



Exploiting the Local Parabolic Landscapes of Adversarial Losses to Accelerate Black-Box Adversarial Attack

Hoang Tran^{}, Dan Lu^{}, and Guannan Zhang^{}^{}

Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
{tranha, lud1, zhangg}@ornl.gov

Abstract. Existing black-box adversarial attacks on image classifiers update the perturbation at each iteration from only a small number of queries of the loss function. Since the queries contain very limited information about the loss, black-box methods usually require much more queries than white-box methods. We propose to improve the query efficiency of black-box methods by exploiting the smoothness of the local loss landscape. However, many adversarial losses are not locally smooth with respect to pixel perturbations. To resolve this issue, our first contribution is to theoretically and experimentally justify that the adversarial losses of many standard and robust image classifiers behave like parabolas with respect to perturbations in the Fourier domain. Our second contribution is to exploit the parabolic landscape to build a quadratic approximation of the loss around the current state, and use this approximation to interpolate the loss value as well as update the perturbation without additional queries. Since the local region is already informed by the quadratic fitting, we use large perturbation steps to explore far areas. We demonstrate the efficiency of our method on MNIST, CIFAR-10 and ImageNet datasets for various standard and robust models, as well as on Google Cloud Vision. The experimental results show that exploiting the loss landscape can help significantly reduce the number of queries and increase the success rate. Our codes are available at <https://github.com/HoangATran/BABIES>.

Keywords: Adversarial attack · Interpolation scheme · Loss landscape

1 Introduction

Deep neural networks (DNN) have been shown to be susceptible to adversarial examples, which are small, human-imperceptible perturbations to the inputs designed to fool the network prediction [14, 33]. Adversarial attacks can be categorized into two main settings: white-box attacks and black-box attacks. In the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20065-6_19.

white-box setting, the attackers have access to all information about the target model and thus can use the model’s gradient to effectively guide the search for adversarial examples [7, 23, 33]. Black-box setting, on the other hand, attacks a model only from classification queries [9, 18, 25]. This type of access requirement is considered more realistic in practice.

Traditionally, black-box methods require a massive amount of queries to find a successful adversarial perturbation. Since each query to the target model costs time and money, query efficiency is a requisite for any practical black-box attack method. Recent years have seen the development of several black-box approaches with significant improved query efficiency [1, 3, 15, 19, 24, 34]. However, current black-box attacks access the target models only at perturbed samples and completely rely on the queries there to update the perturbation at each iteration. To reduce the number of queries, it would be beneficial to be able to make use of these queries to extract more from the models, inferring the loss values and identifying candidate perturbations, where no model query was made. This is a challenging goal: since the landscapes of adversarial losses are often complicated and not well-understood, the accuracy of approximations of the loss values from available model queries is not guaranteed.

In this paper, we develop a new ℓ_2 black-box adversarial attack on frequency domain, which uses an interpolation scheme to approximate the loss value around the current state and guide the perturbation updates. We refer to our method as *Black-box Attack Based on IntErpolation Scheme (BABIES)*. This algorithm is inspired by our observation that for many standard and robust image classifiers, the adversarial losses behave like *parabolas* with respect to perturbations of an image in the Fourier domain, thus can be captured with quadratic interpolation. We treat the adversarial attack problem as a constraint optimization on an ℓ_2 sphere, and sample along geodesic curves on the sphere. If the queries show improvements, we accept the perturbation. If the queries do not show improvement, we will infer a small perturbation from those samples without additional queries. Our method achieves significantly improved query efficiency because the perturbation updates are now informed not only directly from model queries (as in existing approaches), but also from an accurate quadratic approximation of the adversarial loss around the current state. The main contributions of this work can be summarized as follows:

- Theoretical and empirical justifications that the adversarial loss behaves like a parabola in the Fourier domain, but NOT like a parabola in pixel domain.
- Development of BABIES, a random-search-based black-box attack that exploits the parabolic loss landscape to improve the query efficiency.
- Evaluations of BABIES with targeted and untargeted attacks on MNIST, CIFAR-10 and ImageNet datasets with both standard and defended models.

1.1 Related Works

To guide the search for adversarial examples, existing black-box attacks often aim at approximating the gradient, either from the gradient of a surrogate model [26, 27], or from model queries via finite different approximation, zeroth-

order optimization, natural evolution strategies, etc. [1, 4, 8, 9, 18, 19, 34]. Many approaches for reducing the dimension of the search space have been proposed, based on principal component analysis [4], autoencoder [34], and compressed sensing [20]. Our method generates random perturbations on the low frequency domain, similar to SimBA [16] and PPBA [20]. This subspace has been shown to admit a high density of adversarial perturbations [15]. Other strategies for designing random perturbations to guide random-search-based attacks include Square Attack [3], which crafts perturbations with square shape, PRGF [12], which utilizes a transfer-based prior, and GenAttack [2], which uses genetic algorithms. Adversarial examples can also be generated from learning their probability distributions [11, 22] and combinatorial optimization techniques [24].

Our black-box method is concerned with the score-based scenario, where the attacker has access to the output scores of the target classifier. More limited variants of the black-box setting have also been studied, where only access to the top-1 predicted labels is assumed [5, 6, 10, 18]. Recent work [21] considers no-box settings, where the attacker makes no query to the target model but just gathers a small labeled dataset. These forms of attacks are more challenging.

2 Background

Image classification aims to successfully predict what a human sees in an image. The objective of adversarial attack on an image classification model is to introduce a small distortion beyond human perceptibility into the original image to fool the target model. In this work, we consider the score-based black-box attack. We first give the formal definition of the adversarial attack problem under consideration. Let $f : [0, 1]^d \rightarrow \mathbb{R}^K$ be a classifier with d input dimension and K classes, where $f_k(\mathbf{x})$ is the predicted probability that the image \mathbf{x} belongs to class k . The predicted label of the image \mathbf{x} is denoted by

$$h(\mathbf{x}) := \operatorname{argmax}_{k=1, \dots, K} f_k(\mathbf{x}).$$

An adversary aims to generate a perturbed image, denoted by $\hat{\mathbf{x}}$, with a small perturbation that solves the following constrained optimization problem

$$\min_{\hat{\mathbf{x}}} \delta(\mathbf{x}, \hat{\mathbf{x}}) \quad \text{s.t.} \quad \begin{cases} h(\hat{\mathbf{x}}) \neq h(\mathbf{x}) & (\text{untargeted}), \\ h(\hat{\mathbf{x}}) = \hat{y} & (\text{targeted}), \end{cases} \quad (1)$$

where $\delta(\cdot, \cdot)$ measures the perceptual difference between the original image \mathbf{x} and the adversarial image $\hat{\mathbf{x}}$, and \hat{y} is the target label for targeted attacks. The most commonly used definition for δ is the ℓ_2 norm or the ℓ_∞ norm of the distortion $\mathbf{x} - \hat{\mathbf{x}}$. In this work, we will use the ℓ_2 norm, i.e., $\delta(\mathbf{x}, \hat{\mathbf{x}}) := \|\mathbf{x} - \hat{\mathbf{x}}\|_2$, as the distortion metric.

Loss Minimization. For score-based adversarial attack, we can exploit the access to the score function $f(\mathbf{x})$ to define an adversarial loss $L(\hat{\mathbf{x}}, y)$ to guide the

search towards adversarial examples. For untargeted attack, the probability of the class $h(\mathbf{x})$ that the original image \mathbf{x} belongs to is often used as adversarial loss, i.e., $L(\hat{\mathbf{x}}, h(\mathbf{x})) := f_{h(\mathbf{x})}(\hat{\mathbf{x}})$. For targeted attack towards a label \hat{y} , we want to maximize $f_{\hat{y}}(\hat{\mathbf{x}})$, so choose $L(\hat{\mathbf{x}}, \hat{y}) := -f_{\hat{y}}(\hat{\mathbf{x}})$. Since the gradient of the target classifier is unavailable and each query to the model costs time and money, the total number of black-box queries for constructing an adversarial example must not exceed a prescribed budget. Thus, the optimization problem in Eq. (1) is modified to

$$\min_{\hat{\mathbf{x}}} L(\hat{\mathbf{x}}, y) \text{ s.t. } \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \rho, \text{ queries} \leq B, \quad (2)$$

where B is the maximum allowable number of queries and ρ is the constraint on the maximum image distortion. For notational simplicity, we suppress the dependence of L on y and write L as $L(\hat{\mathbf{x}})$ in the rest of the paper.

To solve (2), we employ an iterative random search approach, where at each iteration, we query along a randomly sampled search direction and update the current point based on those queries. When doing a Taylor expansion of the loss with respect to a perturbation δ along any randomly selected direction, i.e.,

$$L(\hat{\mathbf{x}}) = L(\mathbf{x}^*) + \frac{dL}{d\delta}(\mathbf{x}^*)\delta + \frac{d^2L}{d\delta^2}(\mathbf{x}^*)\delta^2 + \mathcal{O}(\delta^3),$$

with \mathbf{x}^* being the current state, it is intuitive to conjecture that the loss would behave like a parabola in the neighborhood of \mathbf{x}^* . However, it is not the case for all perturbation strategies. In the following sections, we show that the adversarial loss behaves like a parabola in the Fourier domain determined by the discrete cosine transform (DCT) [15, 30], but NOT like a parabola in the pixel domain. Then, we develop the BABIES algorithm that exploits the parabolic loss landscape in the frequency domain to improve query efficiency.

3 Theoretical and Empirical Study on the Landscape of the Adversarial Loss

In this section, we investigate the shape of the loss's landscape with respect to two different perturbations, i.e., pixel perturbation and DCT perturbation [15, 30].

Our main observation is that the loss's landscape is closer to a parabola with respect to a DCT perturbation, as shown in Fig. 2 and 3. To theoretically verify such observation, we consider a simplified convolutional neural network (CNN)-based classifier for 1D signals. The length of each signal sample is N . We assume the first two layers of the CNN is a 3×1 convolutional layer followed by a 2×1 max-pooling layer, which is a common setup for CNN-based classifiers. Let $\mathbf{x} = (x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4})$, for $i \in \{0, \dots, N-4\}$, be a 4×1 interior segment of the signal and $\mathbf{w} = (w_1, w_2, w_3)$ be the convolution filter. The output of the convolutional layer, centered at x_{i+2} and x_{i+3} , consists of two entries y_2 and y_3 given by

$$\begin{aligned} y_2 &= w_1x_{i+1} + w_2x_{i+2} + w_3x_{i+3}, \\ y_3 &= w_1x_{i+2} + w_2x_{i+3} + w_3x_{i+4}, \end{aligned}$$

and the output after the ReLU activation is

$$z_2 = \max(y_2, 0), \quad z_3 = \max(y_3, 0), \quad (3)$$

and the output of the max-pooling layer is

$$u = \max(z_2, z_3). \quad (4)$$

The simplified CNN model is visualized in Fig. 1. Note that we choose the 1D case to avoid tedious derivation, but the theoretical intuition is applicable to 2D and 3D cases.

Let us define a perturbed signal as $\mathbf{x} + \delta \mathbf{q}$, where $\mathbf{q} = (q_{i+1}, q_{i+2}, q_{i+3}, q_{i+4})$ is the perturbation direction. The derivative of the adversarial loss $L(\delta)$ (as a function of the perturbation’s magnitude δ) is represented by

$$\frac{dL}{d\delta}(\delta) = \frac{dL}{du} \frac{du}{d\delta}(\delta), \quad (5)$$

and we focus on analyzing the behavior of $du/d\delta$ for both pixel and DCT perturbations.

The Property of $du/d\delta$ Due to Pixel Perturbation. In this case, we perturb the pixel x_{i+2} , i.e., setting $q_{i+2} = 1$, $q_{i+1} = q_{i+3} = q_{i+4} = 0$, to study how $du/d\delta$ behaves as a function of δ . Specifically, $du/d\delta$ under the perturbation of x_{i+2} can be written as

$$\frac{du}{d\delta}(\delta) = \frac{\partial u}{\partial z_2} \frac{\partial z_2}{\partial y_2} w_2 + \frac{\partial u}{\partial z_3} \frac{\partial z_3}{\partial y_3} w_1, \quad (6)$$

which involves the derivatives of the ReLU function and the max-pooling function, e.g.,

$$\frac{\partial u}{\partial z_2} = \begin{cases} 1, & \text{if } z_2 \geq z_3 \\ 0, & \text{otherwise} \end{cases}, \quad \text{and} \quad \frac{\partial z_2}{\partial y_2} = \begin{cases} 1, & \text{if } y_2 > 0 \\ 0, & \text{otherwise} \end{cases},$$

and $\partial u/\partial z_3$, $\partial z_3/\partial y_3$ can be defined similarly. Therefore, $du/d\delta$ can only choose values from the set

$$\frac{du}{d\delta}(\delta) \in \{0, w_1, w_2\}, \quad (7)$$

when perturbing the pixel x_{i+2} by δ . Since y_2, y_3, z_2, z_3 are functions of δ , the value of $du/d\delta$ may “jump” from one value in $\{0, w_1, w_2\}$ to another, because w_1

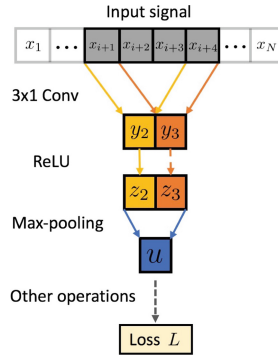


Fig. 1. Illustration of the simplified CNN classifier for verifying our theoretical intuition. We only explicitly write out the first convolutional and max-pooling layers, which is sufficient to verify our theoretical intuition. Other layers are included in “other operations”.

and w_2 may be dramatically different, e.g., $\mathbf{w} = (-1, 5, -1)$ defines a sharpen filter kernel. The maximum jump size could be

$$\left| \frac{du}{d\delta}(\alpha) - \frac{du}{d\delta}(\beta) \right| \leq |w_1| + |w_2|, \quad (8)$$

where $\alpha \neq \beta$ but $|\alpha - \beta|$ is very small. This will eventually lead to the rapid change of the derivative of the total loss $dL/d\delta$ defined in Eq. (5). Figure 2-right illustrates a typical loss landscape with respect to pixel perturbation.

The Property of $du/d\delta$ Due to DCT Perturbation. In this case, all the pixels are perturbed simultaneously. Specifically, the perturbation direction \mathbf{q} is defined by

$$\begin{aligned} q_{i+1} &= \sqrt{\frac{2}{N}} \cos\left(\frac{(2i+1)n\pi}{2N}\right), & q_{i+2} &= \sqrt{\frac{2}{N}} \cos\left(\frac{(2i+3)n\pi}{2N}\right), \\ q_{i+3} &= \sqrt{\frac{2}{N}} \cos\left(\frac{(2i+5)n\pi}{2N}\right), & q_{i+4} &= \sqrt{\frac{2}{N}} \cos\left(\frac{(2i+7)n\pi}{2N}\right), \end{aligned}$$

where $n \in \{0, \dots, N-1\}$ is the selected frequency and N is the total signal length. Then, the derivative $du/d\delta$ is represented by

$$\begin{aligned} \frac{du}{d\delta}(\delta) &= \frac{\partial u}{\partial z_2} \frac{\partial z_2}{\partial y_2} (w_1 q_{i+1} + w_2 q_{i+2} + w_3 q_{i+3}) \\ &\quad + \frac{\partial u}{\partial z_3} \frac{\partial z_3}{\partial y_3} (w_1 q_{i+2} + w_2 q_{i+3} + w_3 q_{i+4}). \end{aligned}$$

Therefore, $du/d\delta$ can only choose values from the set

$$\frac{du}{d\delta}(\delta) \in \{w_1 q_{i+1} + w_2 q_{i+2} + w_3 q_{i+3}, w_1 q_{i+2} + w_2 q_{i+3} + w_3 q_{i+4}, 0\}. \quad (9)$$

As opposed to the pixel perturbation case in Eq. (7), the potential ‘‘jumps’’ of $du/d\delta$ in the DCT domain is much smaller. In fact, the maximum jump size

$$\begin{aligned} \left| \frac{du}{d\delta}(\alpha) - \frac{du}{d\delta}(\beta) \right| &\leq \frac{2\sqrt{2}}{\sqrt{N}} \sin\left(\frac{n\pi}{2N}\right) \left[w_1 \sin\left(\frac{(i+1)n\pi}{N}\right) \right. \\ &\quad \left. + w_2 \sin\left(\frac{(i+2)n\pi}{N}\right) + w_3 \sin\left(\frac{(i+3)n\pi}{N}\right) \right], \end{aligned} \quad (10)$$

where $\alpha \neq \beta$. When perturbing *low-frequency* modes, i.e., n is small, suggested in [15, 30], Eq. (10) can be bounded by $\frac{2\sqrt{2}}{\sqrt{N}} |\sin(\frac{n\pi}{2N})| (|w_1| + |w_2| + |w_3|)$. It is easy to see that this bound is much smaller than the one in Eq. (8) due to the appearance of N (the signal length) in the denominators.

Experimental Illustration. To verify the above intuition, we investigate the landscape of the adversarial loss on untargeted attacks on four different classifiers: (a) standard Inception_v3 on ImageNet [32], (b) ℓ_2 -robust ResNet18 on ImageNet

[29], (c) standard VGG on CIFAR-10 [31], (d) ℓ_2 -robust ResNet50 on CIFAR-10 [13]. For each model, we randomly select 50 images from the corresponding testing sets and define the loss functions as in Background section. We sample 100 1D segments in a neighborhood of each original image, along randomly selected DCT directions and pixel directions, then compute the loss function restricted on them. Then we fit these loss values with parabolas using quadratic regression. The true and approximated landscapes typically found in the DCT and pixel settings are compared in Fig. 2. We observe that the adversarial loss with respect to DCT perturbations is smooth and close to a parabola. On the other hand, the loss function with respect to pixel perturbations shows sharp turns due to the rapid change (jumps) of $dL/d\delta$, therefore cannot be captured by quadratic approximation. This empirical observation is consistent with the above theoretical study.

To show the phenomenon in Fig. 2 is statistically meaningful, we plot in Fig. 3 the correlation between true and approximated loss values given by parabolas on a large number of sample points. To generate each plot, 5000 points are randomly sampled on 100 segments in the neighborhood of each of 50 images (therefore 5000 segments in total). Since the losses on different segments and images are significantly different in value, we normalize them on each segment such that their values lie in $[0, 1]$. Here, we observe strong correlation in DCT setting, confirming that the adversarial losses are generally well-approximated by parabolas in the frequency directions, but much less so in the pixel directions.

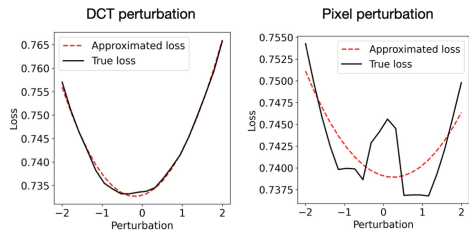


Fig. 2. The landscape of the adversarial loss along DCT directions is often well-behaved and can be fitted with a parabola (**left**). The landscape along pixel direction features sharp turns due to the rapid change (jumps) of $dL/d\delta$ shown in Eq. (8), thus cannot be adequately captured by quadratic approximation (**right**).

4 The BABIES Algorithm

In this section, we present how to exploit the parabolic loss landscape in the DCT domain to develop our BABIES algorithm for black-box attack. Our method consists of two components. Before describing our quadratic interpolation scheme for perturbation updating, we discuss the sampling rule with large step size on the hypersphere.

The Sampling Rule on the Hypersphere. Let us define $\mathcal{B}_\rho := \{\hat{\mathbf{x}} \in [0, 1]^d : \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \rho\}$ and \mathcal{S}_ρ be the boundary of \mathcal{B}_ρ . Let Q be the set of low frequencies extracted by the DCT. Starting from \mathbf{x} , we generate a sequence of iterates $\mathbf{x}^{(k)}$ in \mathcal{B}_ρ which progresses toward an adversarial example. Let ε be the step size

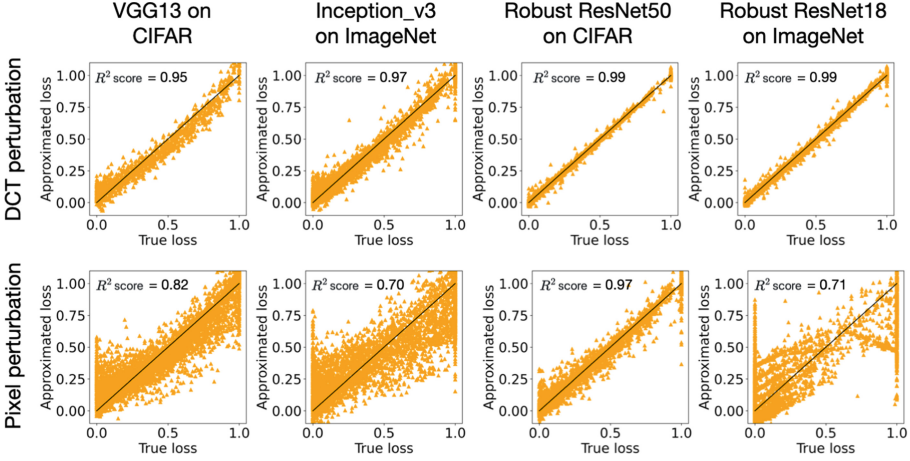


Fig. 3. Scatter plot displays the correlation between true and approximated loss values on a large number of points, sampled from 5000 segments along DCT directions (**top**) and pixel directions (**bottom**). We observe a strong correlation in DCT setting, and much less so along pixel setting. This plot verifies the generality of the example in Fig. 2, that the shape of adversarial losses along DCT directions is close to and can be adequately approximated by a parabola.

parameter and assume $\mathbf{q}^{(k)}$ is the direction sampled from Q , at iteration k , we determine two queries based on $\mathbf{q}^{(k)}$ and ε . When all of $\mathbf{x}^{(k)}$, \mathbf{x}_ε and $\mathbf{x}_{-\varepsilon}$ are in the interior of \mathcal{B}_ρ , i.e., at the beginning of the search, we simply query at

$$\mathbf{x}_{-\varepsilon} = \mathbf{x}^{(k)} - \varepsilon \mathbf{q}^{(k)} \quad \text{and} \quad \mathbf{x}_\varepsilon = \mathbf{x}^{(k)} + \varepsilon \mathbf{q}^{(k)}, \tag{11}$$

and update $\mathbf{x}^{(k)}$ using these queries (see the update rule in the second part). When one or more of $\mathbf{x}^{(k)}$, $\mathbf{x}_{-\varepsilon}$, \mathbf{x}_ε reach the hypersphere \mathcal{S}_ρ , we switch to the sampling rule on the hypersphere, where the queries along the straight line in Eq. (11) is replaced by those along the geodesic curve passing through $\mathbf{x}^{(k)}$ and coplanar to $\mathbf{q}^{(k)}$. We choose $\mathbf{x}_{-\varepsilon}$ and $\mathbf{x}_\varepsilon \in \mathcal{S}_\rho$ so that the *arc length* (instead of standard distance) between them and $\mathbf{x}^{(k)}$ is ε . Denoting $\delta^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$, then the angle between $\delta^{(k)}$ and the line connecting $\mathbf{x}_{-\varepsilon}$ or \mathbf{x}_ε and $\mathbf{x}^{(k)}$ will be ε/ρ . Extract the tangent component of $\mathbf{q}^{(k)}$ as $\tilde{\mathbf{q}}^{(k)} := \mathbf{q}^{(k)} - \langle \mathbf{q}^{(k)}, \delta^{(k)} \rangle \delta^{(k)} / \|\delta^{(k)}\|_2^2$, we arrive at the formula for queries on \mathcal{S}_ρ :

$$\mathbf{x}_{\pm\varepsilon} = \rho \frac{\left(\mathbf{x}^{(k)} \pm \frac{\tilde{\mathbf{q}}^{(k)}}{\|\tilde{\mathbf{q}}^{(k)}\|} \rho \tan\left(\frac{\varepsilon}{\rho}\right) \right)}{\left\| \mathbf{x}^{(k)} \pm \frac{\tilde{\mathbf{q}}^{(k)}}{\|\tilde{\mathbf{q}}^{(k)}\|} \rho \tan\left(\frac{\varepsilon}{\rho}\right) \right\|_2}. \tag{12}$$

The key hyperparameter of our algorithm is the query step size ε . Here, we select relatively large ε for better long-range exploration of adversarial examples. Since the generated samples always lie in \mathcal{B}_ρ , we can use large query steps without concerning with the image distortion. Note that the iterates quickly approach \mathcal{S}_ρ , so

we spend all of the efforts, except the first few iterations, searching the adversarial examples on the hypersphere \mathcal{S}_ρ . We do not fine-tune ε in our experiments. As seen in the next section, the value of step size ε is fixed for each type of target models (standard or robust) and datasets (ImageNet or CIFAR-10 and MNIST), even though ρ significantly varies.

The Update Rule with Quadratic Interpolation. We discuss how to update the iterate from the loss values at three points $\mathbf{x}^{(k)}$, $\mathbf{x}_{-\varepsilon}$ and \mathbf{x}_ε derived from either Eq. (11) or Eq. (12). If one of the queries decreases the loss value, i.e., $\min(L(\mathbf{x}_{-\varepsilon}), L(\mathbf{x}_\varepsilon)) < L(\mathbf{x}^{(k)})$, we accept it as a new state

$$\mathbf{x}^{(k+1)} = \operatorname{argmin}(L(\mathbf{x}_{-\varepsilon}), L(\mathbf{x}_\varepsilon)), \quad (13)$$

and thus make a big step to explore a new region on the hypersphere. When $\min(L(\mathbf{x}_{-\varepsilon}), L(\mathbf{x}_\varepsilon)) \geq L(\mathbf{x}^{(k)})$, the loss function L restricted on the geodesic curve (or straight line if searching within \mathcal{B}_ρ) connecting $\mathbf{x}^{(k)}$, $\mathbf{x}_{-\varepsilon}$ and \mathbf{x}_ε has a local minimizer. Certainly, it is desirable to identify and use this minimizer for the next iterate. Based on our intuition in the previous section, *the idea is to fit a parabola to the three data points to estimate the loss, and assign the minimizer of the parabola to $\mathbf{x}^{(k+1)}$* . The geodesic (or standard) distance between this minimizer and $\mathbf{x}^{(k)}$ can be computed as

$$\gamma = \frac{\varepsilon}{2} \frac{L(\mathbf{x}_\varepsilon) - L(\mathbf{x}_{-\varepsilon})}{L(\mathbf{x}_\varepsilon) - 2L(\mathbf{x}^{(k)}) + L(\mathbf{x}_{-\varepsilon})}. \quad (14)$$

As a result, $\mathbf{x}^{(k+1)} = \mathbf{x}_\gamma$, where \mathbf{x}_γ is defined similarly to \mathbf{x}_ε in Eq. (12) when searching on \mathcal{S}_ρ or Eq. (11) when searching within \mathcal{B}_ρ with γ replaces ε . It is easy to see that the update is not zero when $L(\mathbf{x}_\varepsilon) \neq L(\mathbf{x}_{-\varepsilon})$. Moreover, we use the interpolated loss value

$$L_{\text{interp}} = L(\mathbf{x}^{(k)}) + \frac{1}{8} \frac{(L(\mathbf{x}_\varepsilon) - L(\mathbf{x}_{-\varepsilon}))^2}{L(\mathbf{x}_\varepsilon) - 2L(\mathbf{x}^{(k)}) + L(\mathbf{x}_{-\varepsilon})}, \quad (15)$$

to approximate the current best loss value, instead of querying the loss function at $\mathbf{x}^{(k+1)}$. The present strategy continuously updates the iterates with small

Algorithm 1: BABIES in Pseudocode

```

1: Procedure: BABIES( $\mathbf{x}, \hat{y}, Q, \varepsilon, \rho$ )
2:  $k = 0, \mathbf{x}^{(k)} = \mathbf{x}, \mathcal{L} = L(\mathbf{x}^{(k)})$ 
3: while  $h(\mathbf{x}^{(k)}) \neq \hat{y}$  do
4:   Pick  $\mathbf{q}^{(k)} \in Q$  randomly
5:   if  $\mathbf{x}^{(k)} \notin \mathcal{S}_\rho$  then
6:     Sample  $\mathbf{x}_{-\varepsilon}$  and  $\mathbf{x}_\varepsilon$  using Eq. (11)
7:   else
8:     Sample  $\mathbf{x}_{-\varepsilon}$  and  $\mathbf{x}_\varepsilon$  using Eq. (12)
9:   end if
10:  Compute  $\mathcal{L} = \min(L(\mathbf{x}_{-\varepsilon}), L(\mathbf{x}_\varepsilon))$ 
11:  if  $\mathcal{L} < L(\mathbf{x}^{(k)})$  then
12:    Update  $\mathbf{x}^{(k+1)}$  using Eq. (13)
13:  else
14:    Update  $\mathbf{x}^{(k+1)} := \mathbf{x}_\gamma$ , with  $\gamma$  given
    in Eq. (14)
15:    Update  $\mathcal{L}$  using interpolation in
    Eq. (15)
16:  end if
17:   $k := k + 1$ 
18: end while
19: return  $\mathbf{x}^{(k)}$ 

```

moves, when the random queries cannot find a new candidate. The effectiveness of this strategy relies on the interpolation accuracy. When generating adversarial examples in the frequency domain, the interpolation error is sufficiently small and able to lead the random search towards the loss descent direction (justified by the parabolic landscape of the adversarial loss).

5 Experimental Evaluation

5.1 Results of BABIES on MNIST, CIFAR-10 and ImageNet

We evaluate BABIES-DCT and compare with established algorithms from the literature: Bandits-TD ℓ_2 attack [19], SimBA-DCT [16] and ℓ_2 -Square Attack [3] on the MNIST, CIFAR-10 and ImageNet datasets. We use the default hyperparameters suggested by the authors of the baseline methods. We use the following standard metrics to evaluate the attack performance: the mean number of queries of successful attacks (**Avg. QY**), the median number of query of successful attacks (**Med. QY**) and the success rate (**SR**). Additional evaluations with two other baselines, PPBA [20] and PRGF [12], are provided in the Appendix.

Setup. For MNIST and CIFAR tests, we attack 1,000 correctly labeled images randomly selected from their corresponding testing sets. Evaluation on ImageNet is performed on a set of 1,000 images from the ImageNetV2 [28]. In targeted attack, the target labels are uniformly sampled at random, and the same target labels are used for all methods. The search subspace of BABIES-DCT on ImageNet is set to the first 1/8-th of all frequencies, and includes additional 1/32-th of the next frequencies when all available frequencies are used up without success. Due to the low dimensionality of CIFAR-10 images, we initialize the random search on the first 5/8-th of all frequency, and add an additional 1/8-th

Table 1. The target classifiers and experiment parameters. The first numbers are for *untargeted* attack. Numbers in parentheses are for *targeted* attack. Additional results on sensitivity of BABIES’s performance to hyperparameters are included in Appendix.

Classifiers	Dataset	Type	B	ρ	ϵ
Inception_v3 [32]	ImageNet	Standard	10000 (50000)	5 (12)	2 (3)
ResNet50 [17]	ImageNet	Standard	10000 (50000)	5 (12)	2 (3)
Inception_v3 [32]	CIFAR10	Standard	6144 (6144)	2.4 (4)	2 (2)
VGG13 [31]	CIFAR10	Standard	6144 (6144)	2.4 (4)	2 (2)
ResNet18 [29]	ImageNet	ℓ_2 -robust ($\epsilon = 3$)	10000 (50000)	12 (32)	8 (8)
ResNet50 [29]	ImageNet	ℓ_2 -robust ($\epsilon = 3$)	10000 (50000)	12 (32)	8 (8)
ResNet50 [13]	CIFAR10	ℓ_2 -robust ($\epsilon = 1$)	6144 (6144)	2 (3)	0.5 (0.5)
SmallCNN [35]	MNIST	ℓ_2 -robust ($\epsilon = 0.005$)	10000 (10000)	1 (2)	0.5 (0.5)

of the frequencies at a time, if necessary. We use our method and the other baselines to attack eight pre-trained classifiers (four standard and four ℓ_2 -robust). For each attack, we limit the number of queries (B) the attacker can make and the allowable ℓ_2 distortion (ρ). We use different values of ρ since our experiments span various types of datasets and classifiers. In particular, larger ρ is used if the attacks are more challenging (i.e., on ImageNet dataset, robust models and/or targeted attack). We make minimal tuning of the step size ε , just setting it to be a fraction of ρ . Details of the target classifiers and test parameters here are shown in Table 1. Additional results to show the performance of BABIES is not sensitive with ε are presented in the Appendix.

Results on the Standard Models (Tables 2 and 3). The main comparison results evaluated in the attacks on CIFAR-10 images are reported in Table 2. Here, the median queries of BABIES-DCT are significantly lower than those of other baselines in all of the tests, highlighting that for many images, our method can find an adversarial perturbation in the DCT domain with very few queries. We also achieve the best success rates in three out of four cases. On the average query count, BABIES-DCT and Square-attack each lead in two tests. Both being random-search based method, BABIES-DCT samples from a pre-defined sets of DCT directions while Square-attack crafts random perturbations from a more flexible space. For low dimensional images like CIFAR-10, the set of DCT directions are limited, so eventually all directions will be chosen and recycled. The ability of generating more flexible random directions is an advantage for Square-attack in this case.

Table 3 shows the comparison results on ImageNet. Our method has a very strong performance on the targeted attacks, where it outperforms the others in all three metrics and requires much fewer number of queries (39% and 13% less than the next baseline for Inception_v3 and ResNet50 respectively). On untargeted attacks, the results are more comparable, where BABIES-DCT is slightly better than Square-attack in the success rate, and slightly worse in the number of queries of successful attacks. SimBA does not look particularly strong here, but it should be note that SimBA can find adversarial examples with very small distortions in \mathcal{B}_ρ , while other methods focus on searching on the hypersphere \mathcal{S}_ρ . As such, with the same maximum allowable distortion, SimBA always achieves lowest average distortion on successful attacks.

Table 2. Comparison on attacks against the **standard models** for **CIFAR-10**. BABIES-DCT leads in success rate in 3/4 tests and achieves significantly lower median queries than other baselines in 4/4 tests.

Attack	Inception v3 (untargeted)			Inception v3 (targeted)			VGG13 (untargeted)			VGG13 (targeted)		
	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR
Bandits	409	46	94.5%	817	314	82.9%	199	46	99.6%	601	252	97.7%
Square-attack	170	41	97.5%	345	91	89.4%	96	28	99.9%	130	51	99.3%
SimBA-DCT	203	177	65.9%	603	486	46%	184	140	83.2%	492	356	86.9%
BABIES-DCT	329	17	97.9%	239	62	97.1%	94	5	99.2%	159	47	99.3%

Table 3. Comparison on attacks against the **standard models** for **ImageNet**. For targeted attacks, BABIES outperforms other baselines in all three metrics. For untargeted attacks, the performance of tested methods are more comparable.

Attack	Inception v3 (untargeted)			Inception v3 (targeted)			ResNet50 (untargeted)			ResNet50 (targeted)		
	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR
Bandits	2200	1118	66.6%	21067	19416	77.9%	1238	672	96.6%	4166	3278	100%
Square-attack	1543	777	87.8%	15412	11600	91.3%	1130	563	97.3%	6189	4840	99.5%
SimBA-DCT	1897	1183	78%	16746	13530	82.5%	1673	1115	94.4%	8075	6126	94.6%
BABIES-DCT	1907	908	87.9%	9429	6700	99.2%	1276	742	98%	3670	2923	100%

Results on Defended Models (Tables 4 and 5). Here, none of the evaluated attacks can achieve success rates close to 100%, so we will evaluate them based on the success rate before other metrics, because with low success rate, a method can achieve a low number of queries. For MNIST and CIFAR-10 attacks (Table 4), BABIES-DCT has a significant lead in success rate in three out of four tests (14%–27% to the next baseline). The gap between our method and Bandits is close for untargeted SmallCNN attack, but then, our method posts a much lower average and median query counts. For ImageNet attacks (Table 5), BABIES-DCT leads by a large margin in untargeted ResNet18 (18%), targeted ResNet18 (29%) and untargeted ResNet50 (19%) attacks. Our method is close to Bandits in the targeted ResNet50, but again, it requires much fewer queries. We observe that BABIES shows more significant advantages in attacking defended models, which is consistent with the empirical result (in Fig. 3) that the loss landscape of defended models are closer to parabolas than that of the standard models.

Table 4. Comparison on attacks against the ℓ_2 **robust models** for **CIFAR-10** and **MNIST**. None of the attacks achieve success rates close to 100%, so we evaluate methods on success rate before other metrics. BABIES-DCT leads in success rate in 4/4 tests (and over 14% to the next baseline in 3/4 tests).

Attack	ResNet50, eps=1 (untargeted)			ResNet50, eps=1 (targeted)			SmallCNN, eps=0.005 (untargeted)			SmallCNN, eps=0.005 (targeted)		
	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR
Bandits	2238	1804	45.1%	2992	2919	37.9%	3020	2453	19%	3608	2968	52.5%
Square-attack	1194	302	39%	2464	1978	29.9%	1919	1044	12.9%	2310	1532	41.1%
SimBA-DCT	1280	1232	33.1%	2085	1792	23.8%	1892	2012	8.9%	2440	2660	16.6%
BABIES-DCT	1060	824	59.5%	1489	1420	63.3%	1961	1464	19.6%	2560	1898	67.4%

Table 5. Comparison on attacks against the ℓ_2 -**robust models** for **ImageNet**. Success rate is the most important metric. BABIES-DCT leads in success rate by over 18% to the next baseline in 3/4 tests. For targeted ResNet50, our method is close to Bandits in success rate, but it requires 50% less queries.

Attack	ResNet18, eps=3 (untargeted)			ResNet18, eps=3 (targeted)			ResNet50, eps=3 (untargeted)			ResNet50, eps=3 (targeted)		
	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR	Avg. QY	Med. QY	SR
Bandits	2745	1660	54.3%	13728	12905	33.9%	3952	3558	43.6%	16171	13018	82.4%
Square-attack	1027	209	48.9%	11923	7707	46.5%	1614	758	33.4%	13025	9092	49.9%
SimBA-DCT	4563	3839	50.7%	18321	18843	42.3%	5554	5160	43.2%	19268	19469	21%
BABIES-DCT	2553	1594	71.8%	8283	5700	85.8%	3348	2898	63.2%	8285	5684	83.8%

Qualitative Results (Fig. 4). Since the distortion metric is only an approximation of the imperceptibility, we would like to compare how imperceptible the adversarial images are to the human eye. For that purpose, we selected four images from the targeted attack (on Inception_v3) experiment to explain our observations. The clean images and the distorted images are shown in Fig. 4. All adversarial images have approximately same ℓ_2 distortion norm $\|\delta\|_2 \simeq 12$. It is easy to see that different methods lead to different types of distortion. Even though Bandits is less efficient in our experiments, it generates the most imperceptible adversarial images with comparable ℓ_2 norms. The adversarial images from BABIES-DCT and SimBA-DCT (not shown here) exhibit noticeable wave-like distortions for some images, especially when the background color is light. Finally, Square-attack adds more noticeable sharp distortions. The distortion mass is distinctively concentrated in a set of small squares, coded by the design of searching space by this method.

5.2 Results on Attacking Google Cloud Vision

We perform attacks on the Google Cloud Vision API, which is a powerful image analysis tool capable to identify the dominant entities/objects in images. For an input image, the API provides a list of top concepts appearing in the image, together with their probabilities. We consider an untargeted attack that aims to remove the top 3 objects in the original images. We define the adversarial loss as the maximum of the probabilities of original top 3 concepts, similar to [16, 20], and minimize this loss with BABIES, until an adversarial example is found. We allow maximum 2000 queries for each image and a maximum distortion $\rho = 12.0$ in ℓ_2 norm. Our attack on 50 images randomly selected from ImageNetV2 shows that BABIES achieves an 82% success rate with 205 average queries on successful attacks. In Fig. 5, we show some example images before and after the attack. We observe that the leading concepts in the original images, related to food, laundry, camel and insect, are completely removed in the adversarial images and replaced by

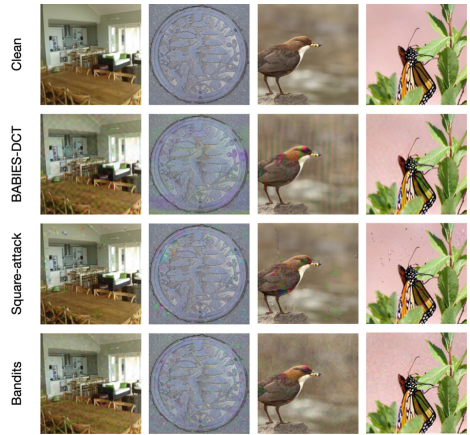


Fig. 4. Qualitative comparison of the imperceptibility of distortion. The distorted images are selected from the targeted attack (on Inception_v3) experiment and have approximately same distortion norm $\|\delta\|_2 \simeq 12$. Bandits generates perturbations with a grainy look and can blend with the background. The wave-like distortions from BABIES-DCT are noticeable for some images. Square attack generates in general more noticeable distortions compared with the other methods.

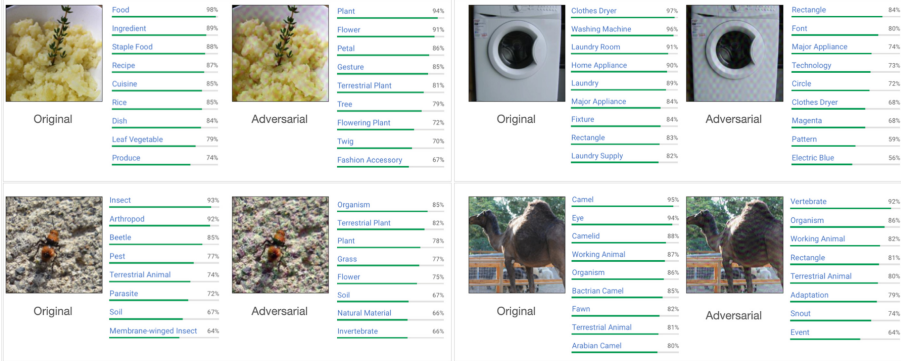


Fig. 5. Example images in our BABIES attack on the Google Cloud Vision API to remove top 3 labels. Labels related to the main objects of original images (clockwise from top left: food, laundry, camel, insect) are completely removed and replaced by less important or incorrect labels.

less important or incorrect labels. This test demonstrates that our method is highly effective against real world systems.

6 Discussion and Conclusion

We propose to exploit the parabolic landscape of the adversarial loss with respect to DCT perturbation to improve the query efficiency of random search methods for black-box adversarial attack. Using a simple quadratic interpolation strategy, we demonstrate that the loss smoothness can greatly improve query efficiency without additional query per iteration. Our algorithm solve a constraint optimization problem on l_2 sphere. Thus we propose to use large query step for better exploration of the search space. Our evaluation shows a remarkable advantage of this strategy.

Our theoretical and empirical study on the landscape of the adversarial loss provides a new angle to investigate the vulnerability of an image classifier. From this perspective, the theoretical insight on the loss landscape may be even more valuable to the community than the BABIES algorithm. For example, an intriguing observation from theoretical study and our experiment is that the relative performance of BABIES-DCT (in comparison to other baselines) is strongest in attacking l_2 -robust models. One possible reason is that the loss landscapes of the defended models are closer to parabolas, which provides a favorable setting for our method. While defended classifiers have been studied extensively recently, l_∞ models have got more attention and less is known about l_2 models. Understanding the properties and possible weakness of l_2 models is an interesting problem we plan to study next. Despite the superior performance, our method has several limitations. First, our method is designed for l_2 attack, which may not outperform the state-of-the-art methods in l_∞ attack. Second, since the

perturbation is made in the Fourier domain, the perturbation is combination of cosine functions which is easier to be distinguished by naked eyes than pixel perturbations, even though the ℓ_2 norm satisfies the constraint.

There are several possible directions to pursue in the future research. One is to investigate the loss smoothness in other spaces, e.g., replacing DCT with wavelet transform. In fact, the idea of Square Attack makes Haar wavelet transform a good candidate to study. An advantage of using wavelet transform is that wavelet is only supported locally, which means perturbing a wavelet mode will result in a smaller distortion than perturbing a globally supported cosine basis. Another area for improvement is to perturb multiple DCT modes within each iteration for more efficient exploration. We leave these directions for future work.

Acknowledgments. This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program; and by the Artificial Intelligence Initiative at the Oak Ridge National Laboratory (ORNL). ORNL is operated by UT-Battelle, LLC., for the U.S. DOE under Contract DE-AC05-00OR22725.

References

1. Al-Dujaili, A., O'Reilly, U.: Sign bits are all you need for black-box attacks. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. OpenReview.net (2020). <https://openreview.net/forum?id=SygW0TEFwH>
2. Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.J., Srivastava, M.: Genattack: practical black-box attacks with gradient-free optimization (2019)
3. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search (2020)
4. Bhagoji, A.N., He, W., Li, B., Song, D.: Practical black-box attacks on deep neural networks using efficient query mechanisms. In: Proceedings of the European Conference on Computer Vision (ECCV), September 2018
5. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 2018 May, Conference Track Proceedings. OpenReview.net (2018). <https://openreview.net/forum?id=SyZI0GWCZ>
6. Brunner, T., Diehl, F., Truong-Le, M., Knoll, A.C.: Guessing smart: biased sampling for efficient black-box adversarial attacks. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 4957–4965. IEEE (2019). <https://doi.org/10.1109/ICCV.2019.00506>
7. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57 (2017)
8. Chen, J., Zhou, D., Yi, J., Gu, Q.: A frank-wolfe framework for efficient and effective adversarial attacks. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial

- Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020, pp. 3486–3494. AAAI Press (2020). <http://aaai.org/ojs/index.php/AAAI/article/view/5753>
9. 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, AISec 2017, pp. 15–26. Association for Computing Machinery, New York (2017). <https://doi.org/10.1145/3128572.3140448>
 10. Cheng, M., Le, T., Chen, P., Zhang, H., Yi, J., Hsieh, C.: Query-efficient hard-label black-box attack: an optimization-based approach. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. OpenReview.net (2019). <https://openreview.net/forum?id=rJlk6iRqKX>
 11. Dolatabadi, H.M., Erfani, S.M., Leckie, C.: Advflow: inconspicuous black-box adversarial attacks using normalizing flows. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020). <https://proceedings.neurips.cc/paper/2020/hash/b6cf334c22c8f4ce8eb920bb7b512ed0-Abstract.html>
 12. Dong, Y., Cheng, S., Pang, T., Su, H., Zhu, J.: Query-efficient black-box adversarial attacks guided by a transfer-based prior. IEEE Trans. Pattern Anal. Mach. Intell. (2021). <https://doi.org/10.1109/TPAMI.2021.3126733>
 13. Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., Tsipras, D.: Robustness (python library) (2019). <https://github.com/MadryLab/robustness>
 14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. CoRR abs/1412.6572 (2015)
 15. Guo, C., Frank, J.S., Weinberger, K.Q.: Low frequency adversarial perturbation. In: Globerson, A., Silva, R. (eds.) Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, 22–25 July 2019. Proceedings of Machine Learning Research, vol. 115, pp. 1127–1137. AUAI Press (2019). <http://proceedings.mlr.press/v115/guo20a.html>
 16. Guo, C., Gardner, J.R., You, Y., Wilson, A.G., Weinberger, K.Q.: Simple black-box adversarial attacks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 2484–2493. PMLR (2019). <http://proceedings.mlr.press/v97/guo19a.html>
 17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
 18. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, Stockholm, Sweden, 10–15 July 2018, vol. 80, pp. 2137–2146. PMLR (2018). <http://proceedings.mlr.press/v80/ilyas18a.html>
 19. Ilyas, A., Engstrom, L., Madry, A.: Prior convictions: black-box adversarial attacks with bandits and priors. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. OpenReview.net (2019). <https://openreview.net/forum?id=BkMiWhR5K7>

20. Li, J., et al.: Projection and probability-driven black-box attack. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
21. Li, Q., Guo, Y., Chen, H.: Practical no-box adversarial attacks against DNNs. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020). <https://proceedings.neurips.cc/paper/2020/hash/96e07156db854ca7b00b5df21716b0c6-Abstract.html>
22. 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: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 3866–3876. PMLR (2019). <http://proceedings.mlr.press/v97/li19g.html>
23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv abs/1706.06083 (2018)
24. Moon, S., An, G., Song, H.O.: Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 4636–4645. PMLR (2019). <http://proceedings.mlr.press/v97/moon19a.html>
25. Narodytska, N., Kasiviswanathan, S.: Simple black-box adversarial attacks on deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1310–1318 (2017). <https://doi.org/10.1109/CVPRW.2017.172>
26. Papernot, N., McDaniel, P., Goodfellow, I.J.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv abs/1605.07277 (2016)
27. 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, ASIA CCS 2017, pp. 506–519. Association for Computing Machinery, New York (2017). <https://doi.org/10.1145/3052973.3053009>
28. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, 09–15 June 2019, vol. 97, pp. 5389–5400. PMLR (2019)
29. Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust imagenet models transfer better? arXiv preprint [arXiv:2007.08489](https://arxiv.org/abs/2007.08489) (2020)
30. Sharma, Y., Ding, G.W., Brubaker, M.: On the Effectiveness of Low Frequency Perturbations. arXiv e-prints [arXiv:1903.00073](https://arxiv.org/abs/1903.00073), February 2019
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015). arxiv.org/abs/1409.1556
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)

33. Szegedy, C., et al.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014). arxiv.org/abs/1312.6199
34. Tu, C.C., et al.: Autozoom: autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: AAAI (2019)
35. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (2019)