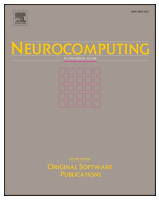
Neurocomputing 610 (2024) 128512

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A review of black-box adversarial attacks on image classification

Yanfei Zhu, Yaochi Zhao *, Zhuhua Hu, Tan Luo, Like He

Hainan University, Haikou, China

ARTICLE INFO

Keywords:
Deep learning
Image classification
Black-box adversarial attacks
Adversarial perturbations

ABSTRACT

In recent years, deep learning-based image classification models have been extensively studied in academia and widely applied in industry. However, deep learning is inherently vulnerable to adversarial attacks, posing security threats to image classification models in security sensitive field, such as face recognition, medical image diagnosis and traffic sign recognition. Especially for black-box adversarial attacks, which can be carried out even without remote model information, the security issues facing deep learning are even more serious. Despite more and more attentions on this issue, existing reviews always analyze black-box adversarial attack only from one perspective, focus on only a certain application field. This paper systematically reviews and discusses existing progress, demonstrating black-box adversarial attacks from multiple perspectives and systematically classifying existing methods. Besides, we also sort out and categorize the application of current black-box adversarial attacks and identify several promising directions for future research.

1. Introduction

In recent years, deep learning has been significantly innovated various fields in academia and industry, driven by the availability of large-scale datasets, the emergence of GPUs, and the advancement of deep network architectures. It has exhibited great successes not only in discriminative tasks such as facial recognition, and speaker identification, but also in intricate and nuanced problems, such as gene sequencing, autonomous driving, and malware detection.

Image classification is a pivotal topic in computer vision. In the early stage, image classification heavily relies on manually designed feature extractors, with serious performance limitation in case of complex scenarios. After the advent of deep learning, image classification became one of the earliest applications of deep learning. Thanks to the high-quality feature representation ability, deep learning has greatly boosted image classification [1–3], making image classification the most vibrant field today.

However, Goodfellow et al. reveal that deep learning models are susceptible to adversarial perturbations [4], posing serious security threats to deep learning-based image classification models in security sensitive field. For instance, adversarial perturbations can potentially result in the recognition errors of medical image, leading to incorrect medical diagnoses and wrong treatment decisions. Similarly, in the case of autonomous driving, adversarial perturbations on specially crafted traffic signs may cause misclassification of traffic signs, resulting in severe safety accidents. Currently, the research on adversarial attack to deep learning-based image classification model has attracted increasing attention.

Adversarial attacks refer to the intentional manipulation of input samples through adding subtle and imperceptible perturbations, which may lead model to output incorrect results with high confidence. Based on the level of access granted to attacker, adversarial attacks can be classified into white-box adversarial attacks and black-box adversarial attacks. For white-box adversarial attacks, such as Fast Gradient Sign Method (FGSM) [4], DeepFool [5], Carlini & Wagner Attack (C&W) [6], etc., the attacker needs to obtain all information about the target model, including its structure, parameters and so on. Therefore, white-box attack require access to the gradients of target model and directly interact with the model. On the contrast, for black-box adversarial attacks, the attacker only needs to access the target model and obtain its output information to launch an attack. Clearly, black-box adversarial attacks are more aligned with real-world attack scenarios, and the research of black-box adversarial attacks is more realistic.

To date, there exists massive research on black-box adversarial attacks [15]. Fig. 1.1 shows the number of publications and citations on black-box adversarial attacks in the field of image classification in Web of Science databases. It can be observed that the number of citations increases year by year, and the number of publications reaches its highest value in 2022. Despite the rapid evolution in this field, there is still lack of systematic study to review and discuss existing progress.

* Corresponding author.
E-mail address: zhyc@hainanu.edu.cn (Y. Zhao).

https://doi.org/10.1016/j.neucom.2024.128512
Received 15 March 2024; Received in revised form 7 August 2024; Accepted 27 August 2024
Available online 30 August 2024

Prior reviews usually analyze black-box adversarial attacks only from one perspective, and only for a certain application field. In this paper, black-box adversarial attack methods are systematically classified from multiple perspectives, and their common application areas are also sorted out and categorized.

In the existing surveys, black-box adversarial attacks usually are classified according to the required attack requirements, such as transferability-based attacks [4,8,9,11,15–19,33,34], score-based attacks [20,22,24,92–94], and decision-based attacks [25,35–38]. However, viewing black-box adversarial attacks from different perspectives can lead to different categorizations, providing a more systematic and detailed understanding. Furthermore, comparing different methods can help us to get the development trends in this field. On consideration of these, this article views black-box adversarial from five perspectives. Apart from the traditional attack requirement-based classification, this article further categorizes black-box attacks into universal attacks and specific attacks based on the perturbation type. Further, according to the perturbation space, black-box adversarial attacks are classified into pixel-space attacks and feature-space attacks in this article. Depending on the attack scenario, we divide black-box adversarial attacks into digital attacks and physical attacks. Moreover, given that generative models are proved to be effective for generating adversarial samples, this paper specifically sorts out distinct generative models based black-box adversarial attacks, including diffusion models that gain popularity in recent years. Besides, while prior work research black-box adversarial attacks focusing on one specific application domain, this paper provides a classification and research status of black-box adversarial attacks across various domains.

The main contributions of this article are as follows:

1. This article demonstrates black-box adversarial attacks in the field of image classification from multiple perspectives and systematically classify existing methods, based on perturbation type, attack requirements in the past two years, perturbation space, attack scenarios, and the type of generative model used.
2. This article introduces black-box attack methods and research status in various security sensitive domains, including face recognition, medical image detection and traffic sign recognition, etc.
3. This article identifies several promising directions for future research.

The rest of this article is organized as follows: Section 2 provides an introduction of basic concepts, including relevant terminology and white-box attack algorithms. Section 3 explores black-box adversarial attack methods from multiple perspectives. Section 4 presents the black-box adversarial attacks in various security sensitive domains. The last part of this article provides future directions and summarize our work in the conclusion.

2. Fundamental concept

2.1. Definition of terms

1. Parameter Gradient : It refers to the gradient of loss function with respect to model parameters. It quantifies how much the loss function would change as the adjustments of model parameters. It guides the optimization process of neural networks.
2. Input Gradient: It refers to the gradients of loss function with respect to input data. It quantifies the impact of variations of individual pixels in input image on the loss function. Input gradient is commonly used for specific tasks, such as interpretability analysis and generating adversarial examples.
3. Gradient-Based Attack: Gradient-Based Attack: A kind of attack methods that require computing input gradient. Most of white-box attacks that have full access to attacked models, and can obtain the input gradient, belongs to gradient-based methods.
4. Gradient-free Attack: A kind of attack methods that do not need computing gradient.
5. Data-free Attack: A kind of attack methods that do not require original training data of attacked models.
6. Logits: The raw output vector outputted by the final layer of model without undergoing an activation function. Each element in this vector represents the score for the corresponding predicted class. It can be transformed into a probability distribution through an appropriate activation function.
7. Probability: The output after applying an activation function on logits, representing the confidence levels of model for different classes. Each element of it indicates the probability of input belonging to the corresponding class. These elements range from 0 to 1, and the sum of all elements equals to 1.
8. Hard Label: The predicted label class outputted by model for input samples, without probability information.
9. Non-targeted Attack: A kind of attack methods that cause the target model to misclassify input data to any wrong class. Its attack objective is expressed by Eq. 1, where $f(\bullet)$ is the target model and $x+\xi$ represents adversarial samples.

$$f(x+\xi) = y' \neq y\# \tag{1}$$

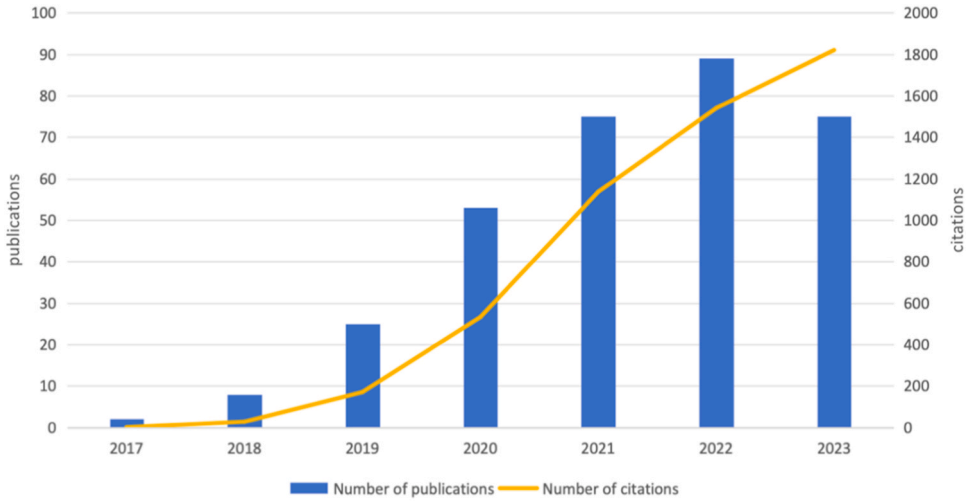


Fig. 1.1. Numbers of publications and citations on black-box adversarial attacks in the field of image classification from 2017 to 2023.

10. Targeted Attack: A kind of attack methods that cause the target model to misclassify input data to the specific, predefined and incorrect category. Its attack objective is shown by Eq. 2, where the target model misclassifies the adversarial sample $x+\xi$ into the wrong class k.

$$f(x+\xi)=y^{'}=k\neq y\# \tag{2}$$

11. Confidence: The degree of certainty that a neural network model assigns to a class for inputs. It is often represented as a score or probability.
12. Attack Success Rate: The probability that an attacker successfully misleads adversarial samples into wrong classes (for non-targeted attacks) or into specific wrong classes (for targeted attacks). It is calculated as the ratio of the number of successful attacks to the total number of attack attempts.
13. ℓ_0 norm: In the field of adversarial attacks, ℓ_0 norm quantifies the number of altered pixels, relative to the original sample.
14. ℓ_1 norm : It also known as Manhattan Norm. The ℓ_1 norm of $x\in\mathbb{R}^n$ is defined as Eq. 3, where $|\bullet|$ is the absolute value.

$$\|x\|_1:=\sum_{i=1}^{n}|x_i|\# \tag{3}$$

15. ℓ_2 norm: It also known as Euclidean Norm. The ℓ_2 norm of $x\in\mathbb{R}^n$ is defined as Eq. 4.

$$\|x\|_2:=\sqrt{\sum_{i=1}^{n}x_i^2}=\sqrt{x^Tx}\# \tag{4}$$

16. ℓ_∞ norm: The definition on $x\in\mathbb{R}^n$ is shown by Eq. 5. It quantifies the maximum value of $x\in\mathbb{R}^n$.

$$\ell_\infty:=\max_{1\leq i\leq n}|x_i|\# \tag{5}$$

17. Robustness: The ability of neural network model to maintain stability and reliability under adversarial attacks.
18. Transferability: The generalization ability of adversarial attacks, which means the attack ability of adversarial samples or adversarial attacks, generated according to one neural network, is also effective on different models, different datasets or in different tasks.

    Transferability: The generalization ability of adversarial attacks, which means the attack of adversarial samples or adversarial attack methods, generated according to one neural network, is also effective on different models, different datasets or in different tasks.
19. Target Model: The attacked model.
20. Substitute Model: The model trained to mimic the prediction behavior of target model. It is also known as proxy model.
21. Pixel Space: The space composed of every pixel in an image. In pixel space, an image is transformed into a matrix or two-dimensional array that can be processed by computers, with each element representing the pixel value at the corresponding position in image.
22. Feature space: The space composed of feature representations of images extracted at different layers of the neural network.

2.2. White-box adversarial attacks

Though, this article focuses on black-box adversarial attacks, we still briefly introduce some white-box attack methods, considering that white-box adversarial attacks usually serve as the foundation of some black-box adversarial attacks. Currently, there is a wide range of white-box attack techniques, including Fast Gradient Sign Method (FGSM), Iterative-FGSM, Basic Iterative Method (BIM), Projected Gradient Descent (PGD), Carlini & Wagner Attack (C&W), Jacobian-based Saliency Map Attack (JSMA), and DeepFool, etc. In this context, below provides a concise overview of white-box adversarial attacks.

FGSM: It is a technique for rapidly generating adversarial samples. Attackers compute the input gradient of model and multiply it by a step size to construct adversarial samples, as shown in Eq. 6, where ϵ represents step size, $sign(\bullet)$ is sign function, $\mathscr{L}(\bullet)$ denotes loss function, and x,y,θ respectively are input, label and parameters

$$x^{adv}=x+\epsilon\bullet sign(\nabla_x\mathscr{L}(x,y;\theta))\# \tag{6}$$

BIM: It is an extension of FGSM, as shown in Eq. 7, where α denotes a small step size. In contrast to FGSM, which is a one-step method that perturbs an image sample with a large step in the direction of maximizing loss function, BIM iteratively increases loss function through multiple small steps. After each iteration, the resulted pixel values are clipped to ensure that the adversarial sample remains within an ϵ-neighborhood of original image.

$$x_0^{adv}=x$$

$$x_{t+1}^{adv}=Clip_{x,\epsilon}\left(x_t^{adv}+\alpha\bullet sign\left(\nabla_x\mathscr{L}\left(x_t^{adv},y;\theta\right)\right)\right)\# \tag{7}$$

C&W: As an optimization-based approach, C&W generates effective adversarial samples with minimal perturbation by optimizing a loss function, as shown in Eq. 8, where δ represents the adversarial perturbation, $D(\bullet,\bullet)$ denotes ℓ_0, ℓ_2, or ℓ_∞ distance metric, and c is a hyperparameter to balance the importance of D and $\mathscr{L}(\bullet)$ is a user-defined loss function.

$$\min_{\delta}D(x,\quad x+\delta)+c\bullet\mathscr{L}(x+\delta)\# \tag{8}$$

$$s.t.\quad x+\delta\in[0,1]^n$$

To ensure producing valid adversarial samples, Eq. 9 is used to constrain the pixel values of $x+\delta$ within $[0,1]^n$, where w is the parameter that needs to be optimized.

$$x+\delta=\frac{1}{2}(\tanh(w)+1)\# \tag{9}$$

DeepFool: It iteratively calculates the shortest distance from a training sample to a classification hyperplane, resulting in the minimal perturbation. For binary classification tasks, the principle can be expressed in Eq. 10.

$$r^*(x_0):=argmin\|r\|_2$$

$$s.t.sign(f(x_0+r))\neq sign(f(x_0))\# \tag{10}$$

where the distance r can be expressed as $\frac{f(x)w}{\|w\|_2^2}$, r* represents the shortest distance from a training sample to the hyperplane, $\hat{k}(x)=sign(f(x))$ represents the label value of classifier. For multi-classification tasks, the algorithm only needs to optimize the Eq. 11, where w_k is the k-th column of w, which is the weight of class k.

$$r^*(x_0):=argmin\|r\|_2$$

$$s.t.\quad \exists k:w_k^T(x_0+r)+b_k\geq w_{\hat{k}(x_0)}^T(x_0+r)+b_{\hat{k}(x_0)}\# \tag{11}$$

3. Black-box adversarial attacks

This article views and classifies black-box adversarial from five different perspectives, including perturbation type, attack requirements, perturbation space, attack scenarios, and the used generative model, as shown in Fig. 3.1. It is necessary to note that research under different

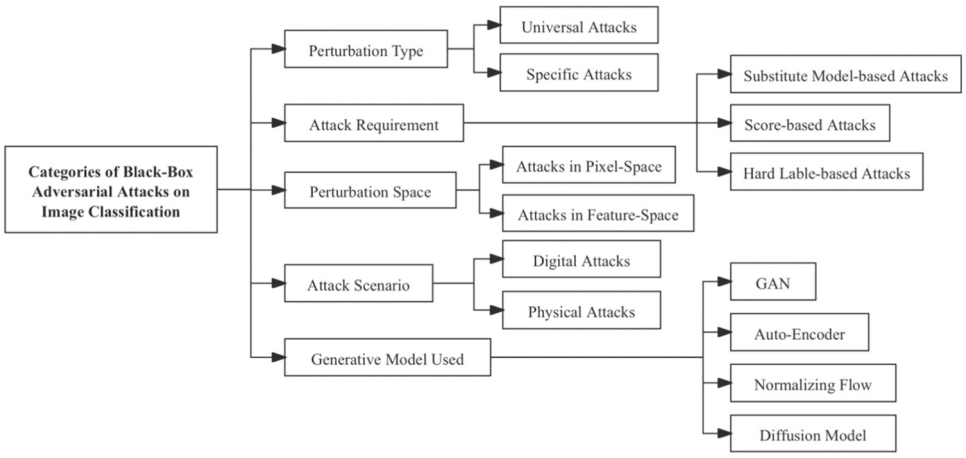


Fig. 3.1. Categories of Black-Box Adversarial Attacks in Image Classification.

classification criteria is not mutually exclusive.

3.1. Classification based on perturbation type

Adversarial attacks mislead the attacked model by adding perturbation to sample. According to the type of perturbation, black-box adversarial attacks can be divided into general black-box adversarial attacks and specific black-box adversarial attacks.

3.1.1. Universal black-box adversarial attacks

Universal black-box adversarial attacks are based on the universal adversarial perturbation (UAP), which is generated in advance and can be applied to all samples. Moosavi-Dezfooli et al. [26] are the first to reveal the existence of universal adversarial perturbations. They define an initial perturbation vector v and iteratively computes the minimal additional perturbation Δv_i for each sample x_i in the training dataset. The new perturbation $v' \leftarrow P_{p,\xi}(v+\Delta v_i)$ is designed to satisfy the constraints $\|v'\| \le \xi$ and causes a misclassification by classifier $\widehat{k}$, which satisfies $\widehat{k}(x_i+v+\Delta v_i) \neq \widehat{k}(x_i)$. Generally, universal black-box adversarial attacks are considered as practical attack methods due to their abilities to generalize across various inputs [27].

The generalization ability of universal black-box adversarial attacks depends on the transferability of UAP. Research shows that a high success rate of white-box adversarial attacks often indicates high transferability. Along this line, Hashemi et al. [32] propose a loss function that focuses on the adversarial energy of the first layer of different model architectures to improve attack success rate. This approach also makes the generated perturbations well transferable to attack other target models, enabling black-box adversarial attacks.

However, it is worth noting that perturbations generated based on transferability may have limitations in terms of their attack effectiveness. To address this issue, Wu et al. [28] introduce a hard-label-based attack method that generates UAP with texture like stripes based on orthogonal matrix. Li et al. [31] propose a customized gradient transformer module to convert UAP into region-homogeneous ones. [48,49] confirm the effectiveness of dynamic attacks on Deep Neural Networks (DNN) using optical tools.

Furthermore, UAP traditionally relies on the original training dataset. In response to this, Zhang et al. [29] regard the logits of DNN as feature representation in a vector form and analyze the mutual influence of clean images and adversarial perturbations based on the Pearson correlation coefficient (PCC). They firstly achieve data-free and targeted universal attack. [30] further explore the possibilities of data-free and non-targeted universal attacks.

3.1.2. Specific black-box adversarial attacks

In contrast to universal black-box adversarial attacks, specific black-box adversarial attacks aim to generate different adversarial perturbations for different samples to achieve accurate and effective attack results. Most research can be classified as this kind of attack, which is introduced afterward.

3.2. Classification based on attack requirement

Though without knowledge of the structure and parameter information of the attacked model, black-box adversarial attacks need to obtain the input's scores, probabilities or labels information of attacked model by means of queries. Based on the obtained information, adversarial samples can be generated and then attack can be launched. According to the dependent information, black-box adversarial attacks can be divided into score-based attacks and hard label-based attacks. Notably, in latest years, using the obtained information, substitute model can be constructed, and on the basis of substitute model, conventional methods in white-box adversarial attacks can be used to generate adversarial samples, so as to achieve black box adversarial attacks. In this section, we will demonstrate score-based and hard label-based black-box adversarial attacks, as well as the substitute model-based adversarial attacks.

3.2.1. Score-based black-box adversarial attacks

Score-based attack refer to that adversaries generate adversarial samples based on model scores (logit or probability). Its primary research objective is to reduce the number of queries while enhancing attack effectiveness.

Score-based attack is first proposed by Chen et al. [20], who employ zeroth-order gradient optimization (ZOO) algorithm to estimate the input gradients of target model. Specifically, ZOO leverages the symmetric difference quotient formula (Eq. 12) to estimate the gradient. After one more query, the coordinate-wise Hessian estimate can be obtained, as shown in Eq. 13.

$$\widehat{g}_i := \frac{\partial f(x)}{\partial x_i} \approx \frac{f(x+he_i)-f(x-he_i)}{2h} \# \tag{12}$$

$$\widehat{h}_i := \frac{\partial^2 f(x)}{\partial x_{ii}^2} \approx \frac{f(x+he_i)-2f(x)+f(x-he_i)}{h^2} \# \tag{13}$$

where e_i represents the perturbation added to original image, h (a hyperparameter) controls the perturbation magnitude, and $f(\bullet)$ is the loss function designed by C&W, as shown in Eq. 14.

$$f(x,t) = \max\left\{\max_{i\neq t}\log[F(x)]_i - \log[F(x)]_t, -\kappa\right\}\# \quad (14)$$

where $F(x)$ refers to the model prediction probability, x is the model input, t is the target class for attack, i represents other classes, and κ is a hyperparameter. Both estimate values $\widehat{g_i}$ and $\widehat{h_i}$ are then utilized in the optimization of C&W.

However, finding the appropriate optimization direction in high-dimensional spaces is not easy, which requires massive queries. To address this, the authors in [22] design an adaptive random gradient estimation strategy using a pretrained autoencoder, which reduces the number of queries without compromising attack effectiveness. Li et al. [92] propose the Projection & Probability-driven Black-box Attack (PPBA) to reduce the solution space. They transform the adversarial optimization problem into a process of recovering frequency-sparse perturbations through constructing a low-frequency constrained sensing matrix, and they use a probability-driven random walk strategy to enhance query efficiency. Andriushchenko et al. [24] introduce a random search that selects locally square regions at random positions, with perturbations approximating the boundary of the feasible set. Croce et al. [93] design versatile framework based on random search algorithms named Sparse-RS, achieving high attack effectiveness and query efficiency for various sparse attack models, including ℓ_0-bounded perturbations, adversarial patches, and adversarial frames. To further enhance query efficiency, NP-Attack [94] utilizes a special encoder-decoder model called Neural Process to characterize image structure information and explore the distribution of adversarial examples around benign inputs.

3.2.2. Hard Label-based black-box adversarial attacks

As attacked models may not output probabilities or logits, score-based attacks may not always align with practical attack environment, therefore some researchers devote to use hard labels in black-box adversarial attacks. Brendel et al. [25] propose an optimization algorithm to generate effective adversarial samples with low perturbation by probing the decision boundaries of attacked model, called decision boundary-based method. The principle of [25] is illustrated in Fig. 3.2. It starts with an adversarial sample that has a large perturbation and successfully misleads the model. Using a random walk method along the decision boundary, the perturbation magnitude is gradually reduced while ensuring adversarial success. This approach is conceptually simple and requires almost no hyperparameter tuning. However, the success rate of the attack depends on the accuracy of the decision boundary, and the attack requires numerous queries, which is challenging in real-world attack scenarios.

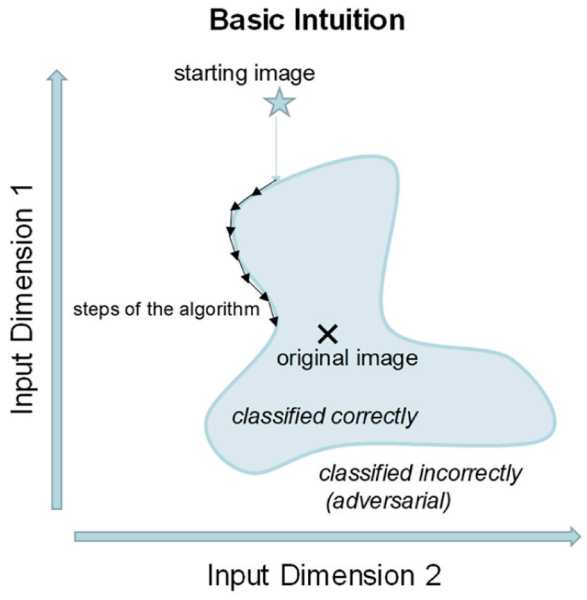


Fig. 3.2. The principle of decision boundary-based attack [25]. The initial sample starts a random walk along the decision boundary, gradually reducing the perturbation while maintaining adversarial success.

For hard label-based attacks, the minimizing of non-continuous step functions is challenging, and often necessitating the random walk strategies, which requires a large number of queries and lacks convergence guarantees. Aiming at this issue, [35] transforms the problem of minimizing non-continuous step functions into a continuous real-valued optimization problem and solves it by zeroth-order gradient optimization algorithms. Building upon this, [37] innovatively develops a series of algorithms based on a novel estimate of gradient direction using binary information at the decision boundary. For the problems that decision-based sparse attacks are difficult to optimize with NP optimization, Vo et al. [38] introduce an evolution-based algorithm, achieving higher attack success rates and query efficiency. Shi et al. [36] propose customized iteration and sampling attack (CISA), bridging transfer-based and decision-based attacks. This method estimates the distance from nearby decision boundary to set stepsize and uses a dual-direction iterative trajectory to find intermediate adversarial samples. It further customizes sampling and noise compression based on the sensitivity of each pixel and finally enhances query efficiency.

Generally, compared to score-based attacks, more researches are being conducted on decision-based attacks, which mainly focuses on optimizing algorithms to improve query efficiency and reduce computational costs.

3.2.3. Substitute model-based black-box adversarial attacks

Papernot et al. [7] propose substitute model-based black-box adversarial attack method. They train a local model (substitute model) to imitate the prediction behavior of the victim model, and generate transferable adversarial samples to attack the remote black-box model. Because adversarial samples, generated based on local model, can often effectively attack victim model [4], substitute model-based adversarial attacks are also called transfer-based methods.

To enhance the transferability of adversarial samples, some research focus on preventing perturbations from overfitting to some specific models and getting stuck in poor local optima. Wang et al. [11] consider the gradient variance of previous iteration to tune the current gradient in each iteration, thereby improving the transferability of adversarial samples. Xiong et al. [15] introduce the stochastic variance reduced ensemble attack (SVRE), which reduces gradient variance to avoid poor local optima. Latest research [34] designs a meta-learning framework, where a meta generator is trained on benign samples and fine-tuned based on adversarial samples and feedback. This framework can yield perturbations with high transferability.

Another solution to improving transferability of adversarial samples is transferring input or mapping feature. Xie et al. [16] utilize random transformations at each iteration to create diverse input patterns and train the substitute model. However, random transformations provide limited diversity due to its simplicity in image manipulation. Byun et al. [17] introduce an Object-Based Diverse Input (ODI) method, where adversarial images are drawn on 3D objects, leading to image misclassification. Wang et al. [33] no longer transform a single image but compute the gradients of input image, which is admixed with a small portion of each add-in image from other categories. In contrast to improving transferability through input image, [19] utilizes the contrastive spectral training to obtain feature extractors on source domain and scrambles the intermediate and final layers of the feature extractor to generate adversarial samples on target domain. However, feature-level attacks often provide inaccurate estimates of neuron importance. Therefore, Zhang et al. [18] propose Neuron Attribution-based Attack (NAA), which entirely attributes the model's output to each neuron in an intermediate layer. By deriving an

approximation scheme of neuron attribution, NAA reduces computation and enhances sample transferability in feature-level attacks.

Though the process of training substitute models is relatively complex, it remains an effective approach. Moreover, the limitations of accessing the original training dataset are mitigated, making substitute model-based adversarial attacks more feasible. This is further discussed in Section 3.5, where Generative Adversarial Networks (GANs) are utilized to carry out data-free attacks.

3.3. Classification based on perturbation space

Adversarial perturbations in back-box adversarial attacks can occur in pixel space or feature space, demonstrated as follows.

3.3.1. Black-box adversarial attacks in pixel space

Adversarial attacks in pixel space directly perturb the pixels in input image to cause misclassification. This type of attack perturbs pixels typically based on their impact on the loss value or classification result, which leads to perturbations spreading across the entire pixel space. To generate imperceptible adversarial samples, ℓ_p norms are used in objection function to constrain the magnitude of perturbations or the number of perturbed pixels. In black-box adversarial attacks, there are a lot of research on attack in pixel-space, such as VT-Attack [11], Admix [33], FE-DaST [13], SVRE [15], etc.

Notably, sparse attack in pixel-space has gained much attention because it generates imperceptible adversarial samples. [39,40] perturb local pixel spaces by minimizing l_0 norm to enhance the transferability and imperceptibility of adversarial samples. Furthermore, He et al. [41] introduce a generator architecture that decouples sparse perturbations into two components: amplitude and position, and jointly optimizes them to generate sparse adversarial samples with relatively strong transferability. Croce et al. [93] propose Sparse-RS framework, which enhances the query efficiency of score-based attacks across multiple models. Lately, due to the gradually increasing attack effectiveness, sparse adversarial attacks have gained increasing attention in adversarial attack community.

3.3.2. Black-box adversarial attacks in feature space

In recent years, some studies explore perturbing feature information of image. For example, Huang et al. [42] generate adversarial perturbations with high-level semantic patterns by training a pre-trained models to learn a low-dimensional embedding and performing efficient search within the embedding space. [44] introduces an adversarial distribution searching-driven attack, which incorporates an auxiliary network to search potential distributions in latent space for generating adversarial perturbations. It also uses edge detection to outline minimal disturbance regions. Cao et al. [45] propose a latent space feature based adversarial attack, which utilizes an adversarial feature extraction network to extract feature maps from the input's latent space and uses these maps as adversarial perturbations.

Additionally, adversarial attacks can be carried out by altering feature information. For instance, [43] injects imperceptible perturbations to feature space through an optimization procedure. Chen et al. [46] utilize stable diffusion to train and generate adversarial samples in latent space after dimensionality reduction by encoder. They also employ attention mechanisms to extract structural information and its correlation with text for generating imperceptible and transferable adversarial samples.

Compared to perturbations obtained through attack in pixel-space, perturbations altering or derived from feature space often have stronger connection to the semantic information of image, which results in higher transferability.

3.4. Classification based on attack scenario

Existing adversarial attacks can be categorized into digital attacks and physical attacks, each of which has different characteristics in the attack targets, implementation methods as well as other factors.

3.4.1. Digital attacks

Digital attacks are aiming for high performance in controlled laboratory settings. Compared to physical attacks, digital attacks are more efficient and cost-effective because they can be rapidly executed in a computer environment without the need for equipment and material associated with real-world deployment. Additionally, attackers can have finer control over the attack process. Most of the previously mentioned attacks fall into the category of digital attacks.

3.4.2. Physical attacks

Due to various constraints imposed by real-world conditions, most digital attacks are difficult to implement in physical environments or not as effective as in controlled laboratory settings. Therefore, physical attacks focus on deploying practical and robust attacks in real-world scenarios. In recent years, the widespread application of DNNs makes research on physical attacks more practical [47].

Currently, physical attacks in the field of image classification primarily focuses on tasks such as facial recognition and road sign recognition [50–52]. In [50], novel adversarial patches are fabricated, which involves designing stickers' positions and rotation angles to make physically viable and inconspicuous adversarial attacks. [51] emphasizes the importance of the content and position of patches, treating them as variables to be optimized simultaneously, using a reinforcement learning framework to find the optimal solution. Additionally, [52] uses shadow for stealthy and non-contact adversarial attacks on traffic signs. Physical adversarial attack methods are gradually evolving towards being black-box, concealed, and non-contact. Chapter four of this paper provides more attack methods in real-world scenarios.

3.5. Classification based on generative model

Owning to powerful generative capabilities, generative model has become an ideal tool for generating synthetic samples that conform to a certain data distribution. Generative models include Generative Adversarial Networks (GANs), Autoencoders, Normalizing Flows, Diffusion Models (DMs), etc. Nowadays, generative model has applied in various fields, certainly including the field of adversarial attacks. This section introduces generative model based black-box adversarial attacks.

3.5.1. GAN based black-box adversarial attacks

Generative Adversarial Networks (GANs) [53,54] consist of a generator and a discriminator. Generator continuously produces higher-quality and more realistic images, as Discriminator attempts to distinguish between generated images and real images. Through the adversarial learning process between the generator and the discriminator, deep representations can be learned without the need for a large amount of labeling, resulting in the generation of high-quality images that conform to the distribution of real images. Nowadays, GANs are widely applied in various fields, including adversarial attacks.

One application of GANs in adversarial attacks is to directly generate adversarial samples. Xiao et al. [55] use the generator of GAN to generate perturbations for any original instances and dynamically train the distillation model of the victim model. [56,57] utilize the latent features of input image for adversarial training. Once the training is complete, the generator can generate corresponding adversarial perturbations for any instance.

Another application of GANs in adversarial attacks is to address the issue of unavailability of original training datasets. Zhou et al. [12] propose to use a multi-branch GAN to generate data for training surrogate model, regarded as discriminator. This approach simultaneously generates data and training surrogate model, which subsequently serves as the basis for generating adversarial samples. To simplify the

architecture of generator and improve attack accuracy, Yu et al. [13] adopt a single-branch generator and design loss functions to alleviate mode collapse in GAN. Zhu et al. [14] further estimate gradient of target models and incorporate it into training. Authors of [10] address the training convergence issue by redesigning the attack framework.

3.5.2. Auto-encoder based black-box adversarial attacks

Auto-Encoders [58] consist of an encoder and a decoder. During the training process, the encoder encodes input data into low-dimensional representation, and the decoder decodes the representation back into the original data. By minimizing the error between the decoded image and the original one, the model is forced to learn important features and eliminate irrelevant and redundant data. Therefore, autoencoders can be used for tasks such as data dimensionality reduction, feature representation, image generation, data denoising etc.

Auto-Encoders also can be used for dimensionality reduction or feature extraction in adversarial attack. For example, in [22], an encoder is used to reduce the dimensionality of input image and obtain a feature space. Then, a pre-trained decoder is used to generate high-dimensional adversarial perturbations based on the feature space, which reduces the number of queries required for score-based attacks. In [46], autoencoder is used to ensure that the training procedure is on the feature space.

3.5.3. Normalizing flow based black-box adversarial attacks

Normalizing Flow [59] is a series of invertible mappings that can transform simple probability distributions into more complex and expressive probability distributions. It is commonly used in generative models, reinforcement learning, variational inference, and other fields. In the context of black-box adversarial attacks, Mohaghegh et al. [60] utilize normalizing flow to model the density of adversarial samples around the target image, which generates adversarial samples, closely following the data distribution of clean images, which enhances the stealthiness of adversarial attacks.

3.5.4. Diffusion model based black-box adversarial attacks

In recent years, there is a growing interest in DMs, which boast image generation and discrimination capabilities comparable to GANs. As a probabilistic generative model, Diffusion Models (DMs) learn the underlying data distribution by destroying and reconstructing image. There are three main types of DM structures [61], namely DDPM [62], NCSN [63], and SDE [64]. They all consist of a forward diffusion process and a reverse denoising process. The forward diffusion process adds noise to image at each step according to predefined hyperparameters until the training image closely meet Gaussian distribution. The denoising process then gradually removes the noise from noisy image, transforming them back into image that conform to the distribution of training data.

Chen et al. [46] are the first to utilize DM as a substitute model to generate imperceptible adversarial samples. Their attack framework is illustrated in Fig. 3.3. This approach involves encoding image using an encoder, training and optimizing framework in the latent space, and then converting the feature image into pixel-wise image through a decoder. Initially, the image undergoes a forward diffusion process, which is showed in the Eq. 15, where $Inverse(\bullet)$ is to transform it into a noisy image.

$$x_t = Inverse(x_{t-1}) = \underbrace{Inverse \circ \cdots \circ Inverse}_{t}(x_0)\# \tag{15}$$

Optimizing a loss function (Eq. 16) during the reverse process induces model to misclassify.

$$\underset{x_t}{\operatorname{argmin}} \mathscr{L}_{attack} = -J\left(x'_0, y; G_\phi\right), where x'_0 = \underbrace{Denoising \circ \cdots \circ Denoising}_{t}(x_t)\# \tag{16}$$

$J(\bullet)$ represents the cross-entropy loss, G_ϕ represents the surrogate model DM, and $Denoise(\bullet)$ signifies the denoising process within diffusion model. Given the robust recognition capabilities demonstrated by the pretrained diffusion model, the authors of [46] posit that misleading this surrogate model can impart strong transferability to adversarial samples. To achieve this, the following loss function (Eq. 17) is designed, where $Var(\bullet)$ calculates the variance of input, $Cross(\bullet)$ denotes the accumulation of all cross-attention maps during denoising process, C represents the name of training image category, and SDM stands for the Stable Diffusion Model. Minimizing this loss function leads the diffusion model to disperse its attention across the training images, disrupting the original strong semantic relationships, which achieves the deception of DM.

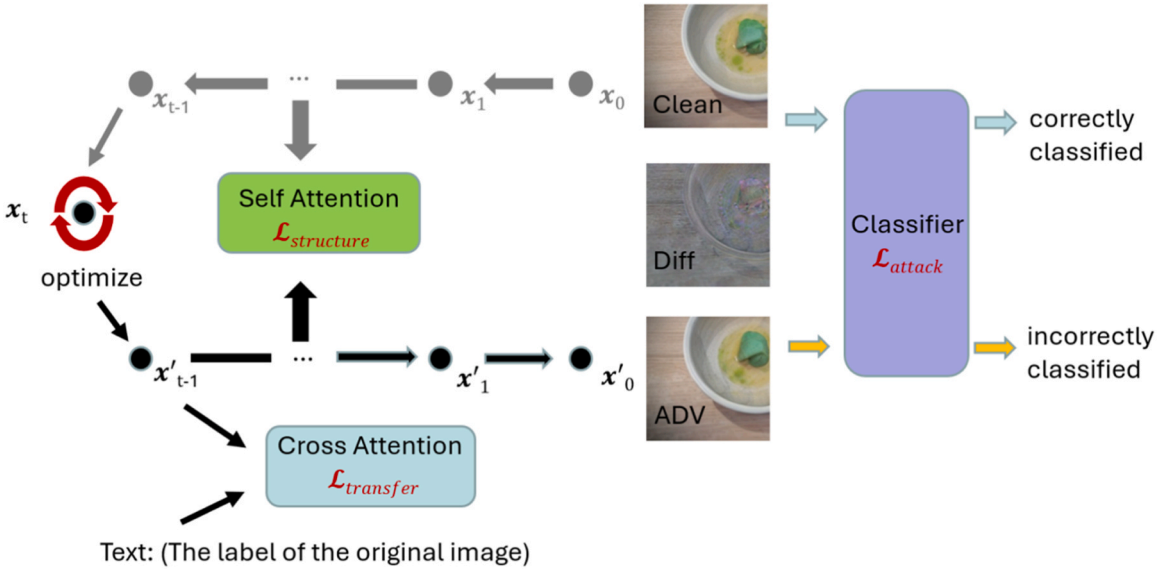


Fig. 3.3. The schematic diagram of Diffusion Models for imperceptible and transferable adversarial attack [46] with stable diffusion [95] and DDIM [21] employed. The training process occurs in the latent space through both forward diffusion and reverse inference processes. Self-attention mechanisms are utilized to preserve structural information of images, while cross-attention mechanism is applied to train DM as a substitute model. An effective generation of adversarial samples is achieved through the well-designed loss function.

$$\underset{x_t}{\operatorname{argmin}} \mathscr{L}_{transfer} = Var(Average(Cross(x_t, t, C; SDM)))\# \tag{17}$$

To ensure that the adversarial samples obtained after denoising is structurally consistent with the original samples, the authors of [46] copy an undisturbed image $x_{t(fix)}$. They ensure the similarity of self-attention maps between $x_{t(fix)}$ and x_t, which are $S_{t(fix)}$ and S_t, at each denoising step, as shown in Eq. 18, and applies this constraint in every denoising step.

$$\underset{x_t}{\operatorname{argmin}} \mathscr{L}_{structure} = \left\| S_t - S_{t(fix)} \right\|_2^2 \# \tag{18}$$

Furthermore, this work also controls the inversion strength of DDIM, retaining only the former denoising steps that add high-level semantic information. This reduction in inversion steps provides ample room for attacks.

The final objective function is the sum of $\mathscr{L}_{attack}$, $\mathscr{L}_{transfer}$, and $\mathscr{L}_{transfer}$ with different weighting factors (Eq. 19).

$$\underset{x_t}{\operatorname{argmin}} \mathscr{L} = \alpha \mathscr{L}_{attack} + \beta \mathscr{L}_{transfer} + \gamma \mathscr{L}_{structure} \# \tag{19}$$

By optimizing the final objective function, DM can be trained to generate adversarial samples that are imperceptible to human eye and possess strong transferability.

Generating high-quality, imperceptible, and effective adversarial samples is always the goal of adversarial attacks, and generative models become effective tools for generating adversarial samples. With technological advancements, increasingly powerful generative models are being proposed, and utilizing these emerging generative models to generate adversarial perturbations or assist in adversarial attacks remains an active and highly promising research area.

4. Black-box adversarial attacks in image security sensitive domains

With the widespread application and deployment of DNN in various industries, black-box adversarial attacks are no longer limited to laboratory settings but pose a serious security threat to the production and life of human, especially in security sensitive domains, such as facial recognition systems, medical image detection, traffic sign recognition. This section presents black-box attack methods in these domains and discusses the robustness evaluation of these models.

4.1. Face recognition

Face recognition is a common application of artificial intelligence in our daily lives, such as automatic payments and access control systems. Although the recognition process is similar with typical image classification tasks, face recognition is often treated separately due to the unique characteristics of face data and the vulnerability to adversarial attacks [65].

Physical attacks targeting face recognition systems tend to use inconspicuous perturbations, such as accessories, to avoid raising suspicion. They also emphasize the imperceptibility to the human eye. Yet, digital attacks are more inclined to leverage specific facial features, such as the position of the eyes and eyebrows, to generate more effective and transferable adversarial examples. In the field of digital attacks, Dong et al. [66] conduct a decision boundary-based attack on face recognition models. They propose an evolutionary attack algorithm that models the local geometry of the search direction and reduce the dimension of the search space to enhance the efficiency of the attack. Authors of [67,68] construct adversarial samples based on the features of image to improve the transferability of the attacks. In particular, Jia et al. [67] perturb on the high-level semantics. They utilize StyleGAN as a facial editing tool and facial attribute vectors as the basis for perturbation. They also propose an important-aware attribute selection strategy to select which facial attributes to perform while preserving the visual appearance of the face. In the field of physical adversarial attacks, the authors of [69,70] respectively construct adversarial perturbations by printing eyeglass frames and using visible light. The authors of [71] propose a method that combines the smoothness loss function and patch-noise for physical attack.

4.2. Medical image detection

Medical imaging [72] is a collection of images that reflects human cells, tissues, and pathological specimens, serving as a vital basis for current medical diagnosis. The accurate processing of medical images is of utmost importance for individual health and even person safety. DNNs gain maturity in the field of medical image processing, encompassing image detection, segmentation, registration, and fusion. Among these, DNN-based image detection refers to training deep learning models using annotated datasets and utilizing these trained models for diagnosis and classification. It represents a specialized form of image classification task. For instance, Esteva et al. [73] employ DNN for early diagnosis and classification of skin cancer.

In the field of medical image detection, due to privacy issues, there is a lack of public standardized datasets for model training, which may lead to overfitting problems. Additionally, standardized medical images exhibit similar characteristics such as fixed organ positions, exposure levels, and backgrounds, making them more susceptible to adversarial attacks compared to regular images. Moreover, some inherent noise is formed in medical images due to equipment parameters, camera exposure [78] and other reasons. All of these make adversarial attacks on medical images show strong transferability across different models. Due to the high demand for accuracy in medical image detection and diagnosis, the robustness of models used against adversarial attacks garner considerable attention from researchers.

The authors of [74,75] demonstrate the vulnerability of medical image classification models to adversarial attacks by applying perturbations to the images using white-box and black-box methods. Most black-box adversarial attacks in the present rely on substitute models, which are used to generate transferable adversarial examples. Pranava et al. [76] perform transfer-based attack on Diabetic Retinopathy 2015 Data Colored Resized and SARS-CoV-2 CT Scan Dataset. Koga et al. [77] utilize SimBA to generate UAPs for medical image classification, exhibiting promising attack effectiveness.

4.3. Traffic sign recognition (TSR)

TSR is a crucial component in autonomous driving and driver assistance system, primarily implemented using deep learning models. It involves capturing road images through cameras and sensors, employing computer vision and deep learning techniques to detect and identify traffic signs on the road. This provides essential information for autonomous driving system or driver's decision, including speed limits, sharp turns, danger warnings, etc. Clearly, the safety of autonomous driving is closely related to the accuracy of TSR.

[7] initially applies black-box adversarial attacks to the TSR task, conducting experiments on the German Traffic Sign Recognition and Detection Benchmark (GTSRB and GTSRD) datasets, achieving generic black-box adversarial attacks. [23] randomly samples a vector from a predefined orthonormal basis to generate adversarial samples. Subsequently, Kumar et al. [79] propose a "multi-gradient" Modified Simple Black-box Attack (M-SimBA) in DNN model for traffic scene perception.

While generic black-box adversarial attacks achieve success in traffic sign datasets like GTSRB, it is not realistic for attackers to iteratively perturb each frame for a high-speed driving car. Additionally, different regions of traffic sign image contribute differently to algorithm's decisions. Thus, generic black-box adversarial attacks are often ineffective in the field of TSR. To address these issues, the authors of [80] employ a soft attention map to highlight those important pixels within a set of training data. They then optimize a UAP based on this information.

Similarly, in [81], the authors collected generic attention patterns from other white-box DNNs to facilitate the transferability of the attack, ultimately achieving a successful black-box attack.

Adversarial attacks on TSR can be applied in real-world environment. Therefore, Woitschek et al. [82] firstly combine a general framework for physical attacks with various black-box attack methods and study the impact of these methods on the success rate of the attack under the same setting. To make these attacks more inconspicuous, Zhong et al. [52] utilize a common natural phenomenon, shadow, to generate adversarial perturbations. This approach enables them to achieve naturalistic and stealthy physical-world adversarial attacks.

In the field of TSR, physical attacks can pose significant harm to lots of vehicles and drivers. Therefore, compared to digital attacks, they are more potentially hazardous. Therefore, there is a need for further research and exploration in the area of physical attacks to understand their potential risks and develop effective countermeasures.

4.4. Other domains

In addition to image classification, there are many other security sensitive domains that are susceptible to the harm of black-box adversarial attacks. These domains include speech recognition, financial fraud detection, malicious software detection, and malicious network traffic detection, etc. This section provides a brief overview of representative black-box adversarial attacks in these domains.

In Automatic Speech Recognition (ASR) systems, attackers can generate adversarial audio samples to mislead the model, potentially allowing attackers to execute malicious commands. Alzantot et al. [83] conduct black-box adversarial attacks on speech classification models by adding small background noise. Taori et al. [84] employ genetic algorithms and gradient estimation for ASR system attacks, while Biolková et al. [85] use a Neural Predictor to estimate minimal perturbations and reduce the number of queries required for the attack. Tong et al. [86] introduce Temporal Natural Evolution Strategies (T-NES) for gradient estimation, resulting in more effective audio perturbation generation.

In the field of financial fraud detection, adversarial samples allow attackers to evade detection and execute fraudulent transactions. While the security of financial fraud detection systems is crucial, research on adversarial attacks in this domain is relatively limited. Authors of [87] assess the effectiveness of various black-box adversarial attacks against credit card fraud detection classifiers and propose an evolutionary algorithm-based method that reduced the number of queries. Additionally, one reason for the limited research in this field is the unique property of financial data. Financial data is heterogeneous and has strong dependencies between features, making it challenging to construct adversarial samples. To address this issue, authors of [88] evaluate the robustness of different classifiers to small perturbations in tabular data and proposed a novel attack through mathematical operations on the features.

In malware detection and network intrusion detection systems (NIDS), adversarial samples enable attackers to conceal malicious software and evade detection. Hu et al. [89] train a substitute detector fit the black-box malware detection system and GAN to minimize the probability of adversarial samples being detected by the substitute detector. Network intrusion detection systems (NIDS) plays a crucial role in network security. Zhu et al. [90] apply GAN to generate adversarial samples and insert synthetic packets into malicious traffic to bypass NIDS detection. By analyzing features of Denial of Service (DoS) samples, Peng et al. [91] propose a method to optimizes a Mahalanobis distance by perturbing the continuous features and discrete features of DoS samples respectively.

Due to the vulnerability of DNN to adversarial attacks, there is a need for further evaluation of the robustness of DNN in various security sensitive domains. It is also essential to conduct domain-specific research based on the characteristics of the data and the types of models in each domain. This research can be helpful in building more robust and secure detection systems and classification models.

5. Future directions

Deep learning models show inherent vulnerabilities to adversarial attacks, leading to extensive research in this area. Black-box adversarial attacks, which are more realistic and practical, gain particular attention. However, due to the lack of knowledge about the target model, most black-box adversarial attacks exhibit transferability only among models with similar structure. Improving transferability is always an important means for the success of black-box attacks and remains a future research direction. Perturbing image samples in feature space is proved to be more effective and concealed, and therefore will be the main attack method. Additionally, sensitive domains such as medical image detection, despite the widespread adoption of DNN, still lack sufficient research on adversarial attacks. This results in a lack of effective evaluation methods for model security and robustness. Therefore, black-box adversarial attacks in these specialized domains are an essential direction for future studies. Considering the complexities of real-world attack scenarios, investigating generic DNN attack methods for different data distributions and tasks is crucial for addressing the challenges of adversarial attacks comprehensively.

6. Conclusion

This article classifies black-box adversarial attacks from different perspectives including perturbation types, attack requirements, perturbation space, attack scenarios, and generators. It provides an introduction and summary of the latest research in these fields. Meanwhile, the applications of black-box adversarial attacks in various security sensitive domains are explored, such as face recognition, medical image recognition, and traffic sign recognition, etc. By presenting different methods and application scenarios, the paper offers a comprehensive and systematic overview of black-box adversarial attacks. Additionally, several potential directions for future research are suggested.

CRediT authorship contribution statement

Yaochi Zhao: Writing – review & editing, Writing – original draft, Supervision. **Yanfei Zhu:** Writing – review & editing, Writing – original draft, Validation, Investigation. **Zhuhua Hu:** Writing – review & editing, Supervision. **Like He:** Writing – review & editing, Investigation. **Tan Luo:** Writing – review & editing.

Declaration of Competing Interest

All authors disclosed no relevant relationships.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62161010, 62361024), the Natural Science Foundation of Hainan Province (623RC446), the Key Research and Development Project of Hainan Province (ZDYF2024GXJS021, ZDYF2022GXJS348), China, the Research Project on Reform of Higher Education and Teaching in Hainan Province, China (Hnjg2022-31).

References

[1] Y. LeCun, B. Boser, J.S. Denker, et al., Backpropagation applied to handwritten zip code recognition[J], Neural Comput. 1 (4) (1989) 541–551.
[2] Y. LeCun, L. Bottou, G.B. Orr, et al., Efficient backprop[M]//Neural networks: Tricks of the trade, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 9–50.

[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks[J], Adv. Neural Inf. Process. Syst. (2012) 25.
[4] I. Goodfellow, Explain. Harness Advers. Ex. arXiv Prepr. arXiv 1412 (2014) 6572.
[5] Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582. IEEE, New York (2016).
[6] Carlini N., Wagner D. Towards evaluating the robustness of neural networks[C]// 2017 ieee symposium on security and privacy (sp). Ieee, 2017: 39-57.
[7] Papernot N., McDaniel P., Goodfellow I., et al. Practical black-box adversarial attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017: 506-519.
[8] I.J. Goodfellow, J. Shlens, C. SzegedyExplaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014. 1412.
[9] N. Narodytska, S.P. Kasiviswanathan, Simple Black-Box Adversarial Attacks on Deep Neural Networks[C], CVPR Workshops 2 (2017) 2.
[10] Zhang J., Li B., Xu J., et al. Towards efficient data free black-box adversarial attack [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15115-15125.
[11] Wang X., He K. Enhancing the transferability of adversarial attacks through variance tuning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1924-1933.
[12] Zhou M., Wu J., Liu Y., et al. Dast: Data-free substitute training for adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 234-243.
[13] M. Yu, S. Sun, FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks[J], Comput. Secur. 113 (2022) 102555.
[14] Zhu Y., Zhao Y., Hu Z., et al. Zeroth-Order Gradient Approximation Based DaST for Black-Box Adversarial Attacks[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2023: 442-453.
[15] Xiong Y., Lin J., Zhang M., et al. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14983-14992.
[16] Xie C., Zhang Z., Zhou Y., et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2730-2739.
[17] Byun J., Cho S., Kwon M.J., et al. Improving the transferability of targeted adversarial examples through object-based diverse input[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15244-15253.
[18] Zhang J., Wu W., Huang J., et al. Improving adversarial transferability via neuron attribution-based attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14993-15002.
[19] Z. Li, W. Wu, Y. Su, et al., CDTA: A Cross-Domain Transfer-Based Attack with Contrastive Learning[C], Proc. AAAI Conf. Artif. Intell. 37 (2) (2023) 1530–1538.
[20] P.Y. Chen, H. Zhang, Y. Sharma, et al., Zoo: Zeroth order optimization based black-box adversarial attacks to deep neural networks without training substitute models [C], Proc. 10th ACM Workshop Artif. Intell. Secur. (2017) 15–26.
[21] Song J., Meng C., Ermon S. Denoising Diffusion Implicit Models[C]//International Conference on Learning Representations. 2020.
[22] Tu C.C., Ting P., Chen P.Y., et al. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 742-749.
[23] C. Guo, J. Gardner, Y. You, et al., Simple black-box adversarial attacks[C]// International Conference on Machine Learning, PMLR (2019) 2484–2493.
[24] M. Andriushchenko, F. Croce, N. Flammarion, et al., Square attack: a query-efficient black-box adversarial attack via random search[C]//European conference on computer vision, Springer International Publishing, Cham, 2020, pp. 484–501.
[25] Brendel W., Rauber J., Bethge M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models[C]//International Conference on Learning Representations. 2018.
[26] Moosavi-Dezfooli S.M., Fawzi A., Fawzi O., et al. Universal adversarial perturbations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1765-1773.
[27] C. Zhang, P. Benz, C. Lin, A survey on universal adversarial attack[J]. arXiv preprint arXiv:2103.01498, 2021..
[28] J. Wu, M. Zhou, S. Liu, Decision-based universal adversarial attack[J]. arXiv preprint arXiv:2009.07024, 2020..
[29] Zhang C., Benz P., Imtiaz T., et al. Understanding adversarial examples from the mutual influence of images and perturbations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14521-14530.
[30] C. Zhang, P. Benz, A. Karjauv, et al., Data-free universal adversarial perturbation and black-box attack[C]//Proceedings of, IEEE/CVF Int. Conf. Comput. Vis. (2021) 7868–7877.
[31] Li Y., Bai S., Xie C., et al. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer International Publishing, 2020: 795-813.
[32] A.S. Hashemi, A. Bär, S. Mozaffari, Hashemi A.S., Bär A., Mozaffari S., et al. Transferable universal adversarial perturbations using generative models[J]. arXiv preprint arXiv:2010.14919, 2020..
[33] Wang X., He X., Wang J., et al. Admix: Enhancing the transferability of adversarial attacks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16158-16167.
[34] F. Yin, Y. Zhang, B. Wu, et al., Generalizable black-box adversarial attack with meta learning[J], IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[35] M. Cheng, T. Le, P.Y. Chen, et al., Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach[C], Int. Conf. Learn. Represent. (2018).
[36] Y. Shi, Y. Han, Q. Hu, et al., Query-efficient black-box adversarial attack with customized iteration and sampling[J], IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2226–2245.
[37] J. Chen, M.I. Jordan, M.J. Wainwright, Hopskipjumpattack: A query-efficient decision-based attack[C], in: ieee symposium on security and privacy (sp), 2020, IEEE, 2020, pp. 1277–1294.
[38] Vo V., Abbasnejad E.M., Ranasinghe D. QUERY EFFICIENT DECISION BASED SPARSE ATTACKS AGAINST BLACK-BOX DEEP LEARNING MODELS[C]// International Conference on Learning Representations. 2021.
[39] Modas A., Moosavi-Dezfooli S.M., Frossard P. Sparsefool: a few pixels make a big difference[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9087-9096.
[40] Croce F., Hein M. Sparse and imperceivable adversarial attacks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4724-4732.
[41] Z. He, W. Wang, J. Dong, Transferable Sparse Adversarial Attack[J]. arXiv preprint arXiv:2105.14727, 2021.
[42] Huang Z., Zhang T. Black-Box Adversarial Attack with Transferable Model-based Embedding[C]//International Conference on Learning Representations. 2019.
[43] Q. Xu, G. Tao, S. Cheng, Towards feature space adversarial attack[J]. arXiv preprint arXiv:2004.12385, 2020.
[44] H. Wang, C. Zhu, Y. Cao, et al., ADSAttack: An Adversarial Attack Algorithm via Searching Adversarial Distribution in Latent Space[J], Electronics 12 (4) (2023) 816.
[45] Cao Y., Zhu C., Wang H., et al. An Adversarial Attack Algorithm based on Edge-Sketched Feature from Latent Space[C]//2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2022: 723-728.
[46] J. Chen, H. Chen, K. Chen, Diffusion Models for Imperceptible and Transferable Adversarial Attack[J]. arXiv preprint arXiv:2305.08192, 2023..
[47] D. Wang, W. Yao, T. Jiang, A survey on physical adversarial attack in computer vision[J]. arXiv preprint arXiv:2209.14262, 2022.
[48] N. Nichols, R. JasperProjecting trouble: Light based adversarial attacks on deep learning classifiers[J]. arXiv preprint arXiv:1810.10337, 2018.
[49] Duan R., Mao X., Qin A.K., et al. Adversarial laser beam: Effective physical-world attack to dnns in a blink[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16062-16071.
[50] X. Wei, Y. Guo, J. Yu, Adversarial sticker: A stealthy attack method in the physical world[J], IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 2711–2725.
[51] X. Wei, Y. Guo, J. Yu, et al., Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks[J], IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[52] Y. Zhong, X. Liu, D. Zhai, et al., Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon[C]//Proceedings of, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2022) 15345–15354.
[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets[J], Adv. Neural Inf. Process. Syst. (2014) 27.
[54] A. Creswell, T. White, V. Dumoulin, et al., Generative adversarial networks: An overview[J], IEEE Signal Process. Mag. 35 (1) (2018) 53–65.
[55] C. Xiao, B. Li, J.Y. Zhu, et al., Generating adversarial examples with adversarial networks[C]. Proceedings of the 27th, International Joint Conference on Artificial Intelligence, 2018, pp. 3905–3911.
[56] S. Jandial, P. Mangla, S. Varshney, et al., Advgan++: Harnessing latent layers for adversary generation[C]//Proceedings of, IEEE/CVF Int. Conf. Comput. Vis. Workshops (2019), 0-0.
[57] Zhao Z., Dua D., Singh S. Generating Natural Adversarial Examples[C]// International Conference on Learning Representations. 2018.
[58] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks[J], Science 313 (5786) (2006) 504–507.
[59] Dinh L., Sohl-Dickstein J., Bengio S. Density estimation using real nvp[J]. arXiv preprint arXiv:1605.08803, 2016.
[60] Mohaghegh Dolatabadi, H. Erfani, S. Leckie, C. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows[J], Adv. Neural Inf. Process. Syst. 33 (2020) 15871–15884.
[61] F.A. Croitoru, V. Hondru, R.T. Ionescu, et al., Diffusion models in vision: A survey [J], IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[62] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models[J], Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[63] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution[J], Adv. Neural Inf. Process. Syst. (2019) 32.
[64] Song Y., Sohl-Dickstein J., Kingma D.P., et al. Score-based generative modeling through stochastic differential equations[J]. arXiv preprint arXiv:2011.13456, 2020.
[65] N. Akhtar, A. Mian, N. Kardan, et al., Advances in adversarial attacks and defenses in computer vision: A survey[J], IEEE Access 9 (2021) 155161–155196.
[66] Dong Y., Su H., Wu B., et al. Efficient decision-based black-box adversarial attacks on face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7714-7722.
[67] S. Jia, B. Yin, T. Yao, et al., Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition[J], Adv. Neural Inf. Process. Syst. 35 (2022) 34136–34147.
[68] F. Zhou, H. Ling, Y. Shi, et al., Improving the Transferability of Adversarial Attacks on Face Recognition With Beneficial Perturbation Feature Augmentation[J], IEEE Trans. Comput. Soc. Syst. (2023).

[69] Sharif M., Bhagavatula S., Bauer L., et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016: 1528-1540.
[70] M. Shen, Z. Liao, L. Zhu, et al., Vla: A practical visible light-based attack on face recognition systems in physical world[J], Proc. ACM Interact., Mob., Wearable Ubiquitous Technol. 3 (3) (2019) 1–19.
[71] Singh I., Araki T., Kakizaki K. Powerful physical adversarial examples against practical face recognition systems[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 301-310.
[72] S. Kaviani, K.J. Han, I. Sohn, Adversarial attacks and defenses on AI in medical imaging informatics: A survey[J], Expert Syst. Appl. 198 (2022) 116815.
[73] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks[J], Nature 542 (7639) (2017) 115–118.
[74] Paschali M., Conjeti S., Navarro F., et al. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer International Publishing, 2018: 493-501.
[75] Asgari Taghanaki S., Das A., Hamarneh G. Vulnerability analysis of chest X-ray image classification against adversarial attacks[C]//Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1. Springer International Publishing, 2018: 87-94.
[76] B.M.S.P. R, Anusree V., Sreeratcha B., et al. Analysis of the effect of black box adversarial attacks on medical image classification models[C]//2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT). IEEE, 2022: 528-531.
[77] K. Koga, K. Takemoto, Simple black-box universal adversarial attacks on deep neural networks for medical image classification[J], Algorithms 15 (5) (2022) 144.
[78] Y. Cheng, F. Juefei-Xu, Q. Guo, Adversarial exposure attack on diabetic retinopathy imagery[J]. arXiv preprint arXiv:2009.09231, 2020.
[79] K.N. Kumar, C. Vishnu, R. Mitra, et al., Black-box adversarial attacks in autonomous vehicle technology[C], in: IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2020, IEEE, 2020, pp. 1–7.
[80] X. Yang, W. Liu, S. Zhang, et al., Targeted attention attack on deep learning models in road sign recognition[J], IEEE Internet Things J. 8 (6) (2020) 4980–4990.
[81] Chi L., Msahli M., Memmi G., et al. Public-attention-based adversarial attack on traffic sign recognition[C]//2023 IEEE 20th Consumer Communications & Networking Conference (CCNC). IEEE, 2023: 740-745.
[82] Woitschek F., Schneider G. Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study[C]//2021 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2021: 481-487.
[83] Alzantot M., Balaji B., Srivastava M. Did you hear that? adversarial examples against automatic speech recognition[J]. arXiv preprint arXiv:1801.00554, 2018.
[84] Taori R., Kamsetty A., Chu B., et al. Targeted adversarial examples for black box audio systems[C]//2019 IEEE security and privacy workshops (SPW). IEEE, 2019: 15-20.
[85] M. Biolková, B. NguyenNeural Predictor for Black-Box Adversarial Attacks on Speech Recognition[J]. arXiv preprint arXiv:2203.09849, 2022.
[86] C. Tong, X. Zheng, J. Li, et al., Query-efficient black-box adversarial attacks on automatic speech recognition[J], IEEE/ACM Trans. Audio, Speech, Lang. Process. (2023).
[87] Kumar N., Vimal S., Kayathwal K., et al. Evolutionary adversarial attacks on payment systems[C]//2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2021: 813-818.
[88] Agarwal A., Ratha N.K. Black-Box Adversarial Entry in Finance through Credit Card Fraud Detection[C]//CIKM Workshops. 2021.
[89] Hu W., Tan Y. Generating adversarial malware examples for black-box adversarial attacks based on GAN[C]//International Conference on Data Mining and Big Data. Singapore: Springer Nature Singapore, 2022: 409-423.
[90] Y. Zhu, L. Cui, Z. Ding, et al., Black box attack and network intrusion detection using machine learning for malicious traffic[J], Comput. Secur. 123 (2022) 102922.
[91] Peng X., Huang W., Shi Z. Adversarial attack against dos intrusion detection: An improved boundary-based method[C]//2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019: 1288-1295.
[92] Li J., Ji R., Liu H., et al. Projection & probability-driven black-box attack[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 362-371.
[93] Croce F., Andriushchenko M., Singh N.D., et al. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(6): 6437-6445.
[94] Y. Bai, Y. Wang, Y. Zeng, et al., Query efficient black-box adversarial attack on deep neural networks[J], Pattern Recognit. 133 (2023) 109037.
[95] Rombach R., Blattmann A., Lorenz D., et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.