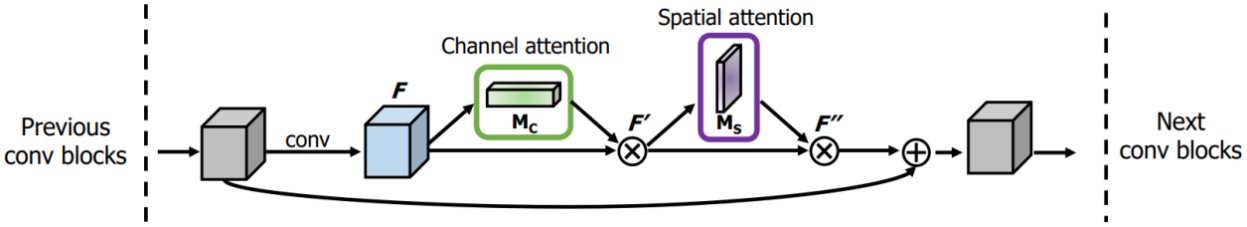


Figure 1. Network structure of CBAM-ResNet50.

In order to compare the effect of models before and after ResNet50 combined with CBAM attention mechanism, we have visualized the thermodynamic diagram. **Figure 2** shows the visualization effect before and after ResNet50 combined with CBAM attention mechanism. It can be concluded from the figure that the thermal map combined with CBAM is more accurate for the image feature area and has a larger range.

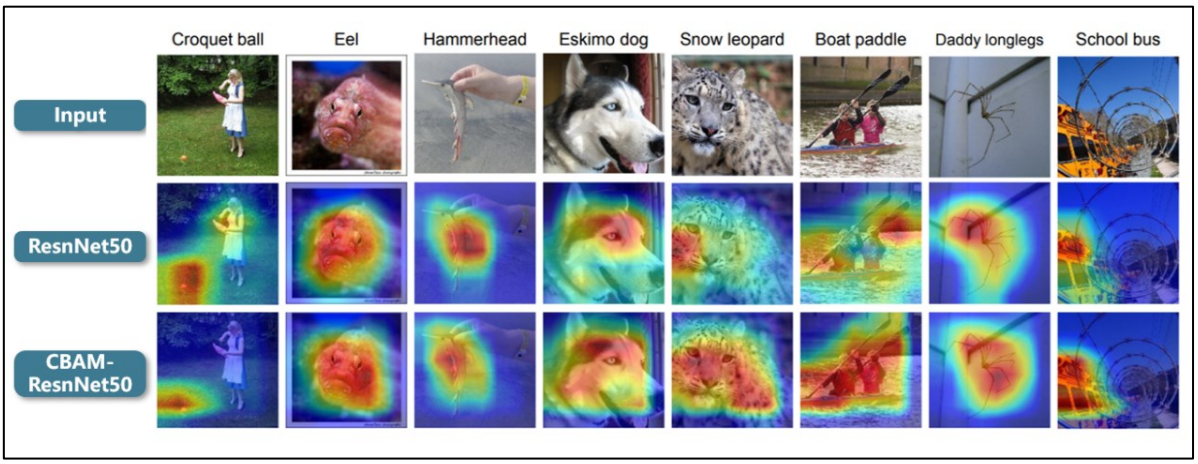


Figure 2. Visual renderings before and after ResNet50 combined with CBAM attention mechanism.

3.2. FDRT enhancement of image random transform frame based on frequency domain to enhance transferability

The generation of adversary samples depends on the original input clean images. When the number of original clean images is small and the types are limited, it is very easy to cause "over fitting" phenomenon, which leads to weak transferability when transferring to other models. In recent years, data enhancement methods such as image random transformation have been introduced into the generation of adversary samples and have been widely used. However, most of the image transformation methods are based on the spatial domain, and they are lack of enough randomness, thus easy to be defended targetedly, and leading to weak transferability. Therefore, this paper

introduces FDRT (Frequency domain random transformation), an image random transformation framework based on frequency domain, into the generation of adversary samples. In each iteration of adversary sample generation, FDRT randomly selects one of the four transformation methods of gaussian blur with equal probability, sharpening, rotation and scaling, and applies it to the image with probability $P = 0.5$. This random transformation increases the diversity and transferability of adversary samples, so that attackers better perform the attack.

3.2.1. Build image space domain and frequency domain conversion framework

Since our frequency domain image random transformation method needs to be converted within the frequency domain and spatial domain, this section first introduces the image space domain and frequency domain transformation methods we use.

1) DCT transformation

Discrete Cosine Transform (DCT) is used in our model. For a gray-scale image $f(x, y)$ of size $\mathrm{N} \times \mathrm{N}$, its DCT transform can be expressed by Equation (10):

$$F(u,v) = \frac{1}{\sqrt{N}} C(u)C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) cos\left[\frac{(2x+1)u\pi}{2N}\right] cos\left[\frac{\left((2y+1)u\pi\right)}{2N}\right] \quad (10)$$

where in, $C(u)$ $C(v)$ is the DCT coefficient. Through DCT transformation, the spatial domain pixels of the image will be converted into coefficients in the frequency domain. These coefficients describe the changes of different frequencies in the image. The larger the coefficient, the higher the frequency appears in the image.

2) IDCT transformation

The inverse transform from frequency domain to space domain is also called inverse discrete cosine transform (IDCT). IDCT is the process of remapping the frequency domain coefficients after DCT transformation back to the original spatial domain. Similar to DCT transform, IDCT transform is also a linear transform. Assuming that a DCT coefficient matrix X with the size of $\mathrm{N} \times \mathrm{N}$ has been obtained, the IDCT transformation Y of this matrix can be calculated by Equation (11):

$$Y_{u,v} = \frac{1}{2N} a(u)a(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} X_{u,v} cos\left[\frac{(2x+1)u\pi}{2N}\right] cos\left[\frac{\left((2y+1)u\pi\right)}{2N}\right] \quad (11)$$

Compared with the Equation of DCT transform, the coefficient matrix of IDCT transform is divided by $2\mathrm{N}$, and a product of $a(u)$ and $a(v)$ is added. This is to ensure the orthogonality and energy conservation of IDCT transform.

3.2.2. Random transformation of frequency domain image to enhance transferability

To enhance the diversity and transferability of adversary samples, we introduce a method that boosts their robustness through random image transformations during each iteration of sample generation. More precisely, the image first undergoes conversion from spatial to frequency domain, and a random selection is made among four transformation methods: Gaussian blur, sharpening, rotation, and scaling. A probability value of $P = 0.5$ (based on experimentation) is then applied to the image. Afterward, the frequency domain is converted back to the spatial domain, and the iteration process is repeated. Through the random combination of these four transformations methods, the diversity and robustness of the adversary samples can be

increased, thus improving their transferability. The following describes the four transformation methods of gaussian blur, sharpening, rotation and scaling in the frequency domain.

Firstly, the Gaussian fuzzy transformation in frequency domain is introduced. Gaussian blur is one of the most basic filtering methods in image processing, which is used to reduce noise and smooth images. In the frequency domain, Gaussian blur can be achieved by low-pass filtering the Fourier transformed image. Gaussian blur can be expressed by Equation (12):

$$H(u,v) = e^{\frac{D^2(u,v)}{2\sigma^2}} \tag{12}$$

$D(u,v)$ represents the distance function in the frequency domain, while σ denotes the standard deviation of the Gaussian kernel.

The second is sharpening transform in frequency domain. Sharpening transformation can enhance the edges and details of the image, and reduce image blur and distortion. In the frequency domain, sharpening can be achieved by high pass filtering the Fourier transformed image in the frequency domain. Sharpening can be expressed by Equation (13):

$$G(u,v) = \big(1 + kH(u,v)\big)F(u,v) \tag{13}$$

where $F(u,v)$ is the frequency domain representation of the original image, k is the sharpening coefficient, and $H(u,v)$ is the sharpening filter.

The third is the frequency domain rotation transform. The rotation in frequency domain can be realized by rotating the image after Fourier transform. The rotation transformation can be expressed by Equation (14):

$$G(u,v) = F\big(u^{'}, v^{'}\big) \tag{14}$$

where (u', v') is the coordinate after rotation transformation, which is calculated by Equation (15):

$$u^{'} = ucos\theta + vsin\theta, v^{'} = -usin\theta + vcos\theta \tag{15}$$

Finally, the frequency domain scaling transform. Scaling can be achieved by interpolating the Fourier transformed image in the frequency domain. Scaling transformation can be expressed by Equation (16):

$$G(u,v) = F(ku, kv) \tag{16}$$

where *k* is the scale. **Figure 3** shows an example of Gaussian blur, sharpening, rotation and scaling of an image in the frequency domain. The example image is an image converted from the frequency domain back to the spatial domain.

The main concept of FDRT involves transforming the image from the spatial domain to the frequency domain during each iteration of generating adversary samples. One of the four transformation methods (Gaussian blur, sharpening, rotation, or scaling) is randomly selected for image enhancement, with a probability of P = 0.5. The enhanced image in the frequency domain is then converted back to the spatial domain and used as the input image for the next iteration, enabling the generation of transferable adversary samples. Based on I-FGSM algorithm, Algorithm 1 mainly shows the pseudo code of the main framework of FDRT-I-FGSM algorithm.

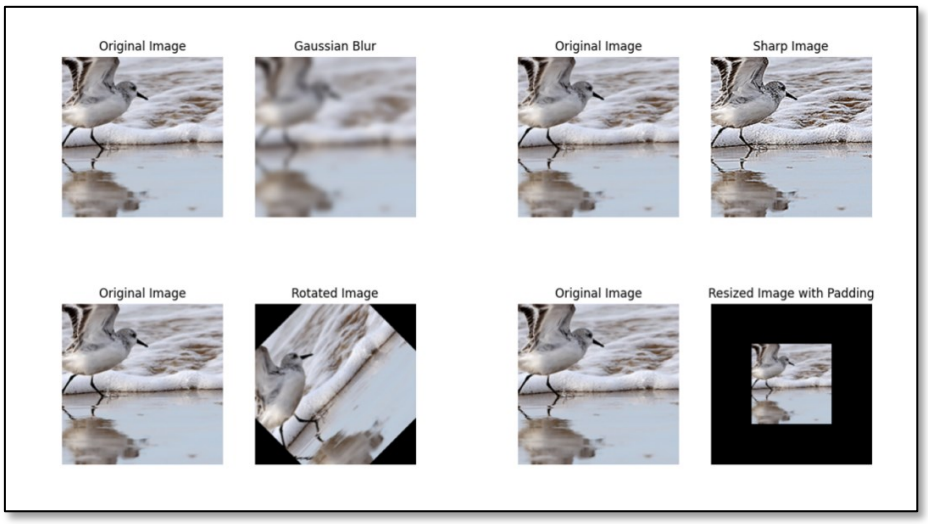


Figure 3. Example of Gaussian blur, sharpening, rotation and scaling in frequency domain.

Algorithm 1 FDRT-I-FGSM

Input: clean sample x, normalized to [-1,1], its corresponding real label is y^{true}, the maximum infinite norm perturbation is ε, the iteration step is α, and the iteration round is T， The classification function is f, the loss function is J, the transformation from spatial domain to frequency domain is *DCT* transformation, the transformation from frequency domain to spatial domain is *IDCT* transformation, the image transformation function is θ, and the transformation probability is P

Output: the adversary sample is x^{adv}

1: Initialize adversary sample $x_0^{adv} = x$； $\alpha = \varepsilon/T$

2: for $t = 0$ to $T - 1$ do:

3: Input image x into classification function *f* and calculate gradient information $\nabla_x J(\theta, x, y^{true})$;

4: Update the adversary sample and cut it $x_{t+1}^{adv} = Clip_{x,\varepsilon}\left\{x_t^{adv} + \alpha sign\left(\nabla_x J(\theta, x_t^{adv}, y^{true})\right)\right\}$;

5: *DCT* (x_{t+1}^{adv}) transformation will be performed on the adversary sample;

6: Perform random transformation of frequency domain image $\theta(DCT(x_{t+1}^{adv}), P = 0.5)$;

7: *IDCT* $(\theta(DCT(x_{t+1}^{adv}), P = 0.5)$ transformation will be performed on the adversary sample;

8: return $x^{adv} = x_{t+1}^{adv}$

Transformation of frequency domain images is shown in **Figure 4**. Input a clean image, carry out back propagation after obtaining the prediction result, and move forward in the direction of maximum loss to generate adversary perturbation. The perturbation is added to the clean image, and then the DCT frequency domain transform is performed. In the frequency domain, the Gaussian blur, sharpen, rotation and scaling transform are randomly selected, and the transformation probability is 0.5. After the transformation is completed, the IDCT is switched from the frequency domain to the spatial domain, and then the iterative input is performed again, and the final adversary sample can be obtained until the stop condition is met.

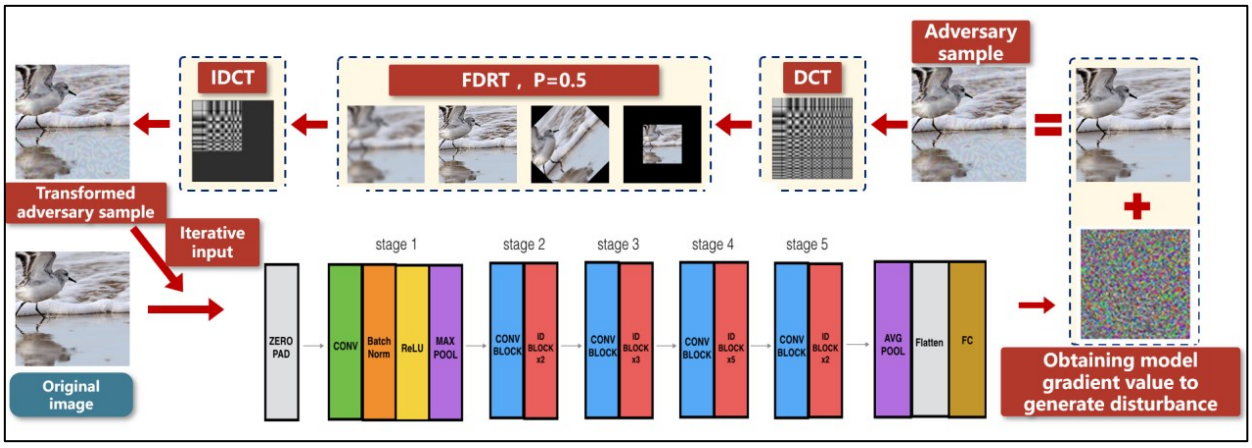


Figure 4. Schematic diagram of generating adversary samples based on random transformation of frequency domain image.

4. Experiment

In this section, the ResNet50 model with CBAM attention mechanism is used as the attacked white-box model. At the same time, the frequency domain FDRT random transformation framework proposed by us is introduced to generate adversary samples. The finally generated adversary samples are transferred to other models as black-box models to explore the change of transferability success rate and the specific impact of different parameters on transferability success rate.

4.1. Experimental setup

Based on the ImageNet dataset, this article selected 1000 samples as test data to ensure a 100% recognition success rate for the model.

The evaluation of adversarial samples is one of the important means to evaluate the robustness of deep learning models. Attack success rate and image visualization analysis are the two main indicators for adversarial sample evaluation.

Attack success rate refers to the success rate of attackers using adversarial sample attack models. In general, the attack success rate can be represented by Equation (17).

$$ASR = \frac{\sum_{i=1}^{n}[f(x_i + \delta_i) \neq y_i]}{n} \tag{17}$$

Among them, n is the number of test samples, x_i is the original test sample, δ_i is the adversarial perturbation added to x_i, f is the classifier of the deep learning model, and y_i is the true label of sample x_i. If the adversarial perturbation can successfully change the classification results of the model, i.e. $f(x_i + \delta_i) \neq y_i$, then this sample is considered a successful attack. The higher the success rate of the attack, the poorer the robustness of the model, and the stronger the performance of the attack algorithm.

Image visualization analysis can help researchers better understand adversarial samples. In general, it is possible to compare the original image with the adversarial sample. This method can provide a more intuitive understanding of the impact of adversarial perturbations on the original image, as well as the differences between adversarial samples and the original image. Another visualization method is to use thermal maps to determine the focus of attention of the model, in order to better

understand adversarial samples.

This article will discuss the maximum perturbation value ε set to 20/255, except for FGSM, the perturbation step size of attack algorithms is set to 2/255, the number of iterations is set to 10, the momentum attenuation factor is 1, and the transformation probability $P = 0.5$.

4.2. Attack experiment

In this experiment, two white-box models, ResNet50 and CBAM-ResNet50, were selected and used to generate adversarial samples. At the same time, two black-box models were selected: Inception V3 and Inception-ResNet V2, which were unable to understand their internal structure. Then five adversarial sample algorithms will be used: FGSM, I-FGSM, MI-FGSM, PGD, and NI-FGSM to attack these models. Among them, experiments without perturbation and with random noise were added as a comparison. The experimental results are shown in **Table 1**.

Table 1. CBAM + FDRT white-box and black-box attack experiment results.

	Attack	**ResNet50**	**Inception-V3**	**Inception-ResNet V2**	**Average**
	No Perturbation	0%	0%	0%	-
	Random Noise	5.4%	2.1%	0.9%	
ResNet50	FGSM	76.5%	24.6%	28.3%	26.45%
	I-FGSM	89.3%	29.0%	29.9%	29.45%
	MI-FGSM	97.0%	33.5%	34.1%	33.8%
	PGD	100%	36.7%	35.2%	35.95%
	NI-FGSM	100%	48.9%	50.4%	49.65%
CBAM-ResNet50	FDRT-FGSM	81.3%	35.3%	34.1%	34.7%
	FDRT-I-FGSM	89.9%	39.4%	38.9%	39.15%
	FDRT-MI-FGSM	100%	42.3%	44.7%	43.5%
	FDRT-PGD	100%	45.7%	48.0%	46.85%
	FDRT-NI-FGSM	100%	60.5%	59.6%	60.05%

As can be seen from **Table 1**, on the white-box model ResNet50, after the combination of CBAM attention mechanism and FDRT framework, the attack performance adversary samples have been significantly improved. For example, when FDRT-FGSM is used, the attack success rate increases from 76.5% to 81.3%. Similarly, when using FDRT-I-FGSM and FDRT-MI-FGSM, the attack success rate has been significantly improved, from 89.3% and 97.0% to 89.9% and 100%, respectively. On the black-box models Inception V3 and Inception-ResNet V2, the transferability performance against various attacks has also been improved after the combination of CBAM ResNet50 and FDRT frameworks. For example, when FDRT-FGSM and FDRT-I-FGSM are used, the attack success rate increases from 24.6% and 29.0% to 35.3% and 39.4% respectively. When FDRT-MI-FGSM is used, the attack success rate increases from 34.1% to 42.3%. When FDRT-NI-FGSM is used, the attack success rate on the black-box model increases from 48.9% and 50.4% to 60.5% and 59.6%, respectively. Finally, we also calculate the average transferability attack

success rate, and the results show that the transferability attack success rate is significantly improved after adding CBAM attention mechanism and FDRT framework to each attack. Therefore, the combination of CBAM attention mechanism and FDRT framework can significantly improve the attack success rate and transfer attack performance and the attack success rate is higher than that of using one of the methods alone.

Transferability analysis

The CBAM attention mechanism improves model robustness and the transferability of adversarial samples by guiding the model to emphasize crucial features in the input image. This reduces the model's vulnerability to attacks. CBAM combines channel attention and spatial attention, making the model more focused on meaningful areas in the image, thereby improving the robustness of the model. By generating perturbations that ultimately lead to classification errors in the robust region of the image, CBAM can enhance the performance of transfer attacks by enabling the generated perturbations to have better generalization performance on other models.

Secondly, the frequency domain-based image random transformation framework FDRT can further improve the transferability of performance of transfer attack. FDRT destroys local feature information in adversarial samples by randomly transforming images in the frequency domain. This transformation diversifies the adversarial samples in multiple iterations, forcing the model to generate stronger adversarial perturbation that are more robust and diverse. The frequency domain transformation mainly targets the high-frequency parts (usually details) in the image, and by randomizing these details, it disrupts the local features of the original image, making it more difficult for the adversarial sample features to be captured and recognized by the model.

Therefore, the combination of CBAM and FDRT utilizes the advantages of attention mechanism and frequency domain transformation, which can further enhance the performance of transfer attacks. The attention mechanism can guide the model to focus on important image features from a global perspective, improve the robustness of the model, and force attacks to generate perturbations in key areas of the image during the generation of perturbations; The frequency domain transformation can destroy local features in the image, and make the generated adversarial perturbation more random and diverse, thereby making the adversarial samples more robust and transferable. In summary, the combination of these two methods can improve the robustness of adversarial perturbation at both global and local levels, thereby enhancing the performance of transfer attacks.

4.3. Hyper-parameters

In order to further explore how to improve the performance of transfer attacks, this section of the experiment investigated the effects of iteration number, and transformation probability P on the transferability of adversarial samples.

4.3.1. Impact of iteration times on transferability

Based on the previous experimental setup, this paper gradually increases the number of iterations (starting from 2, in steps of 4, with the maximum perturbation

value of 20/255), and studies the relationship between the number of iterations and the success rate of transfer attack. FGSM algorithm belongs to one-time iterative algorithm, so it does not need to be discussed. **Figure 5** shows the relationship between the number of iterations and the success rate of transfer attack. The results show that with the increase of iteration rounds, the success rate of white-box attack of FDRT-I-FGSM method is significantly improved, while the success rate of black-box transfer attack is slightly decreased at about 10. This is because as the number of iterations increases, the degree of fitting against samples in the classification model will be higher and higher, so the transferability will decline. However, for FDRT-MI-FGSM, FDRT-PGD and FDRT-NI-FGSM, the success rate of black-box attack has been significantly improved with the increase of the number of iterations, indicating that the influence factors in these methods can better match the number of iterations, such as momentum method and NAG. However, the increase in the number of iterations also means an increase in computational overhead. Therefore, the number of iterations is selected as 10 to ensure the high success rate of white-box and black-box attacks and the fast generation rate of adversary samples.

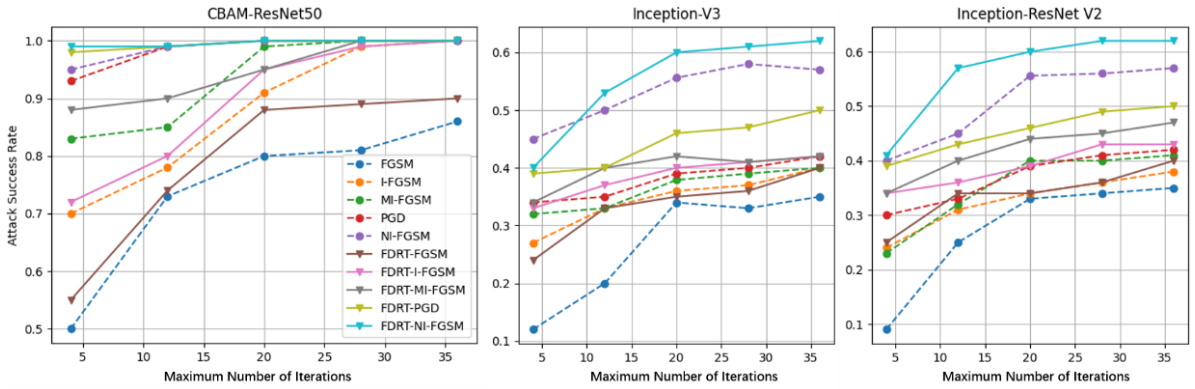


Figure 5. The impact of the maximum number of iterations on the success rate of transferability attack.

4.3.2. Influence of frequency domain transformation probability *P* on transferability

The previous part has verified the effectiveness of the CBAM + FDRT scheme proposed in this paper. In this section, the transformation probability *P* is further studied in the above experiments. In order to explore the impact of transformation probability *P* on the success rate of transferability, the value range of transformation probability *P* is fixed in the range of 0 to 1, and the step size is 0.1. For the sake of experimental preciseness (eliminating the influence of attention mechanism), only ResNet50 model is used here as the white-box model, and the adversary sample algorithm is FGSM, I-FGSM, MI-FGSM, PGD and NI-FGSM after adding FDRT framework. The experimental results are shown in **Figure 6**.

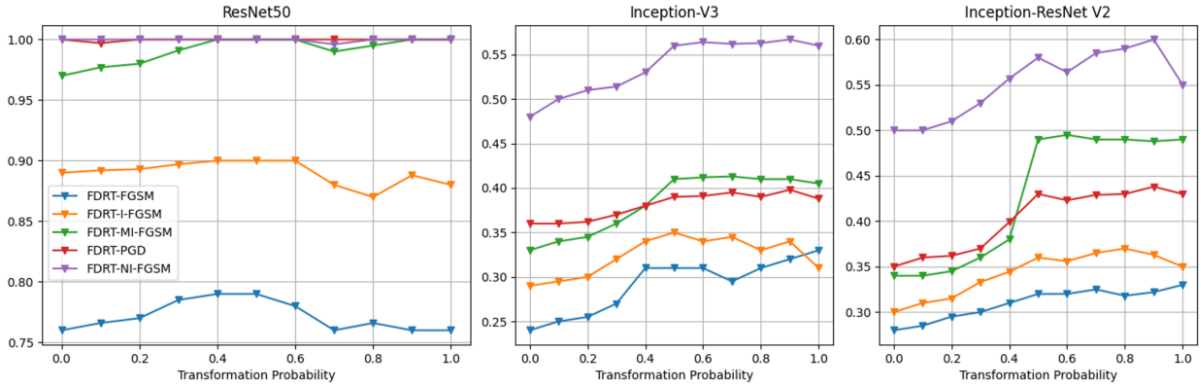


Figure 6. Influence of frequency domain transformation probability P on transferability attack success rate.

It can be seen from **Figure 6** that in the white-box model ResNet50, with the gradual increase of frequency domain transformation probability P, the attack success rate increases slightly. When $P = 0.5$, the attack success rate is the highest. When $P > 0.5$, the attack success rate is floating, and the FDRT-I-FGSM attack success rate has a significant downward trend. In black-box attack, with the increase of frequency domain transformation probability P, the attack success rate increases significantly, and the average increase is about 5%. In particular, in the Inception-ResNet V2 model, the attack success rate of FDRT-MI-FGSM increases the most when P changes from 0.4 to 0.5. In general, whether it is a white-box attack or a black-box attack, the highest attack success rate is when $P = 0.5$. Therefore, this paper selects the frequency domain transformation probability $P = 0.5$ as the best parameter to achieve the best attack effect.

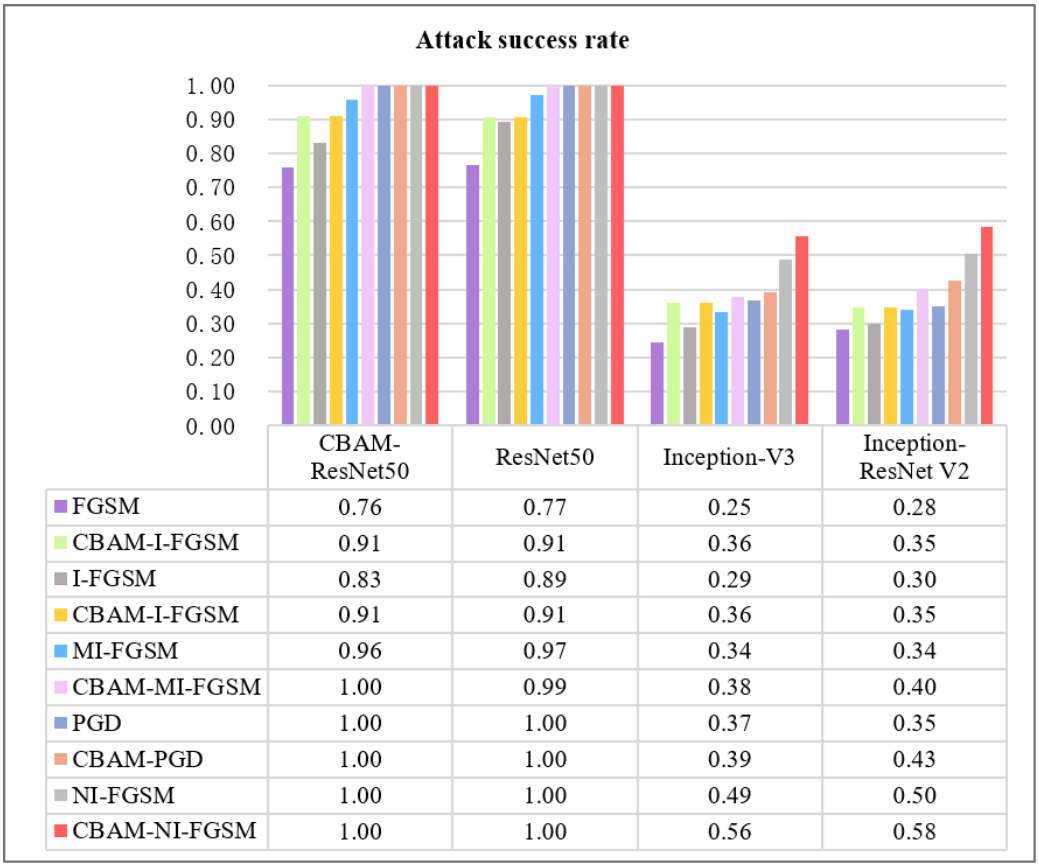
4.4. Ablation experiment

4.4.1. Analysis of experimental results of introducing attention mechanism to attack

Since the experiment of attacking ResNet50 only with FDRT method has been discussed in the above experiment, the experiment in this section only needs to discuss the transferability experiment of introducing CBAM-ResNet50 model. In this experiment, two white-box models, ResNet50 and CBAM-ResNet50, were selected and used to generate adversary samples. At the same time, two black-box models are selected: Inception V3 and Inception ResNet-V2. These two black-box models cannot understand their internal structures. Then we use five sample algorithms: FGSM, I-FGSM, MI-FGSM, PGD and NI-FGSM to attack these models. The experimental results are shown in **Figure 7**. According to the experimental results in **Figure 7**, the following conclusions can be drawn:

1) After the ResNet50 model is used as the white-box model, and the five adversarial sample attack algorithms of FGSM, I-FGSM, MI-FGSM, PGD and NI-FGSM are used to generate adversary samples, the transferability attack on the CBAM-ResNet50 model will slightly reduce the attack success rate. However, when the CBAM-ResNet50 model is used as a white-box model to attack the ResNet50 model, the attack success rate increases slightly, which indicates that the CBAM-ResNet50 model is more robust to transferability attacks.

2) From the perspective of the black-box model, under the attack using ResNet50 and CBAM- ResNet50 as the white-box model, CBAM-ResNet50 model has a higher transfer attack success rate than ResNet50 model. At the same time, the adversary samples generated by NI-FGSM algorithm have the highest attack accuracy.
3) Compared with the improved transferability of CBAM-ResNet50 model, the CBAM+FDRT method exhibits even better performance enhancement on attack transferability, validating its effectiveness.



	CBAM-ResNet50	ResNet50	Inception-V3	Inception-ResNet V2
FGSM	0.76	0.77	0.25	0.28
CBAM-I-FGSM	0.91	0.91	0.36	0.35
I-FGSM	0.83	0.89	0.29	0.30
CBAM-I-FGSM	0.91	0.91	0.36	0.35
MI-FGSM	0.96	0.97	0.34	0.34
CBAM-MI-FGSM	1.00	0.99	0.38	0.40
PGD	1.00	1.00	0.37	0.35
CBAM-PGD	1.00	1.00	0.39	0.43
NI-FGSM	1.00	1.00	0.49	0.50
CBAM-NI-FGSM	1.00	1.00	0.56	0.58

Figure 7. Comparative experimental results of attacks with CBAM attention mechanism.

4.4.2. Visual analysis and comparison

In order to judge whether the perturbation generated by the model with attention mechanism is more specific to the robustness characteristics of the image, we use ResNet50 model and CBAM-ResNet50 model to generate five adversary sample algorithms for an image respectively, and conduct heatmap visual analysis on the generated adversary samples. Visualization of samples using heatmap can more intuitively show the contribution of each pixel to the classification results, as shown in **Figure 8**. We compare ResNet50 model and CBAM-ResNet50 model, and finds that the two models show different color distributions on the heatmap. Specifically, in the ResNet50 model, it is found that the model's attention has been diverted. The attention should be on the target object in the original image, while the attention in the adversary sample has been diverted to the background unrelated to the classification, which can also explain why the adversary sample can successfully attack the model,

because the model's attention is not at all on the feature points of the objects in the image. In the CBAM-ResNet50 model, it can be found that the attention has not been diverted, and with the strengthening of the algorithm, the attention has converged. The main reason is that the CBAM mechanism makes the model more focused, and makes the added adversary perturbation more focused, and all of them are on important features. This result also verifies the effectiveness of CBAM mechanism in improving the robustness of the model.

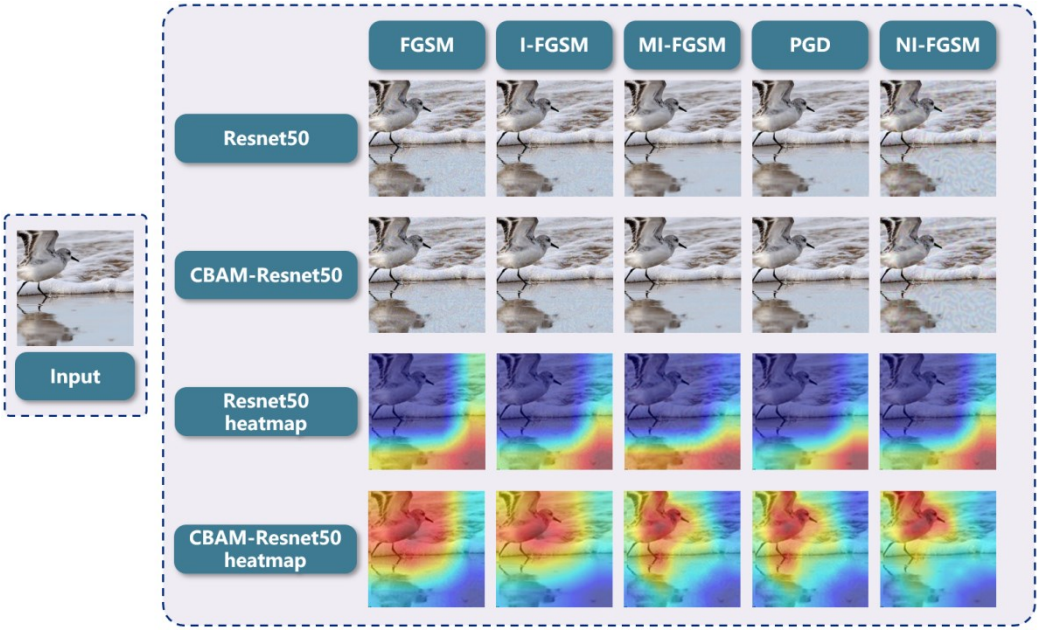


Figure 8. Visualization heatmap of adversarial samples before and after adding attention mechanism.

5. Conclusion

While the white-box attack algorithm is becoming more and more perfect, the more realistic black-box transfer attack in the real scene is subject to the poor transferability effect of the adversary samples produced by the current attack algorithm. Aiming at this urgent problem of poor transferability of adversary samples, we introduce ResNet50 model based on CBAM attention mechanism as the enhanced white-box model, an image random transformation framework FDRT based on frequency domain is proposed to improve the transferability of adversary sample. Experimental results demonstrate that the suggested method for enhancing adversary sample transferability leads to substantial improvements in the transferability of both the existing basic attack algorithm, the enhanced algorithm, and the state-of-the-art (SOTA) attack algorithm. Specifically, the proposed method based on attention mechanism and frequency domain transformation, also considering the human eye awareness and computational overhead, can improve the average black-box transferability attack success rate of the three models from 49.65% of NI-FGSM to 60.05% of FDRT-NI-FGSM. The success rate of black-box attacks exceeds 6% on average across the five methods. The experimental findings demonstrate that our proposed approach greatly enhances the performance of transfer attacks while maintaining imperceptibility to the human eye.

Author contributions: Conceptualization, MF, methodology MF, CZ and JH; software, CZ and JH; validation, CZ, JH, MX and MF; formal analysis, MX and MF; investigation, CZ and JH; resources, MF; data curation, JH; writing—original draft preparation, CZ and JH; writing—review and editing, MX and MF; visualization, CZ and JH; supervision, MF; project administration, MF; funding acquisition, MF. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:1312.6199. 2013
2. Wang J, Yin Z, Jiang J, et al. Attention-guided black-box adversarial attacks with large-scale multi-objective evolutionary optimization. International Journal of Intelligent Systems. 2022; 37(10): 7526–7547. doi: 10.1002/int.22892
3. Aslan MF, Celik Y, Sabanci K, et al. Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. International Journal of Intelligent Systems and Applications in Engineering. 2018; 6(4): 289–293. doi: 10.18201/ijisae.2018648455
4. Viaene S, Baesens B, Van Gestel T, et al. Knowledge discovery in a direct marketing case using least squares support vector machines. International Journal of Intelligent Systems. 2001; 16(9): 1023–1036. doi: 10.1002/int.1047
5. AL-Rousan N, Mat Isa NA, Mat Desa MK, et al. Integration of logistic regression and multilayer perceptron for intelligent single and dual axis solar tracking systems. International Journal of Intelligent Systems. 2021; 36(10): 5605–5669. doi: 10.1002/int.22525
6. Alarab I, Prakoonwit S. Adversarial Attack for Uncertainty Estimation: Identifying Critical Regions in Neural Networks. Neural Processing Letters. 2021; 54(3): 1805–1821. doi: 10.1007/s11063-021-10707-3
7. Wang Y, Yang G, Li T, et al. Optimal mixed block withholding attacks based on reinforcement learning. International Journal of Intelligent Systems. 2020; 35(12): 2032–2048. doi: 10.1002/int.22282
8. Andriushchenko M, Flammarion N. Understanding and improving fast adversarial training. Advances in Neural Information Processing Systems. 2020; 33: 16048–16059
9. Zhang Y, Liang P. Defending against whitebox adversarial attacks via randomized discretization. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. PMLR. 2019; 684–693
10. Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses. In: Proceedings of the International Conference on Learning Representations. 2018
11. Kwon H, Lee S. Ensemble transfer attack targeting text classification systems. Computers & Security. 2022; 117: 102695. doi: 10.1016/j.cose.2022.102695
12. Zhang Y, Tan Y, Chen T, et al. Enhancing the Transferability of Adversarial Examples with Random Patch. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence; July 2022. doi: 10.24963/ijcai.2022/233
13. Zhang C, Benz P, Karjauv A, et al. Investigating Top-k White-Box and Transferable Black-box Attack. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2022. doi: 10.1109/cvpr52688.2022.01466
14. Li Z, Yin B, Yao T, et al. Sibling-Attack: Rethinking Transferable Adversarial Attacks against Face Recognition. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2023. doi: 10.1109/cvpr52729.2023.02359
15. Zhu H, Sui X, Ren Y, et al. Boosting transferability of targeted adversarial examples with non-robust feature alignment. Expert Systems with Applications. 2023; 227: 120248. doi: 10.1016/j.eswa.2023.120248
16. Zhang J, Wu W, Huang J, et al. Improving Adversarial Transferability via Neuron Attribution-based Attacks. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2022. doi: 10.1109/cvpr52688.2022.01457
17. Liu Y, Cheng Y, Gao L, et al. Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022. doi: 10.1109/cvpr52688.2022.01468

18. Fu C, Li S, Yuan X, et al. Ad2Attack: Adaptive Adversarial Attack on Real-Time UAV Tracking. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA); 23 May 2022. doi: 10.1109/icra46639.2022.9812056
19. Zhang Y, Shin SY, Tan X, et al. A Self-Adaptive Approximated-Gradient-Simulation Method for Black-Box Adversarial Sample Generation. Applied Sciences. 2023; 13(3): 1298. doi: 10.3390/app13031298
20. Lin Y, Zhao H, Tu Y, et al. Threats of Adversarial Attacks in DNN-Based Modulation Recognition. In: Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications; July 2020. doi: 10.1109/infocom41043.2020.9155389
21. Sun X, Cheng G, Li H, et al. Exploring Effective Data for Surrogate Training Towards Black-box Attack. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2022. doi: 10.1109/cvpr52688.2022.01492
22. Cai Z, Song C, Krishnamurthy S, et al. Blackbox attacks via surrogate ensemble search. Advances in Neural Information Processing Systems. 2022; 35: 5348–5362
23. Chen J, Jordan MI, Wainwright MJ. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP); May 2020. doi: 10.1109/sp40000.2020.00045
24. Wang Z, Guo H, Zhang Z, et al. Feature importance-aware transferable adversarial attacks. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); October 2021. doi:10.1109/iccv48922.2021.00754
25. Huang Q, Katsman I, Gu Z, et al. Enhancing Adversarial Example Transferability with an Intermediate Level Attack. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); October 2019. doi: 10.1109/iccv.2019.00483
26. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. 2014
27. Kurakin A, Goodfellow IJ, Bengio S. Adversarial Examples in the Physical World. Artificial Intelligence Safety and Security. 2018; 99–112. doi: 10.1201/9781351251389-8
28. Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks with Momentum. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2018. doi: 10.1109/cvpr.2018.00957
29. Lin J, Song C, He K, et al. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. International Conference on Learning Representations. 2020
30. Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks. In: Proceedings of the International Conference on Learning Representations; 2018
31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv preprint arXiv:1706.03762. 2017
32. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); 2018: 3–19
33. Xie C, Zhang Z, Zhou Y, et al. Improving Transferability of Adversarial Examples with Input Diversity. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2019. doi: 10.1109/cvpr.2019.00284
34. Huang Y, & Kong AWK. Transferable adversarial attack based on integrated gradients. arXiv preprint arXiv:2205.13152.
35. Zhao A, Chu T, Liu Y, et al. Minimizing Maximum Model Discrepancy for Transferable Black-box Targeted Attacks. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2023. doi: 10.1109/cvpr52729.2023.00788
36. Feng Y, Wu B, Fan Y, et al. Boosting Black-Box Attack with Partially Transferred Conditional Adversarial Distribution. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022. doi: 10.1109/cvpr52688.2022.01467