Boosting Decision-based Black-box Adversarial Attacks with Random Sign Flip

Weilun Chen[1,2], Zhaoxiang Zhang[1,2,3⋆], Xiaolin Hu[4], and Baoyuan Wu[5,6]

[1] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
[4] Tsinghua University
[5] The Chinese University of Hong Kong, Shenzhen
[6] Tencent AI Lab
{chenweilun2018, zhaoxiang.zhang}@ia.ac.cn, xlhu@mail.tsinghua.edu.cn, wubaoyuan1987@gmail.com

Abstract. Decision-based black-box adversarial attacks (decision-based attack) pose a severe threat to current deep neural networks, as they only need the predicted label of the target model to craft adversarial examples. However, existing decision-based attacks perform poorly on the l_∞ setting and the required enormous queries cast a shadow over the practicality. In this paper, we show that just randomly flipping the signs of a small number of entries in adversarial perturbations can significantly boost the attack performance. We name this simple and highly efficient decision-based l_∞ attack as Sign Flip Attack. Extensive experiments on CIFAR-10 and ImageNet show that the proposed method outperforms existing decision-based attacks by large margins and can serve as a strong baseline to evaluate the robustness of defensive models. We further demonstrate the applicability of the proposed method on real-world systems.

Keywords: adversarial examples, decision-based attacks

1 Introduction

Deep neural networks are susceptible to *adversarial examples* [6, 48, 51]. In terms of image classification, an imperceptible adversarial perturbation can alter the prediction of a well-trained model to any desired class [9, 44]. The effectiveness of these maliciously crafted perturbations has been further demonstrated in the physical world [4, 18, 30], leading to growing concerns about the security of widely deployed applications based on deep neural networks, especially in sensitive areas, *e.g.*, financial service, autonomous driving, and face verification. Developing

⋆ Corresponding Author

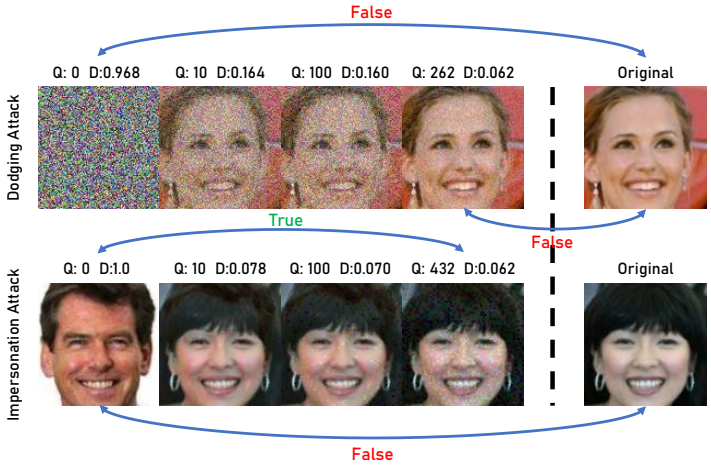


Fig. 1. Examples of attacking the face verification API in Tencent AI Open Platform. **Q** denotes the query number. **D** denotes the l_∞ distance towards the original image. True means the API classifies the two images as the same identity, and false means not the same. Best viewed in color with zoom in.

adversarial attacks under various settings provides great insights to understand and resistant the vulnerability of deep neural networks.

A broad range of existing works on adversarial attacks [21, 37] mainly focus on the *white-box* setting, where the adversary is capable to access all the information about the target model. While white-box attacks serve as an important role to evaluate the robustness, nearly all the real-world models are not completely exposed to the adversary. Typically, at least the model structures and internal weights are concealed. We refer to this case as the *black-box* setting. One approach to make black-box attacks feasible is to utilize the transferability of adversarial examples [33, 34, 38, 39, 46]. Whereas, these *transfer-based black-box* attacks need a substitute model and suffer from a low attack success rate when conducting targeted attacks. Instead of relying on the substitute model, some methods successfully conduct both untargeted and targeted attacks via accessing the confidence scores or losses of the target model [11, 27, 28, 36]. Despite the efficiency of these *score-based black-box* attacks, their requirements are still too hard in some real-world cases. Perhaps the most practical setting is that only the final decision (top-1 label) can be observed, noted as the *decision-based black-box* setting [7]. Even under this rather restrictive setting, deep neural networks are extremely vulnerable [7, 10, 12, 27]. However, the considerable query numbers required by existing decision-based attacks diminish their applicability, not to mention the poor performance on undefended and weak defensive models under the l_∞ setting. Designing a versatile and efficient decision-based l_∞ adversarial attack is still an open problem.

In this paper, we focus on the challenging decision-based black-box l_∞ setting and consider deep neural networks used for image classification as the models being attacked. Inspired by the cruciality of the sign vector of adversarial perturbations, we propose a simple and highly efficient attack method and dub **Sign Flip Attack**. Our method does not need to estimate gradients and works in an iterative manner. In each iteration, we first reduce the l_∞ distance by projection, then randomly flip the signs of partial entries in the current adversarial perturbation. The core of our method is the novel random sign flip step, which iteratively adjusts the sign vector of the current perturbation to get closer to a good one. Extensive experiments on two standard datasets CIFAR-10 [29] and ImageNet [15] demonstrate the superiority of our method. Results on 7 defensive models indicate that our method can serve as a strong baseline to evaluate the robustness. We further apply our method on real-world systems to show its practical applicability. Examples of attacking a face verification API are shown in Fig. 1.

2 Related Work

A variety of white-box attacks [9, 19, 21, 37, 40] have been developed since the vulnerability of deep neural networks discovered by Szegedy et al. [44]. The gradient-based nature of powerful white-box attacks is leveraged by quite a few defenses. However, Athalye et al. [3] has shown that most of them can be defeated. Defenses based on robust optimization [32, 35, 53] are the most effective, while still perform limited. In what follows, we will concentrate on the recent advances of black-box attacks.

Transfer-based black-box attacks. One intriguing property of adversarial examples is their good transferability [44]. Papernot et al. [39, 40] constructed a local surrogate model by querying the target model, then used the resulting one to generate adversarial examples. Liu et al. [34] showed that adversarial examples crafted on an ensemble of models have better transferability. Dong et al. [16] integrated momentum into the iterative white-box attacks and achieved a superior black-box attack performance. However, transfer-based black-box attacks are not effective to generate targeted adversarial examples [11] and perform poorly on Ensemble Adversarial Training [45]. Besides, it is arduous to find or train an appropriate substitute model for a real-world deployed application.

Score-based black-box attacks. In the score-based black-box setting, the adversary can obtain the corresponding predicted probability or loss by querying the target model. Chen et al. [11] applied zeroth order optimization to craft adversarial examples. A similar technique was used in [5], different from [11] they used the estimated gradient to perform the fast gradient sign method (FGSM) [21] and its iterative variant [30]. Ilyas et al. [27] used the natural evolutionary strategy (NES) for gradient estimation, then performed the projected gradient descent (PGD) [35]. The method can be further improved by exploiting two kinds of gradient priors [28]. Instead of gradient approximation, some attacks work in a gradient-free manner [1, 2, 22, 36]. Recently, several works [14, 20, 23, 31, 42]

consider additional models to improve query efficiency. Albeit effective, the applicability of these attacks is still restricted by their requirement of accessing the continuous outputs of the target model.

Decision-based black-box attacks. Brendel et al. [7] proposed the first effective decision-based black-box attack on deep neural networks, named the Boundary Attack, which starts from an adversarial point and performs random walks on the decision boundary while keeping adversarial. Ilyas et al. [27] extended their score-based method to this setting by estimating a proxy score. Cheng et al. [12] reformulated the original problem to a continuous version and applied zeroth order optimization. In [13], the same continuous problem was considered, however it computes the sign of the directional derivative instead of the magnitude, which leads to fast convergences. Recently, Chen et al. [10] proposed an unbiased estimate of the gradient direction at the decision boundary to improve the Boundary Attack. In each iteration, the adversarial example first approaches the boundary via a binary search, then moves along the estimated gradient direction to deviate from the decision boundary. In [17], an evolutionary attack method was proposed against face recognition systems. These methods generally require enormous queries and have poor performance under the l_∞ setting.

3 Approach

In this section, we first introduce preliminaries about adversarial examples and specify the threat model. Then, we present the proposed decision-based black-box l_∞ adversarial attack, which we dub **Sign Flip Attack** (SFA).

3.1 Preliminaries

We consider an image classifier $f : \mathbb{R}^d \to \mathbb{R}^k$ based on deep neural networks as the target model. For a given image $\boldsymbol{x} \in [0,1]^d$ and its corresponding true label y, $f(\boldsymbol{x})_i$ denotes the probability that $\boldsymbol{x}$ belongs to class i, and $c(\boldsymbol{x}) = \arg\max_{i\in\{1,\dots,k\}} f(\boldsymbol{x})_i$ refers to the predicted label. We only consider images that are correctly classified. The goal of the adversary is to find an adversarial perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ such that $c(\boldsymbol{x}+\boldsymbol{\delta}) \neq y$ (untargeted attacks) or $c(\boldsymbol{x}+\boldsymbol{\delta}) = t$ ($t \neq y$, targeted attacks) , and $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$. Here, ϵ refers to the allowed maximum perturbation. We choose l_∞ distance to depict the perceptual similarity between the natural and adversarial images.

Suppose we have a suitable loss function $L\big(f(\boldsymbol{x}), y\big)$, *e.g.*, cross entropy loss, then we can formulate the task of generating untargeted adversarial examples as a constrained optimization problem:

$$\max_{\boldsymbol{\delta}} L\big(f(\boldsymbol{x}+\boldsymbol{\delta}), y\big) \quad s.t.\ \|\boldsymbol{\delta}\|_\infty \leq \epsilon. \tag{1}$$

The constrained problem in Eq. 1 can be efficiently optimized by gradient-based methods under the white-box setting, such as PGD, which is an iterative method using the following update:

$$\boldsymbol{x}^n = \Pi_{B_\infty(\boldsymbol{x},\epsilon)}\big(\boldsymbol{x}^{n-1} + \eta \mathbf{sgn}(\nabla_{\boldsymbol{x}} L(f(\boldsymbol{x}^{n-1}), y))\big). \tag{2}$$

Here, $B_\infty(\boldsymbol{x}, \epsilon)$ refers to the l_∞ ball around $\boldsymbol{x}$ with radius ϵ and Π is the projection operator. Specifically, in this case, the projection is an element-wise clip:

$$\left(\Pi_{B_\infty(\boldsymbol{x},\epsilon)}(\boldsymbol{x}')\right)_i = \min\{\max\{\boldsymbol{x}_i - \epsilon, \boldsymbol{x}'_i\}, \boldsymbol{x}_i + \epsilon\}. \tag{3}$$

3.2 Threat Models

We consider adversarial attacks under the decision-based black-box setting, with l_∞ distance as the similarity constraint. That is, the adversary has no knowledge about the network architecture, internal weights, intermediate outputs or the predicted probability $f(\cdot)$, and can solely obtain the final decision $c(\cdot)$ of the target model by querying. As the value of $f(\cdot)$ can not be directly obtained, we consider the following constrained problem:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_\infty \quad s.t.\ \phi(\boldsymbol{x} + \boldsymbol{\delta}) = 1. \tag{4}$$

Here, $\phi : \mathbb{R}^d \to \{0, 1\}$ is an indicator function, which takes 1 if the adversarial constraint is satisfied, that is, $c(\boldsymbol{x}+\boldsymbol{\delta}) \neq y$ in untargeted attacks and $c(\boldsymbol{x}+\boldsymbol{\delta}) = t$ in targeted attacks. $\phi(\boldsymbol{x})$ can be computed by querying the target model. The goal of the adversary is to find a successful adversarial perturbation in as few queries as possible. Thus in practice, we set a maximum distortion ϵ. Once $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, we stop the attack and report the query number.

3.3 Sign Flip Attack

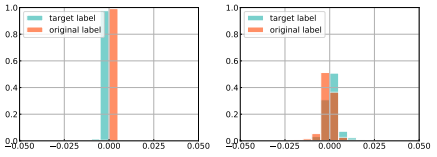
The basic logic of our method is simple. For an image-label pair $\{\boldsymbol{x}, y\}$, we start from a perturbation[1] $\boldsymbol{\delta}$ with a large l_∞ norm such that $\phi(\boldsymbol{x} + \boldsymbol{\delta}) = 1$, then iteratively reduce $\|\boldsymbol{\delta}\|_\infty$ while keeping adversarial. Next, we will mainly discuss the targeted version of the proposed method, and t is the target label. The extension to the untargeted setting is straightforward.

In each iteration, we first add a random noise $\boldsymbol{\eta}$ to the current perturbation $\boldsymbol{\delta}$, then project the new perturbation onto a smaller l_∞ ball, which can be formalized as:

$$\boldsymbol{\eta} \sim \{-\alpha, \alpha\}^d, \quad \boldsymbol{\delta}_p = \Pi_{B_\infty(0,\epsilon'-\alpha)}(\boldsymbol{\delta} + \boldsymbol{\eta}). \tag{5}$$

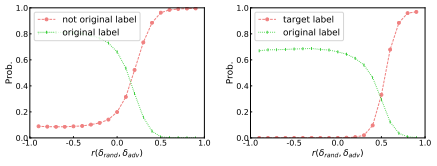
Here, $\epsilon' = \|\boldsymbol{\delta}\|_\infty$ and $0 < \alpha < \epsilon'$. $\boldsymbol{\delta}_p$ is the generated perturbation after the project step. α is an adjustable hyperparameter, which controls the shrinking magnitude of $\|\boldsymbol{\delta}\|_\infty$. Intuitively, the project step often leads to decreases in $f(\boldsymbol{x}+\boldsymbol{\delta})_t$ and increases in $f(\boldsymbol{x}+\boldsymbol{\delta})_y$, as the adversarial example gets closer to a natural image classified with a high probability. In Fig. 2 (a), we plot the distribution of the probability increments $\Delta_p f_t = f(\boldsymbol{x} + \boldsymbol{\delta}_p)_t - f(\boldsymbol{x} + \boldsymbol{\delta})_t$ and $\Delta_p f_y = f(\boldsymbol{x} + \boldsymbol{\delta}_p)_y - f(\boldsymbol{x} + \boldsymbol{\delta})_y$, it shows that $\Delta_p f_t$ is always negative and $\Delta_p f_y$ is always

[1] Finding an initial perturbation is easy, any image which is from a different class(untargeted attacks) or a specific class(targeted class) can be taken as an initial adversarial example.



(a) $\Delta_p f_y$ and $\Delta_p f_t$ (b) $\Delta_s f_y$ and $\Delta_s f_t$

Fig. 2. Distribution of the relative probability changes on the target and original label. We perform targeted SFA to 100 images from ImageNet on DenseNet-121. (a) the project step. (b) the random sign flip step. Note, we only consider the relative probability changes on successful trails, *i.e.*, $\phi(\boldsymbol{x}+\boldsymbol{\delta}_p) = 1$ and $\phi(\boldsymbol{x}+\boldsymbol{\delta}_s) = 1$. For comparison, the success rates for the random sign flip step and the project step are 24.2% and 49.3% respectively.



(a) untargeted (b) targeted

Fig. 3. Relationship between the sign match rate and the predicted probability. The experiments are conducted on 1,000 images from ImageNet. $\boldsymbol{\delta}_{adv}$ is an $\epsilon = 0.031$ l_∞ adversarial perturbation generated by 20-step PGD. $\boldsymbol{\delta}_{rand}$ is randomly selected from $\{-\epsilon, \epsilon\}^d$. "original label", "target label" and "not original label" denote $f(\boldsymbol{x}+\boldsymbol{\delta}_{rand})_y$, $f(\boldsymbol{x}+\boldsymbol{\delta}_{rand})_t$ and $\max_{i\neq y} f(\boldsymbol{x}+\boldsymbol{\delta}_{rand})_i$, respectively.

positive. If $f(\boldsymbol{x}+\boldsymbol{\delta}_p)_t$ is less than $f(\boldsymbol{x}+\boldsymbol{\delta}_p)_y$ or any other entry in $f(\boldsymbol{x}+\boldsymbol{\delta}_p)$, we reject $\boldsymbol{\delta}_p$ to hold the adversarial constraint.

As the project step always reduces $f(\boldsymbol{x}+\boldsymbol{\delta})_t$ and increases $f(\boldsymbol{x}+\boldsymbol{\delta})_y$, we expect to get a new perturbation using a single query, denoted by $\boldsymbol{\delta}_s$, which satisfies the following properties: 1) if $\boldsymbol{\delta}_s$ does not violate the adversarial constraint, it should have a high probability to acquire a positive $\Delta_s f_t$ and a negative $\Delta_s f_y$; 2) the l_∞ norm of $\boldsymbol{\delta}_s$ is equal to $\|\boldsymbol{\delta}\|_\infty$. It is not desirable to alleviate the probability changes introduced by the project step at the cost of a greater l_∞ norm.

Depicting the distribution of $\boldsymbol{\delta}_s$ is arduous. Fortunately, adversarial examples found by PGD frequently lie on the vertices of l_∞ balls around natural images [36] and one can modify the sign vector of a given adversarial perturbation to generate a new one [28]. These discoveries suggest that searching among the vertices of l_∞ balls may have a higher success rate than searching in the l_∞ balls. Our experiments conducted on CIFAR-10 support this claim, the untargeted attack success rates for these two random sampling strategies are 43.0% and 18.2% respectively. Thus, we conjecture that one has a high probability to get a qualified $\boldsymbol{\delta}_s$ by randomly changing the sign vector. Inspired by this, we propose the **random sign flip** step. In each iteration, we randomly select partial coordinates (*e.g.*, 0.1%), then flip the signs of the corresponding entries in $\boldsymbol{\delta}$. Suppose $\boldsymbol{s} \in \{0,1\}^d$ and $\boldsymbol{p} \in (0,1)^d$, the random sign flip step can be formulated as:

$$s_i \sim \textbf{Bernoulli}(p_i), \quad \boldsymbol{\delta}_s = \boldsymbol{\delta} \odot (1 - 2\boldsymbol{s}), \tag{6}$$

where s_i and p_i denote the i-th element of $\boldsymbol{s}$ and $\boldsymbol{p}$ respectively. $\odot$ is the Hadamard product. $\boldsymbol{p}$ is another crucial hyperparameter, which controls the sign flip probability of each entry in $\boldsymbol{\delta}$. We will discuss how to adjust these two hyperparameters later. Same as the project step, $\boldsymbol{\delta}_s$ which violates the adversarial constraint will be rejected. A simple illustration of the random sign flip

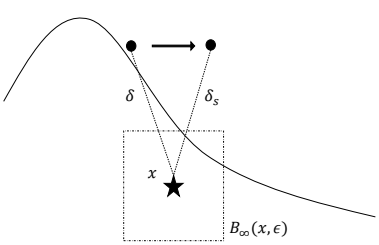


Fig. 4. Illustration of the random sign flip step. $\boldsymbol{\delta}_s$ is the generated perturbation of $\boldsymbol{\delta}$ after the random sign flip step, which also has a larger distance towards the decision boundary.

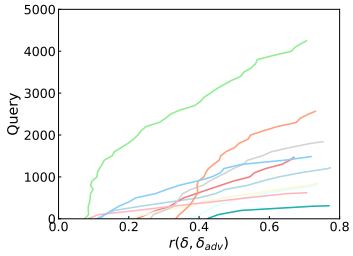


Fig. 5. The change of the sign match rate between $\boldsymbol{\delta}_{adv}$ and $\boldsymbol{\delta}$ during querying. $\boldsymbol{\delta}_{adv}$ is generated by targeted PGD with the last $\boldsymbol{\delta}$ as the initial start. Each curve represents an image chosen from CIFAR-10.

step is shown in Fig. 4 . Clearly, the random sign flip step does not alter the l_∞ norm, $\|\boldsymbol{\delta}_s\|_\infty = \|\boldsymbol{\delta}\|_\infty$. In practice, we plot the distribution of $\Delta_s f_t$ and $\Delta_s f_y$ in Fig. 2 (b). It can be seen that $\boldsymbol{\delta}_s$ generated by the random sign flip step does have a certain probability to get a positive $\Delta_s f_t$ and a negative $\Delta_s f_y$.

Why does the random sign flip step work? Let $r(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = \frac{\mathbf{sgn}(\boldsymbol{\delta}_1)^\top \mathbf{sgn}(\boldsymbol{\delta}_2)}{d}$ be the sign match rate of $\boldsymbol{\delta}_1 \in \mathbb{R}^d$ and $\boldsymbol{\delta}_2 \in \mathbb{R}^d$. We find that,

1) There exist a large body of diverse adversarial sign vectors for most images. We choose 1,000 images from ImageNet. For each image-label pair $\{\boldsymbol{x} \in \mathbb{R}^d, y \in \mathbb{R}\}$, we generate 100 adversarial perturbations using 20-step targeted PGD with random starts and calculate the maximum sign match rate between any two of them. The results show that the maximum sign match rate is lower than 0.1 for 87.8% of images (82.5% for untargeted attacks).
2) For an image-label pair $\{\boldsymbol{x}, y\}$, consider a random vector $\boldsymbol{\delta}_{rand} \in \{-\epsilon, \epsilon\}^d$ and a "good" adversarial perturbation[2] $\boldsymbol{\delta}_{adv} \in [-\epsilon, \epsilon]^d$, $r(\boldsymbol{\delta}_{rand}, \boldsymbol{\delta}_{adv})$ has a negative correlation with $f(\boldsymbol{x} + \boldsymbol{\delta}_{rand})_y$ and a positive correlation with $f(\boldsymbol{x} + \boldsymbol{\delta}_{rand})_t$ (or $\max_{i \neq y} f(\boldsymbol{x} + \boldsymbol{\delta}_{rand})_i$), as presented in Fig. 3.

As a large number of diverse adversarial sign vectors exist, we assume that there is a "good" adversarial perturbation $\boldsymbol{\delta}_{adv} \in \mathbb{R}^d$ which has a relatively high sign match rate with the current perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$. Although we do not know the actual $\boldsymbol{\delta}_{adv}$, we can alter the sign match rate between $\boldsymbol{\delta}_{adv}$ and $\boldsymbol{\delta}$ through the random sign flip step. Then according to the second point mentioned above, what we prefer is to get a new perturbation with a higher sign match rate. In fact, once our method attacks successfully, we can use the resulting perturbation as an initial start and perform targeted PGD to construct such a $\boldsymbol{\delta}_{adv}$. We plot the change of the sign match rate between $\boldsymbol{\delta}_{adv}$ and $\boldsymbol{\delta}$ during querying in Fig. 5. It shows a clear uptrend. Thus, the random sign flip step serves as a tool to

[2] $f(\boldsymbol{x} + \boldsymbol{\delta}_{adv})_t$ (or $\max_{i \neq y} f(\boldsymbol{x} + \boldsymbol{\delta}_{adv})_i$) is very close to 1.

make the sign match rate become higher. But how large is the probability of acquiring a higher sign match rate after the random sign flip step?

Let $r(\boldsymbol{\delta}, \boldsymbol{\delta}_{adv}) = r\,(r \in [-1, 1])$ and $m = (1-r) \cdot d/2$ denote the total number of coordinates where $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_{adv}$ have reverse signs. We perform one random sign flip step. That is, we randomly select $k\,(k \ll \min(m, d-m))$ coordinates and flip the corresponding signs in $\boldsymbol{\delta}$, resulting in a new vector $\boldsymbol{\delta}_s$. Then, we have

$$P(r(\boldsymbol{\delta}_s, \boldsymbol{\delta}_{adv}) > r) = \frac{\sum_{i=\lceil \frac{k+1}{2} \rceil}^{k} \binom{m}{i} \cdot \binom{d-m}{k-i}}{\binom{d}{k}}, \tag{7}$$

where $\binom{\cdot}{\cdot}$ is the binomial coefficient. It can be easily proven that $P(r(\boldsymbol{\delta}_s, \boldsymbol{\delta}_{adv}) > r)$ is smaller than $P(r(\boldsymbol{\delta}_s, \boldsymbol{\delta}_{adv}) < r)$ when r is larger than 0. Thus, it is no surprise that lots of $\boldsymbol{\delta}_s$ are rejected during the optimization as mentioned in Fig. 2. Even though, our method performs much better than existing methods, see Section 4 for detailed information.

The complete Sign Flip Attack (SFA) is summarized in Algorithm 1. SFA works in an iterative manner. Given a correctly classified image, we first initialize the perturbation which satisfies the adversarial constraint, then push the initial adversarial example close to the original image through a binary search. In each iteration, there are two steps requiring a total of 2 queries. One is the project step described in Eq. 5 , which reduces the l_∞ norm of the current perturbation. The other is the random sign flip step described in Eq. 6 , which has a relatively high probability to generate a better adversarial perturbation. We clip adversarial examples to a legitimate range before querying the target model, and reject unqualified perturbations. Next, we will discuss several techniques to improve query efficiency.

Dimensionality reduction. In general, attacking deep neural networks with high-dimensional inputs requires hundreds of thousands of queries under the black-box setting. Several works [2, 11] have demonstrated that performing dimensionality reduction can boost attack efficiency. We adopt this strategy into our method. To be specific, we define $\boldsymbol{p}$ and $\boldsymbol{\eta}$ in the lower dimensionality, *i.e.*, $\boldsymbol{p} \in (0, 1)^{d'}$, $\boldsymbol{\eta} \in \{-\alpha, \alpha\}^{d'}$, $d' < d$, and choose bilinear interpolation $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ as the upscaling function. Then,

$$\boldsymbol{\eta} \leftarrow T(\boldsymbol{\eta}), \quad s_i \sim \mathbf{Bernoulli}(p_i),\ \boldsymbol{s} \leftarrow \mathbf{sgn}(T(\boldsymbol{s})). \tag{8}$$

With a suitable d', we can obtain higher attack success rates under limited queries. Note, our method works well even without dimensionality reduction.

Hyperparameter adjustment. Our method has two hyperparameters α and $\boldsymbol{p}$ corresponding to two steps in each iteration. We dynamically adjust these two hyperparameters according to the success rate of several previous trails. For the project step, if the success rate is higher than 70%, we increase α by multiplying a fixed coefficient (*e.g.*, 1.5). If the success rate is lower than 30%, we reduce α by dividing the same coefficient. For the random sign flip step, if the success rate is higher than 70%, we increase $\boldsymbol{p}$ by adding a fixed increment (*e.g.*, 0.001). If the success rate is lower than 30%, we reduce $\boldsymbol{p}$ by subtracting the same

Algorithm 1 Sign Flip Attack

Input: indicator function ϕ, original image $\boldsymbol{x}$, threshold ϵ

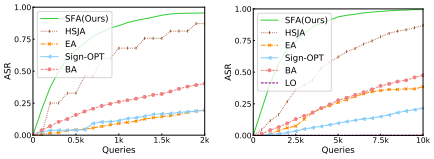
Initialize $\boldsymbol{\delta} \in \mathbb{R}^d$, $\alpha \in \mathbb{R}_+$, $\boldsymbol{p} \in (0,1)^d$;
$\boldsymbol{\delta} \leftarrow$ **BinarySearch**($\boldsymbol{x}$, $\boldsymbol{\delta}$, ϕ)
$\epsilon' = \|\boldsymbol{\delta}\|_\infty$
while $\epsilon' > \epsilon$ **do**
  $\boldsymbol{\eta} \sim \{-\alpha, \alpha\}^d$
  $\boldsymbol{\delta}_p = \Pi_{B_\infty(0, \epsilon' - \alpha)}(\boldsymbol{\delta} + \boldsymbol{\eta})$
  if $\phi(\boldsymbol{x} + \boldsymbol{\delta}_p) = 1$ **then**
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta}_p$
  end if
  $s_i \sim$ **Bernoulli**(p_i)
  $\boldsymbol{\delta}_s = \boldsymbol{\delta} \odot (1 - 2\boldsymbol{s})$
  if $\phi(\boldsymbol{x} + \boldsymbol{\delta}_s) = 1$ **then**
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta}_s$
  end if
  Adjust $\boldsymbol{p}$, α
  $\epsilon' = \|\boldsymbol{\delta}\|_\infty$
end while
return $\boldsymbol{x} + \boldsymbol{\delta}$

increment. In practice, we adjust α and $\boldsymbol{p}$ once every 10 iterations. Besides, we also adjust p after each successful random sign flip step by $\boldsymbol{p} \leftarrow \boldsymbol{p} + 0.0001(1 - 2\boldsymbol{s})$. In this way, we roughly adjust each entry in $\boldsymbol{p}$. We set $\alpha = 0.004$ (around one pixel $1/255$), and set the initial flip probability for each coordinate as $p_i = 0.001$. We bound $\boldsymbol{p} \in [0.0001, 0.01]^d$ in each step, as we only want to flip a rather small number of entries.

4 Experiments

We compare Sign Flip Attack (SFA) with a comprehensive list of decision-based black-box attacks: Boundary Attack (BA) [7], Label-Only Attack (LO) [27], HopSkipJumpAttack (HSJA) [10], Evolutionary Attack (EA) [17] and Sign-OPT Attack (Sign-OPT) [13]. We use the l_∞ version of each baseline method if it exists. For BA and EA, we use their original versions. All experiments are conducted on two standard datasets, CIFAR-10 [29] and ImageNet [15].

The maximum l_∞ distortion and limited query numbers will be specified in each part. For our method, we use the same hyperparameters (described in Section 3.3) across all images. For baseline methods, we use their default settings. We quantify the performance in terms of three dimensions: attack success rate (ASR), average queries (AQ) and median queries (MQ). Average queries and median queries are calculated over successful trails. For some experiments, we also provide the results achieved by strong white-box attacks, *e.g.*, 100-step PGD. Additional results are provided in the supplementary material.

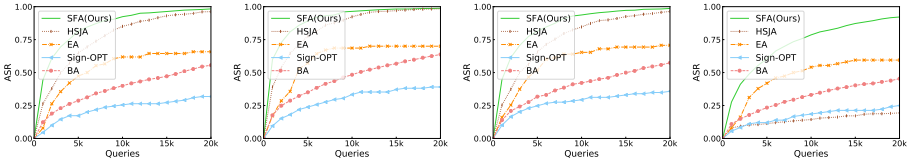


(a) (b)

Fig. 6. Comparison of the attack success rates over the number of queries of various methods on **CIFAR-10** with $\epsilon = 0.031$. (a) untargeted attacks. (b) targeted attacks.

Table 1. Results on **CIFAR-10**. The query limits for untargeted and targeted attacks are 2,000 and 10,000, respectively. For Lable-Only Attack (LO), we only consider targeted attacks, since it is mainly designed for this setting. **SFA wo/SF** indicates SFA without Random Sign Flip.

Method	Untargeted			Targeted		
	ASR	AQ	MQ	ASR	AQ	MQ
BA [7]	40.2%	809	660	47.5%	4,629	4,338
LO [27]	-	-	-	0.2%	3,533	3,533
Sign-OPT [13]	19.7%	886	730	21.5%	5,515	5,438
EA [17]	19.4%	1,037	1,076	38.0%	4,139	3,611
HSJA [10]	87.1%	680	503	86.9%	3,731	3,197
SFA wo/SF	9.1%	-	-	0.2%	-	-
SFA(Ours)	**95.4%**	**409**	**282**	**99.4%**	**1,807**	**1,246**



(a) (b) (c) (d)

Fig. 7. Untargeted attack success rates versus numbers of queries on **ImageNet** with four different model architectures. (a) ResNet-50. (b) VGG-16. (c) DenseNet-121. (d)Inception-v3.

4.1 Attacks on Undefended Models

CIFAR-10 setup. We use the first 1000 images of the validation set. 953 of them are correctly classified by our trained ResNet-18 [24]. For untargeted attacks, we set the maximum queries to 2,000. For targeted attacks, we set the maximum queries to 10,000. Following the protocol in [7], the target label is set to $t = (y+1) \bmod 10$ for an image with label y. To ensure a fair comparison, we use the same initial perturbation for all methods in targeted attacks. As a convention, We bound the maximum l_∞ distortion to $\epsilon = 0.031(8/255)$.

ImageNet setup. To verify the robustness of attack methods against different network architectures, we consider four prevailing model, ResNet-50 [24], VGG-16 [41], DenseNet-121 [25], and Inception-v3 [43]. For untargeted attacks, we randomly select 1,000 images and set the maximum queries to 20,000. For targeted attacks, we randomly select 200 images due to time concerns and set the maximum queries to 100,000. The target label is randomly chosen across 1,000 classes. Again, the same initial perturbation is applied for all methods for targeted attacks. We bound the maximum l_∞ distortion to $\epsilon = 0.031$. For our method, we apply dimensionality reduction described in Section 3.3 . We simply set $d' = d/4$, *e.g.*, $d = 224 \times 224 \times 3$, $d' = 112 \times 112 \times 3$.

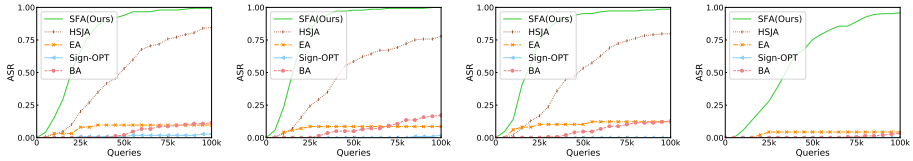


Fig. 8. Targeted attack success rates versus numbers of queries on **ImageNet** with four different model architectures. (a) ResNet-50. (b) VGG-16. (c) DenseNet-121. (d)Inception-v3.

We report the results for CIFAR-10 in Table 1. Untargeted and targeted attack results for ImageNet are shown in Table 2 and Table 3. The corresponding cumulative attack success rates of the number of queries are presented in Fig. 6, 7 and 8, respectively. Compared with existing methods, our method consistently achieves higher attack success rates across different datasets and various network architectures for both untargeted and targeted settings while requiring half or even fewer queries. In Table 3, we notice that the average and median queries of EA are better than ours. This is due to the average and median queries can be influenced by the attack success rates. As an instance, with Inception-v3 as the target model, our method achieves a 95.8% attack success rate, 91.5% higher than EA. On the 4.3% of images that EA attacks successfully, the average and median queries of our method are 6,657 and 5,604, respectively. Our method reduces more than 60% of queries compared to EA. A detailed comparison with EA and HSJA is provided in the supplementary material.

The importance of Random Sign Flip. We study the effect of the random sign flip step. In Table 1, it can be seen that the random sign flip step boosts the attack success rate. With the random sign flip step, the success rates for untargeted and targeted attacks are 95.4% and 99.4%. However, without it, the attack success rates drop to 9.1% and 0.2%, even lower than random sampling.

Dimensionality reduction helps attack efficiency. We also study the effect of dimensionality reduction. We provide the results without dimensionality reduction in Table 2 and Table 3. The results show that, with the help of dimensionality reduction, it is able to achieve a higher attack success rate with fewer queries. Note that the results achieved by our method without dimensionality reduction are still much better than existing methods.

4.2 Attacks on Defensive Models

To investigate the effectiveness of decision-based black-box attacks against defenses, we conduct experiments on 7 defense mechanisms: Adversarial Training [35], Thermometer Encoding [8], Bit Depth Reduction [50], FeatDenoise [49], FeatScatter [52], KWTA [47] and TRADES [53]. For the first 3 defense mechanisms, we compare our method with BA [7], Sign-OPT [13], EA [17] and HSJA

Table 2. Results of **untargeted** attacks on **ImageNet**.The maximum number of queries sets to 20,000. **SFA wo/DR** indicates SFA without dimensionality reduction.

Method	ResNet-50 [24]			VGG-16 [41]			DenseNet-121 [25]			Inception-v3 [43]		
	ASR	AQ	MQ	ASR	AQ	MQ	ASR	AQ	MQ	ASR	AQ	MQ
BA [7]	55.5%	6,547	4,764	63.7%	5,848	3,906	57.4%	6,285	4,170	45.3%	6,404	4,830
Sign-OPT [13]	31.8%	5,929	3,448	39.0%	4,984	3,624	35.7%	4,669	2,432	25.0%	6,548	5,153
EA [17]	65.7%	4,004	2,578	70.0%	3,016	2,886	70.6%	37,13	2,752	59.4%	4,247	2,836
HSJA [10]	96.1%	4,370	2,883	98.3%	3,044	1,554	96.3%	4,394	2,883	19.4%	5,401	2,538
SFA wo/DR	96.9%	3,193	1,570	**99.0%**	2,112	820	97.5%	2,972	1,652	91.1%	**3,759**	**2,134**
SFA(Ours)	**98.2%**	**2,712**	**1,288**	98.7%	**1,754**	**636**	**98.6%**	**2,613**	**1,200**	**92.1%**	4,501	2,602

Table 3. Results of **targeted** attacks on **ImageNet**. The maximum number of queries sets to 100,000. **SFA wo/DR** indicates SFA without dimensionality reduction.

Method	ResNet-50 [24]			VGG-16 [41]			DenseNet-121 [25]			Inception-v3 [43]		
	ASR	AQ	MQ	ASR	AQ	MQ	ASR	AQ	MQ	ASR	AQ	MQ
BA [7]	11.4%	62,358	57,336	17.1%	62,480	67,658	12.8%	58,879	56,646	3.4%	84,266	85,202
Sign-OPT [13]	2.7%	54,523	51,319	1.9%	91,172	82,492	0.0%	-	-	0.0%	-	-
EA [17]	9.6%	19,682	20,435	8.5%	12,126	8,534	12.2%	17,820	7,195	4.3%	18,164	15,362
HSJA [10]	84.5%	44,188	41,205	77.8%	39,400	36,172	79.7%	41,319	36,964	0.0%	-	-
SFA wo/DR	98.6%	29,440	26,208	98.5%	23,216	21,076	98.6%	28,151	25,824	**97.0%**	37,169	38,056
SFA(Ours)	**99.3%**	22,538	19,380	**99.2%**	16,627	15,008	**98.6%**	20,331	17,762	95.8%	36,681	32,210

[10]. For other defense mechanisms, we only compare with the existing state-of-the-art method HSJA [10].

The overall results are presented in Table 4 and Table 5. Because the average and median queries can be affected by the attack success rate for an effective method, we also report the results achieved by our methods on the images that each baseline method successfully fools, please see the values in brackets. Our method significantly outperforms the existing decision-based attacks by large margins on nearly all evaluation metrics. In what follows, we will make a detailed discussion.

Firstly, our method performs much more stable than state-of-the-art decision-based attacks when confronting defenses that cause obfuscated gradients [3]. Thermometer Encoding [8] and Bit Depth Reduction [50] are two types of defenses that leverage per pixel quantization to resistant adversarial examples. Although these two defenses have been completely broken by current white-box attacks, it is still difficult to defeat them for decision-based attacks. For untargeted attacks on Thermometer Encoding, our method achieves a 92.3% attack success rate, while the highest achieved among existing methods is a mere 34.1%.

Secondly, our method can serve as a strong baseline to evaluate the robustness of defensive models. FeatScatter [52] and KWTA [47] are two recently published defenses. Both of them have shown superior resistance against white-box attacks than Adversarial Training [35] and done certain sanity checks such as transfer-based black-box attacks. However, according to our results, they reduce the performance of the original Adversarial Training. As presented in Table 5, our method has 52.3% and 74.7% attack success rates, around 29% and 41% higher

Table 4. Attack performance against Adversarial Training [35], Thermometer Encoding [8], and Bit Depth Reduction [50]. The limited query budgets set to 100,000, 50,000 and 50,000, respectively. The values in brackets denote results achieved by our method on the images that baseline methods successfully fool.

Method	CIFAR-10 [29]						ImageNet [15]		
	Adv. Training [35]			Thermometer [8]			Bit Depth [50]		
	ASR	AQ	MQ	ASR	AQ	MQ	ASR	AQ	MQ
BA [7]	5.6%	1,811(1,294)	880(242)	11.0%	286(208)	134(122)	13.5%	165(173)	142(90)
Sign-OPT [13]	6.3%	5,908(324)	2,505(202)	8.0%	3,583(262)	1,285(122)	15.1%	2,284(154)	179(90)
EA [17]	12.1%	7,675(4,442)	1,594(1,462)	34.1%	5,254(752)	4,459(310)	75.3%	9,689(3,245)	5,626(858)
HSJA [10]	34.0%	14,829(7,694)	4,450(2,972)	13.1%	5,030(200)	310(132)	8.4%	6,664(243)	159(76)
SFA(Ours)	41.6%	15,512	5,486	92.3%	7,024	3,386	91.2%	7,227	2,100
BPDA [3](white-box)	**50.9%**	-	-	**100%**	-	-	**100%**	-	-

Table 5. Attack performance against TRADES [53], FeatScatter [52], KWTA [47] and FeatDenoise [49]. The limited query budgets set to 100,000. The values in brackets denote results achieved by our method on the images that HSJA successfully fools.

Method		PGD(white-box)	SFA(Ours)			HSJA [10]		
		ASR	ASR	AQ	MQ	ASR	AQ	MQ
CIFAR-10 [29]	TRADES [53]	**34.0%**	29.8%	7,714	2,470	25.2%	13,115(4,372)	3,569(1,844)
	FeatScatter [52]	23.2%	**52.3%**	10,254	3,956	42.0%	14,393(4,441)	5,222(2,318)
	KWTA+AT [47]	33.6%	**74.7%**	16,935	3,660	35.4%	9,953(2,568)	2,187(688)
ImageNet [15]	FeatDenoise [49]	**88.0%**	51.3%	18,616	7,848	44.0%	23,866(13,620)	13,151(5,478)

than PGD, respectively. Their model accuracies are actually lower than ones in Adversarial Training.

Thirdly, for those defensive models which indeed increase the robustness, our method obtains better results than other decision-based attacks while still falls behind of white-box attacks. On CIFAR-10, for Adversarial Training [35] and TRADES [53], our method is slightly worse than PGD. Whereas, on ImageNet, the gap between our method and PGD is quite large. PGD achieves an 88.0% attack success rate against FeatDenoise [49], around 31% higher than our method. Attacking models with high dimensional inputs is arduous for *all* decision-based attacks. Our method takes a steady step towards closing the gap.

4.3 Attacks on Real-world Applications

In this section, we investigate the applicability of decision-based attacks on real-world systems.

The target models are the face verification[3] API and food[4] API in Tencent AI Open Platform. For the face verification API, we set the similarity score threshold to 70. If the output is larger than 70, the decision is True — the two images are from the same identity, otherwise False. We choose 10 pairs of images from the Labeled Face in the Wild [26] dataset. We bound the maximum

[3] https://ai.qq.com/product/face.shtml#compare
[4] https://ai.qq.com/product/visionimgidy.shtml#food

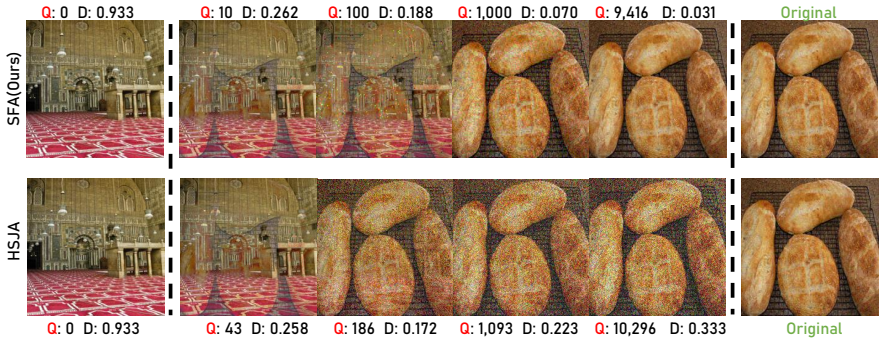


Fig. 9. An example of attacking the food API in Tencent AI Open Platform. **Q** denotes the query number. **D** denotes the l_∞ distance towards the original image. Green means the API classifies the image as food, red is otherwise. Best viewed in color with zoom in.

l_∞ distortion to $\epsilon = 0.062(16/255)$ and set the maximum query number to 5,000. The numbers of successfully attacked pairs of BA, Sign-OPT, EA, LO, HSJA, and our method are 1, 0, 5, 0, 2 and 9, respectively. The food API takes a single image as input and determines whether the input is about food. Our method successfully invades this API. We present an example in Fig. 9. The original and initial images are chosen from ImageNet.

5 Conclusion

In this paper, we proposed the Sign Flip Attack, a simple and highly efficient decision-based black-box l_∞ adversarial attack to craft adversarial examples. We introduced the novel random sign flip step to search for a better adversarial perturbation during the optimization, boosting the attack success rate and query efficiency. Comprehensive studies on CIFAR-10 and ImageNet demonstrate that our method has significantly outperformed existing methods. Experiments on several defensive models indicate the effectiveness of our method in evaluating the robustness. Additionally, we applied our method to attack real-world applications successfully. These promising results suggest that our method can be viewed as a strong baseline to facilitate future research.

Acknowledgement

This work was supported in part by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (No. 61836014, No. 61761146004, No. 61773375).

References

1. Al-Dujaili, A., O'Reilly, U.M.: Sign bits are all you need for black-box attacks. In: Proceedings of International Conference on Learning Representations (2020)
2. Alzantot, M., Sharma, Y., Chakraborty, S., Srivastava, M.: Genattack: Practical black-box attacks with gradient-free optimization. arXiv preprint arXiv:1805.11090 (2018)
3. Athalye, A., Carlini, N., Wagner, D.A.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proceedings of International Conference on Machine Learning (2018)
4. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397 (2017)
5. Bhagoji, A.N., He, W., Li, B., Song, D.X.: Practical black-box attacks on deep neural networks using efficient query mechanisms. In: Proceedings of the European Conference on Computer Vision (2018)
6. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: Proceedings of the Joint European conference on machine learning and knowledge discovery in databases. pp. 387–402. Springer (2013)
7. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: Proceedings of International Conference on Learning Representations (2018)
8. Buckman, J., Roy, A., Raffel, C., Goodfellow, I.: Thermometer encoding: One hot way to resist adversarial examples. In: Proceedings of International Conference on Learning Representations (2018)
9. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
10. Chen, J., Jordan, M.I., Wainwright, M.: Hopskipjumpattack: A query-efficient decision-based attack. arXiv preprint arXiv:1904.02144 (2019)
11. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. pp. 15–26. ACM (2017)
12. Cheng, M., Le, T., Chen, P.Y., Yi, J., Zhang, H., Hsieh, C.J.: Query-efficient hard-label black-box attack: An optimization-based approach. In: Proceedings of International Conference on Learning Representations (2018)
13. Cheng, M., Singh, S., Chen, P.Y., Liu, S., Hsieh, C.J.: Sign-opt: A query-efficient hard-label adversarial attack. Proceedings of International Conference on Learning Representations (2020)
14. Cheng, S., Dong, Y., Pang, T., Su, H., Zhu, J.: Improving black-box adversarial attacks with a transfer-based prior. In: Advances in Neural Information Processing Systems. pp. 10934–10944 (2019)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
16. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9185–9193 (2018)

17. Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7714–7722 (2019)
18. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 (2017)
19. Fan, Y., Wu, B., Li, T., Zhang, Y., Li, M., Li, Z., Yang, Y.: Sparse adversarial attack via perturbation factorization. In: Proceedings of European Conference on Computer Vision (2020)
20. Feng, Y., Wu, B., Fan, Y., Li, Z., Xia, S.: Efficient black-box adversarial attack guided by the distribution of adversarial perturbations. arXiv preprint arXiv:2006.08538 (2020)
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of International Conference on Learning Representations (2014)
22. Guo, C., Gardner, J.R., You, Y., Wilson, A.G., Weinberger, K.Q.: Simple black-box adversarial attacks. arXiv preprint arXiv:1905.07121 (2019)
23. Guo, Y., Yan, Z., Zhang, C.: Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In: Advances in Neural Information Processing Systems. pp. 3825–3834 (2019)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
25. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2261–2269 (2017)
26. Huang, G., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. Tech. rep. (10 2008)
27. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: Proceedings of International Conference on Machine Learning (2018)
28. Ilyas, A., Engstrom, L., Madry, A.: Prior convictions: Black-box adversarial attacks with bandits and priors. In: Proceedings of International Conference on Learning Representations (2019)
29. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical Report (2009)
30. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. In: Proceedings of International Conference on Learning Representations (2016)
31. Li, Y., Li, L., Wang, L., Zhang, T., Gong, B.: Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: Proceedings of International Conference on Machine Learning (2019)
32. Li, Y., Wu, B., Feng, Y., Fan, Y., Jiang, Y., Li, Z., Xia, S.: Toward adversarial robustness via semi-supervised robust training. arXiv preprint arXiv:2003.06974 (2020)
33. Li, Y., Yang, X., Wu, B., Lyu, S.: Hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations. arXiv preprint arXiv:1906.09288 (2019)
34. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: Proceedings of International Conference on Learning Representations (2016)

35. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of International Conference on Learning Representations (2017)
36. Moon, S., An, G., Song, H.O.: Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In: Proceedings of International Conference on Machine Learning (2019)
37. 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 (2016)
38. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
39. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519. ACM (2017)
40. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P). pp. 372–387. IEEE (2016)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
42. Suya, F., Chi, J., Evans, D., Tian, Y.: Hybrid Batch Attacks: Finding black-box adversarial examples with limited queries. In: USENIX Security Symposium (2020)
43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
44. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of International Conference on Learning Representations (2013)
45. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: Proceedings of International Conference on Learning Representations (2018)
46. Wu, D., Wang, Y., Xia, S.T., Bailey, J., Ma, X.: Skip connections matter: On the transferability of adversarial examples generated with resnets. In: Proceedings of International Conference on Learning Representations (2020)
47. Xiao, C., Zhong, P., Zheng, C.: Resisting adversarial attacks by k-winners-take-all. In: Proceedings of International Conference on Learning Representations (2020)
48. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1369–1378 (2017)
49. Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
50. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: Proceedings of Network and Distributed System Security Symposium. Internet Society (2018)
51. Xu, Y., Wu, B., Shen, F., Fan, Y., Zhang, Y., Shen, H.T., Liu, W.: Exact adversarial attack to image captioning via structured output learning with latent variables. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4135–4144 (2019)

52. Zhang, H., Wang, J.: Defense against adversarial attacks using feature scattering-based adversarial training. In: Advances in Neural Information Processing Systems (2019)
53. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: Proceedings of International Conference on Machine Learning (2019)