Polishing Decision-based Adversarial Noise with a Customized Sampling

Yucheng Shi[1,2], Yahong Han[1,2*], Qi Tian[3]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Lab of Machine Learning, Tianjin University, Tianjin, China
[3]Noah's Ark Lab, Huawei Technologies
{yucheng, yahong}@tju.edu.cn, tian.qi1@huawei.com

Abstract

As an effective black-box adversarial attack, decision-based methods polish adversarial noise by querying the target model. Among them, boundary attack is widely applied due to its powerful noise compression capability, especially when combined with transfer-based methods. Boundary attack splits the noise compression into several independent sampling processes, repeating each query with a constant sampling setting. In this paper, we demonstrate the advantage of using current noise and historical queries to customize the variance and mean of sampling in boundary attack to polish adversarial noise. We further reveal the relationship between the initial noise and the compressed noise in boundary attack. We propose Customized Adversarial Boundary (CAB) attack that uses the current noise to model the sensitivity of each pixel and polish adversarial noise of each image with a customized sampling setting. On the one hand, CAB uses current noise as a prior belief to customize the multivariate normal distribution. On the other hand, CAB keeps the new samplings away from historical failed queries to avoid similar mistakes. Experimental results measured on several image classification datasets emphasizes the validity of our method.

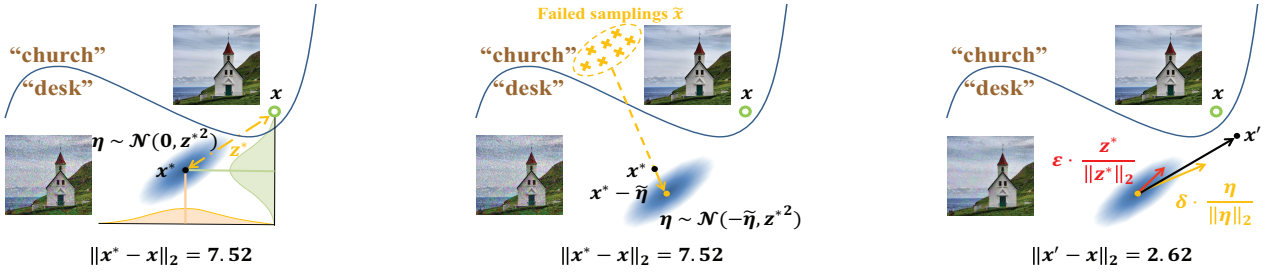
1. Introduction

Adversarial examples [29, 22] have revealed the inherent vulnerability of deep neural networks (DNNs). Based on attackers' knowledge of the target model [23], adversarial attacks can be divided into white-box attacks and black-box attacks. In black-box attack, attackers can only query target model and get the hard-label predictions without access to complete knowledge of target model. Transfer-based attacks [13, 18, 8, 9, 25], decision-based attacks [32, 10, 2, 25], and attacks based on zeroth order optimization [5, 20, 31] are three mainstream black-box attacks. Among them, decision-based attacks squeeze out noise by randomly searching in the input space of original image. It requires neither substitute models as transfer-based attacks nor a thorough query to the target model as zeroth order optimization, and can generate adversarial noise with relatively small magnitude under limited queries. Several recent studies [2, 1, 7] indicates that, a combination of transfer-based attacks and decision-based attacks achieves the state-of-the-art black-box attack effect.

Constructing adversarial examples is not to simply fool DNNs, but to quantitatively evaluate the robustness of target model. By continuously polishing adversarial perturbation, we can gradually achieve an accurate evaluation of the minimum noise magnitude for misclassification. For an image classifier, the minimum noise required to misclassify each image, the reasonable query direction in each stages of one attack process, and even the sensitivity of each pixel in an image are all different [11]. Therefore, an accurate evaluation on the robustness of one target model requires customization of each image and its attack process. White-box attacks directly model the correlation between pixels and categories by back-propagation [13, 3]. For black-box attacks, the only clue available for attackers is the historical query, which is an unbiased characterization of each pixel's noise sensitivity. However, most existing decision-based attacks [1, 2] use constant sampling settings independent of historical queries or current noise, which severely hinders the efficiency of noise polishing.

Failed samplings in decision-based attacks contain location information of decision boundaries [12]. Although failed samplings, i.e., samples fall in the true category, cannot be directly used to compress noise, they depict directions with greater probability across the decision boundary. Since we want as many samples as possible that fall in the other side of decision boundary, this information can be used to customize sampling process and keep new samples away from directions with a high probability of failure. But existing decision-based attacks always sample on a constant distribution, and never change the sampling during the

*Corresponding author.



(a) Variance Customize (b) Mean Customize (c) Sample and Polish

Figure 1. Flowchart of CAB attack. Blue curve represents decision boundary between 'church' and 'desk' categories. In each plot, green circle and black dot represent the original image and current adversarial example, respectively. Yellow crosses in (b) are set of historical failed samplings $\tilde{x}$. CAB customizes the variance of the normal distribution by current noise z^* in (a) and its mean by opposite direction of historical failed samples $-\tilde{\eta}$ in (b). Then CAB perform random sampling on the customized distribution to form spherical direction. Finally, spherical (yellow arrow in (c)) and source directions (red arrow) are combined to form a polished adversarial example x' in (c).

attack process. In addition, the stepsize of single-step modification in existing decision-based attacks is also a constant value. As the noise magnitude decreases, the success rate of query will gradually decreases, and the efficiency of noise polishing will be further affected with a constant stepsize.

In this paper, we show that in order to minimize the noise magnitude after one single step of boundary attack, the variance of multivariate normal distribution should be linearly correlated to the absolute value of current noise as shown in Fig. 1 (a), instead of using unit variance for each dimension. Moreover, we analyze the advantages of initializing adversarial noise with transfer-based attack over random initialization through the monotonicity of noise compression. We adjusted the strategy of stepsize customization based on this property of noise compression. Under the guidance of current noise and historical queries, we propose Customized Adversarial Boundary (CAB), a decision-based attack customizes sampling distribution in line with noise sensitivity of each pixel. CAB customizes the mean of sampling distribution by historical failed samples, as demonstrated by yellow crosses in Fig. 1 (b). In this way, new samples are guided away from directions with high failure rates. Experiments on Imagenet [24], Tiny-Imagenet [1], MNIST [19], and CIFAR-10 [17] show that CAB achieves the smaller median noise magnitude than other decision-based attacks under the same query limitation.

We summarize our contributions as follows:

(1) We show that in order to maximize the noise reduction expectation, the variance of sampling in boundary attack should be proportional to current noise.

(2) Based on the monotonicity of noise compression for boundary attack, we improve its stepsize adjustment and use transfer-based attack to customize the initial noise.

(3) We develop CAB, a decision-based attack that exploits current noise and failed samples to customize the normal distribution in the sampling process. Extensive experiments on several datasets and models demonstrate the superior performance of CAB over other decision-based attacks.

2. Related Work

When there is no access to gradients of the target model, transfer-based attacks, decision-based attacks, and attacks based on zeroth order optimization give three different solutions for the black-box scenario. In this paper, we mainly discuss the first two methods and their combinations.

2.1. Transfer-based Attack

Transfer-based attacks fool DNNs by exploiting transferability between substitute model and target model [21]. The effect of transfer-based attack can be influenced by ensemble adversarial training [30] of the target model. A more reasonable strategy is to divide black-box attack into two phases: firstly generate adversarial examples as starting points by transfer-based attacks, and further compress their redundant noise by decision-based attacks [25, 2].

2.2. Decision-based Attack

Decision-based attacks sample in the neighborhood of the original image to seek smaller noise magnitude without crossing decision boundaries. Decision-based attacks do not rely on substitute models, but use various strategies to find adversarial examples. Most decision-based attacks require an initial adversarial example that has already been misclassified as starting point. Several state-of-the-art decision-based attacks are introduced in the following.

Whey Optimization. Whey optimization [25] divides adversarial noise into groups to reduce noise magnitude.

The greedy searching process of Whey tends to fall into local optimum after several steps of compression, reducing searching efficiency of later stage.

Boundary Attack. Boundary attack [32] starts from an adversarial example and search along two directions simultaneously, namely spherical direction and source direction:

$$x_{t+1} = x_t + \delta \cdot \frac{\eta}{\|\eta\|_2} + \varepsilon \cdot \frac{x - x_t}{\|x - x_t\|_2}, \quad \eta \sim \mathcal{N}(0, I) \quad (1)$$

where x_t is the adversarial example with smallest noise after t steps of boundary attack. η and $(x - x_t)$ refer to the direction of spherical and source direction, respectively. δ is the stepsize of spherical direction, and ε is the stepsize of source direction. Because of the indiscriminate use of standard normal distribution for each dimension, boundary attack cannot evaluate and exploit the differences of noise sensitivity between pixels.

Biased Boundary Attack. The Biased Boundary Attack [2] replaces the normal distribution in boundary with Perlin distribution, concentrating on low-frequency domain of input space to make the adversarial example more 'natural'.

Evolutionary Attack. The Evolutionary Attack [10] reduces the dimension of sampling space by bilinear interpolation and restricting noise to the central part of images. Evolutionary attack performs better in tasks involving strong prior knowledge such as face recognition.

There are some other attacks involving zeroth order optimization [5, 20, 31, 4, 6]. They are mainly aiming at black-box scenario where score of each category can be obtained or with relatively sufficient query budget. In this paper, we only discuss black-box scenario with limited queries and the target model only outputs hard label.

3. Proposed Method

3.1. Notation

Consider a target model based on DNN under black-box attack: $F : X^N \to Y^C$, where X represents the input space, N is the dimension ($N = Width \times Height \times Channel$ for image data) and Y represents the classification space with C categories. Suppose x^* is an adversarial example with smallest noise magnitude that we have found. The objective of decision-based attack can be described as:

$$\max_{x'} \ \|x^* - x\|_2 - \|x' - x\|_2, \quad s.t. \quad F(x) \neq F(x') \,, \quad (2)$$

where x and x' represent the original image and the new adversarial example generated after this step, respectively. We replace the adversarial examples x^* and x' with the sum of original image x and adversarial noise, z^* and $z^* + z$, where z^* and z are the current adversarial noise with minimum magnitude and the noise added after this step, respectively. Since x and x^* are fixed, the objective function in Eqn. (2) can be equivalently reformulated as:

$$\min_{z} \ \|z^* + z\|_2, \quad s.t.\ F(x) \neq F(x + z^* + z) \,, \quad (3)$$

Note that the ℓ_2 distance is calculated under the premise that adversarial examples are misclassified by the target model. The ℓ_2 norm is chosen as the distance metric because it can more accurately characterize the robustness of one model than ℓ_∞ norm [11].

3.2. Variance and Noise Reduction

In this section, we will formally demonstrate that expectation of noise reduction maximizes when variance of normal distribution in sampling linearly correlates to the absolute value of current noise.

Consider the one-step noise update of boundary attack in Eqn. (1), we rewrite z as $\delta \cdot \frac{\eta}{\|\eta\|_2} + \varepsilon \cdot \frac{x - x^*}{\|x - x^*\|_2}$, and $(1 - \frac{\varepsilon}{\|z^*\|_2})$ as α. Since $x - x^* = -z^*$, we have

$$\begin{aligned} & \|z^* + z\|_2 \\ = \ & \|z^* + \delta \cdot \frac{\eta}{\|\eta\|_2} + \varepsilon \cdot \frac{-z^*}{\| - z^*\|_2}\|_2 \\ = \ & \|\alpha \cdot z^* + \delta \cdot \frac{\eta}{\|\eta\|_2}\|_2 \\ = \ & \sqrt{\|\alpha \cdot z^*\|_2^2 + 2 \cdot \delta \cdot \alpha \cdot (z^* \bullet \frac{\eta}{\|\eta\|_2}) + \|\delta \cdot \frac{\eta}{\|\eta\|_2}\|_2^2} \end{aligned}$$

where $\bullet$ denotes the standard inner product. Since z^*, ε and δ are all fixed, $\|\delta \cdot \frac{\eta}{\|\eta\|_2}\|_2^2 \equiv \delta^2$, object of Eqn. (3) is actually to minimize $z^* \bullet \frac{\eta}{\|\eta\|_2}$. By the Cauchy-Schwarz Inequality we have

$$- \ \|z^*\|_2 \cdot \ \|\eta\|_2 \leq z^* \bullet \eta \leq \ \|z^*\|_2 \cdot \ \|\eta\|_2 \quad (4)$$

with $\beta = \frac{\eta}{\|\eta\|_2}$, this yields

$$\begin{aligned} & \|z^* + z\|_2^2 \\ \geq \ & \|\alpha \cdot z^*\|_2^2 - 2 \cdot \alpha \cdot \delta \cdot (\|z^*\|_2 \bullet \|\beta\|_2) + \|\delta \cdot \beta\|_2^2 \\ = \ & (\|\alpha \cdot z^*\|_2 - \|\delta \cdot \beta\|_2)^2 \\ \Longrightarrow \ & \|z^* + z\|_2 \geq \big| \|\alpha \cdot z^*\|_2 - \|\delta \cdot \beta\|_2 \big| \end{aligned}$$

The equality holds when $z^* = -k\beta, k \in \mathcal{R}^+$. In other words, the magnitude of total noise $\|z^*\|_2$ after boundary attack is minimized when the direction of $\frac{\eta}{\|\eta\|_2}$ and current noise z^* are exactly reversed. Suppose that $\eta \sim \mathcal{N}(0, \Sigma)$ is a N-dimensional random vector that follows the normal distribution with zero mean and covariance matrix $\Sigma = diag(\sigma_1^2, \cdots, \sigma_N^2)$ (in boundary attack, $\Sigma = I_N$). Each element of η is a univariate normal distribution, the mean is set to zero for better exploration in the sampling space [10]. Since $\beta = \frac{\eta}{\|\eta\|_2}$ and $\beta_i = \frac{\eta_i}{\sqrt{\eta_i^2 + \cdots + \eta_N^2}}$. The ratio of β_i^2 's

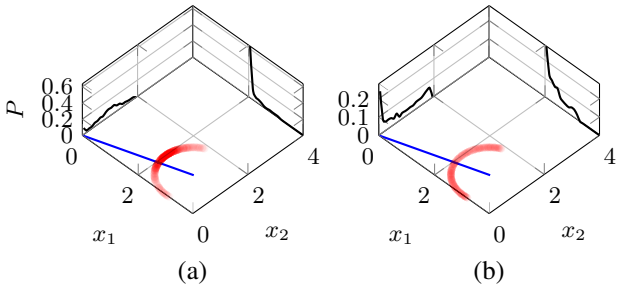


Figure 2. Distribution of z with $\sigma_1 : \sigma_2 = 3 : 1$ (a) and $\sigma_1 : \sigma_2 = 1 : 1$ (b) when $N = 2$.

expectation satisfies

$$\begin{aligned} & E(\beta_1^2) : E(\beta_2^2) : \cdots : E(\beta_N^2) \\ = \; & E(\eta_1^2) : E(\eta_2^2) : \cdots : E(\eta_N^2) \\ = \; & Var(\eta_1) : Var(\eta_2) : \cdots : Var(\eta_N) \\ = \; & \sigma_1^2 : \sigma_2^2 : \cdots : \sigma_N^2 \end{aligned}$$

As one kind of rejection sampling [12], the boundary attack will only query the target model when noise reduction regard to one sampling is greater then zero, i.e., $z^* \bullet \eta \leq 0$. Therefore, the expectation of new noise x' after one-step boundary attack minimizes when $\sigma_i \propto \left|z_i^*\right|, 1 \leq i \leq N$.

To show the the influence of σ on noise reduction more intuitively, we visualize the distribution of x' in two-dimensional space in Fig. 2. Blue vectors represent $x^* = (3, 1)$. Red marks indicate the distribution of x' after 1000 samplings under normal distribution with $\sigma_1 : \sigma_2 = 3 : 1$ (a) and $\sigma_1 : \sigma_2 = 1 : 1$ (b) centered on $(3, 1)$. The deeper the red, the denser samples are in the neighborhood. Black line charts at $x_1 = 0$ and $x_2 = 4$ are independent probability distribution $\mathbf{P}$ of x_2 and x_1, respectively. When the variance ratio of two dimensions $\sigma_1 : \sigma_2 = x_1^* : x_2^*$, x' concentrates in the opposite direction of x^* in (a). However, with an equal variance for each dimension in (b), x^* evenly distributes in all directions, which impedes the efficient polishing of noise. This relationship suggests that, compared to standard normal distribution for all dimensions, sampling over normal distribution customized by current noise increases the expectation of noise reduction.

3.3. Customization of Initial Noise and Stepsize

In this section, we customize the initial noise and stepsize by analyzing the monotonicity of noise compression in decision-based attack. Under the assumption that misclassification probability increases monotonically with the distance from the original image, the final noise magnitude is positively correlated with the initial noise magnitude. On the one hand, this explains the effectiveness of initializing adversarial noise by transfer-based attacks. On the other hand, we adapt the strategy of customize stepsize in boundary attack based on this feature.

According to the setting in [11], we denote $\rho_{F,x}(\lambda)$ as the misclassification probability of target model F for a random point with the distance λ from original image x:

$$\rho_{F,x}(\lambda) = \mathbb{P}_{z \sim \lambda \mathbb{S}}\{F(x) \neq F(x+z)\} \tag{5}$$

where $\lambda\mathbb{S}$ denotes the uniform measure on the sphere surface centered at 0 and of radius λ. In other words, the set of points satisfie $\|z\|_2 = \lambda$. Since the risk of misclassification generally increases with the distance from the original image, we assume $\rho_{F,x}(\lambda)$ (abbreviated as $\rho(\lambda)$) increases monotonically with λ within a certain range $\triangle_{adv}(x; F) \leq \lambda \leq \triangle_{unif,\xi}(x; F)$:

$$\begin{aligned} \forall \quad & \lambda_1, \lambda_2 \in [\triangle_{adv}(x; F), \triangle_{unif,\xi}(x; F)], \\ & \lambda_1 \geq \lambda_2 \rightarrow \rho(\lambda_1) \geq \rho(\lambda_2) \end{aligned}$$

As defined in [11], $\triangle_{adv}(x; F)$ denotes the ℓ_2 norm of the global smallest adversarial perturbation that causes misclassification. And $\triangle_{unif,\xi}(x; F)$ denotes the $\xi-$robustness of F to random uniform noise. For simplicity, we assume that after b steps of queries in decision-based attack, the noise magnitude of adversarial example x' to be queried is subject to a uniform distribution from $\triangle_{adv}(x; F)$ to the current smallest noise magnitude $\|z^*\|_2$.

Proposition 1 *Assume that misclassification probability $\rho(\lambda)$ increases monotonically with λ within a certain range $\triangle_{adv}(x; F) \leq \lambda \leq \triangle_{unif,\xi}(x; F)$. For any two adversarial examples x_1' and x_2' about original image x, if the corresponding noise magnitude satisfies $\triangle_{adv}(x; F) \leq \lambda_2 < \lambda_1 \leq \triangle_{unif,\xi}(x; F)$, the expected noise magnitude after one step of decision-based attack satisfies $\mathbb{E}(\lambda_2) < \mathbb{E}(\lambda_1)$.*

Proof. Consider the relationship between noise magnitudes $\lambda = \|z\|_2$ and misclassification probabilities $\rho(\lambda)$, the expected noise magnitude after $b + 1$ steps is:

$$\mathbb{E}(\lambda) = \int_{\triangle_{adv}(x;F)}^{\lambda} a\rho(a) \frac{1}{\lambda - \triangle_{adv}(x; F)} da \tag{6}$$

Since it is impossible to get the specific value of $\triangle_{adv}(x; F)$ before the decision-based attack completely converges, and $\rho(\lambda) = 0$ when $\lambda \leq \triangle_{adv}(x; F)$, we rewrite Eqn. (6) as:

$$\mathbb{E}(\lambda) = \int_0^{\lambda} a\rho(a) \frac{1}{\lambda} da \tag{7}$$

For x_1' and x_2' satisfies $\triangle_{adv}(x; F) \leq \lambda_2 < \lambda_1 \leq \triangle_{unif,\xi}(x; F)$, the difference between the expected noise magnitude after one step of decision-based attack is:

$$\begin{aligned} & \mathbb{E}(\lambda_1) - \mathbb{E}(\lambda_2) \\ = \; & \int_0^{\lambda_1} a\rho(a) \frac{1}{\lambda_1} da - \int_0^{\lambda_2} a\rho(a) \frac{1}{\lambda_2} da \\ = \; & \frac{1}{\lambda_1} \int_{\lambda_2}^{\lambda_1} a\rho(a) da - (\frac{1}{\lambda_2} - \frac{1}{\lambda_1}) \int_0^{\lambda_2} a\rho(a) da \end{aligned}$$

Given $\lambda_2 < \lambda_1$ and monotonicity of $\rho(\lambda)$, we have:

$$\begin{aligned} & \mathbb{E}(\lambda_1) - \mathbb{E}(\lambda_2) \\ \geq \; & \frac{1}{\lambda_1}\int_{\lambda_2}^{\lambda_1} a\rho(\lambda_2)da - (\frac{1}{\lambda_2} - \frac{1}{\lambda_1})\int_0^{\lambda_2} a\rho(\lambda_2)da \\ = \; & \frac{\rho(\lambda_2)}{2\lambda_1}(\lambda_1^2 - \lambda_2^2) - \frac{\lambda_2^2\rho(\lambda_2)}{2}(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}) \\ = \; & \frac{\rho(\lambda_2)}{2\lambda_1}(\lambda_1^2 - \lambda_1\lambda_2) > 0 \end{aligned}$$ □

From Proposition 1, we show that under the assumption that misclassification probability increases monotonically with the distance from the original image, the expected noise magnitude after one step also increases monotonically with the initial noise magnitude. The decision-based attack process satisfies memorylessness, i.e., the current noise is determined only by noise of last step [32]. Therefore, the monotonicity of noise compression satisfies transitivity in multiple steps. In other words, when using the same decision-based attack and querying the target model for the same times, the expected final noise magnitude is positively correlated with the initial noise magnitude.

This explains the effectiveness of combining transfer-based attack with decision-based attack in black-box scenario. Decision-based methods such as boundary attack use random noise as initial noise, whose magnitude is much larger than that generated by transfer-based attacks, thus the final noise is also larger under the same number of queries. We follow this black-box attack setup that uses adversarial examples generated by transfer-based attack as the starting point of decision-based attack.

In addition, as the noise magnitude is continuously compressed, the possibility of misclassification for new query will gradually decrease if the stepsize δ and ε in Eqn. (1) for spherical and source direction remain the same. In order to compensate for the decrease in success rate of querying, we introduce the exponential scheduling to dynamically customize the stepsize in both directions:

$$\delta_s = \delta_0\varphi^s, \quad \varepsilon_s = \varepsilon_0\varphi^s, \tag{8}$$

where s represents the number of successful queries so far, δ_s and ε_s are stepsizes of spherical and source direction after s successful queries. δ_0 and ε_0 are initial stepsizes. $\varphi \in (0, 1)$ is a decay factor for stepsize scheduling. As the distance between nearest adversarial example and the original image is shortened, the stepsize of new query is also reduced. This exponential scheduling strategy balances the noise compression rate with the query success rate, enabling stepsize customization for different images and for different query phases of one image.

3.4. CAB Attack

For boundary attack which randomly searches in a large input space, reducing the sampling space is critical to the efficiency of noise polishing. Evolutionary Attack [10] reduces sampling space by bilinear interpolation and by limiting adversarial noise to the center of an image. Distinguishing the sensitivity of pixels to noise by relative position may be effective for images with a single structure (e.g., face recognition images) or a small size, but not for larger and more complex ones. The current noise z^* is a more unbiased characterization of the sensitivity of pixels compared to artificial rules in [10]. Therefore, we only adjust noise on pixels where the current noise magnitude is already large:

$$H(z, r) = \underset{\hat{z}\subset z^*, |\hat{z}|/|z^*|=r}{\arg\max} \sum_{z\in\hat{z}} |z|\,, \tag{9}$$

$$T_i = \begin{cases} 1, \; if \; z_i^* \in H, \\ 0, \; else. \end{cases} \tag{10}$$

where $\hat{z}$ is a set of pixels in z^* with the largest absolute values, $r \in (0, 1)$ is the ratio of the pixel numbers in $\hat{z}$ and z^*. Specifically, we pick pixels with the largest absolute value in z^* according to ratio r, and form up a mask T to

Algorithm 1 Customized Adversarial Boundary

Input: Target DNN $F(x)$ and adversarial example x^*
Original image x and its label y
Max querying number B, pixel retention rate r
Initial stepsize of spherical direction δ_0
Initial stepsize of source direction ε_0
Decay factor for stepsize scheduling φ
Output: Adversarial example x' with compressed noise

$W \leftarrow [\,]$, $s \leftarrow 0$;
for b in 1 to B **do**
  if $W \neq \varnothing$ **then**
    $z^* \leftarrow x^* - x$;
    // Sample over a customized normal distribution;
    $\eta \sim \mathcal{N}(-\frac{1}{|W|}\sum\tilde{\eta}, z^{*2}) \quad s.t. \quad \tilde{\eta} \in W$;
  else
    $\eta \sim \mathcal{N}(0, z^{*2})$;
  end if
  // Pick up pixels with the largest absolute value in z^*;
  $H(z, r) = \arg\max_{\hat{z}\subset z^*, |\hat{z}|/|z^*|=r} \sum_{z\in\hat{z}}|z|$;
  Construct T by H according to Eqn. (10);
  $x' = x^* + T \bullet (\delta_s \cdot \frac{\eta}{\|\eta\|_2} - \varepsilon_s \cdot \frac{z^*}{\|z^*\|_2})$;
  if $y \neq F(x')$ **then**
    // Sampling is successful, noise is compressed;
    $x^* \leftarrow x'$, $W \leftarrow [\,]$;
    $s \leftarrow s + 1$, $\delta_s \leftarrow \delta_{s-1}\varphi$, $\varepsilon_s = \varepsilon_{s-1}\varphi$;
  else
    // Sampling is failed, update the failed sampling set;
    $W = W \cup \eta$;
  end if
end for
return x'.

filter out the less sensitive area of the new noise.

Under the guidance of current noise, CAB attack adaptively assigns the variance ratio of the normal distribution for each dimension according to the conclusion drawn from Section 3.2 and Fig. 2, and selects area most sensitive to noise to further reduce sampling space. Both processes take advantage of historical successful samplings. Although existing decision-based attacks directly discard failed samplings, they in fact contain information about decision boundaries. We amend the distribution's mean of next sample to keep away failed samplings:

$$\eta \sim \mathcal{N}(-\frac{1}{K}\sum_{j=1}^{K}\tilde{\eta_j}, z^{*2}),$$
$$s.t. \qquad F(x^* + \delta \cdot \frac{\tilde{\eta}}{\|\tilde{\eta}\|_2} + \varepsilon \cdot \frac{x - x^*}{\|x - x^*\|_2}) = F(x),$$

where K is the total number of failed samplings on the current adversarial example x^*, and $\tilde{\eta_j}$ is the normal random vector used at the j-th failed sampling. We maintain a sampling record of adversarial examples and save all the failed samplings about the current adversarial example x^* as $\tilde{x}$. The record is continually updated until a successful sampling occurs, i.e., the noise is further compressed. Since in the later stage of decision-based attack, the success rate of sampling decreases with the reduction of noise magnitude, maintaining the record can keep new samplings away from the historical failed ones. Algorithm 1 details CAB attack.

4. Experiments

4.1. Setup

CAB attack is tested on Tiny-Imagenet [1] and Imagenet [24] datasets with image size of $64 \times 64 \times 3$ and $224 \times 224 \times 3$, respectively. In experiment, we add adversarial noise to the validation set of Imagenet and Tiny-Imagenet, containing 50000 and 10000 images respectively, and input to eight different target models: Resnet-18 [14], Inception-v3 [28], Inception-Resnet v2 [27], NASNet [34], Resnet-101 , Dense-161[16] , VGG19 [26] and SENet-154 [15].

As for evaluation criteria, we compare CAB with other attacks by median noise magnitude:

$$mid = median(\{\|x' - x\|_2 \mid x \in \mathbf{X}\}), \qquad (11)$$

where x is an original image in the test set $\mathbf{X}$. x' is the adversarial example found that is closest to x. A smaller median ℓ_2 noise magnitude indicates that the attack method can better polish the adversarial noise under the same number of queries. It is worth noting that adversarial examples are rounded before being input to the target model for a more realistic black-box attack setting.

4.2. Comparison of Attack Effect

We report the comparison of median noise magnitude using different decision-based attacks in Table 1. The ℓ_2 norm of initial noise is shown in the first row. Last five rows in the table represent median noise magnitude under five different decision-based attacks. Two columns represent two datasets, Tiny-Imagenet (left) and Imagenet (right). For four models of each dataset, we use Curls [25], a state-of-the-art transfer-based iterative method, to attack each substitute-target model pair and input the generate adversarial example as starting point of decision-based attacks. Each element in the 6×2 table is a 4×4 matrix, where each row represents the substitute model used by Curls method and each column represents the target model. Therefore, elements on the diagonal are actually results of polished noise magnitude in white-box attack setting. mid of each attack method are calculated under the same amount of queries $B = 300$. Pixel retention rate $r = 0.2$ for our CAB attack. Stepsizes of spherical direction and source direction are $\delta_0 = 0.1, \varepsilon_0 = 0.003$ for Boundary, Biased Boundary, Evolutionary and CAB attacks. Decay factor φ is set to 0.99. For BBA [2], we use the version that incorporates information from a substitute model at each step.

It can be seen from Table 1 that CAB achieves the smallest noise magnitude on different target models with the same number of queries on all the black-box attacks, or the off-diagonal elements in each 4×4 matrix. Compared with boundary attack, CAB has a significant decrease on median noise magnitude, which validates the effectiveness of customizing the distribution with current noise and failed samplings. Since the magnitude of white-box noise (diagonal elements in each 4×4 matrix) tends to be rather small (less than 1/10 of black-box noise in ℓ_2 norm), the strategy used by Boundary that sampling over a standard normal distribution equally for each dimension is more suitable. As a result, white-box noise magnitude of CAB on Imagenet is slightly bigger than that of Boundary.

4.3. Ablation Study

The max query number B is the key parameter in CAB. We use inc-v3 as substitute model and res-18 as target model for three different transfer-based attacks, Curls [25], I-FGSM [18] and VR-IGSM [33], to generate initial adversarial examples. Curves of noise magnitude as B increases is shown in Fig. 3. A higher B provides decision-based attacks with more opportunities to fine-tune the adversarial noise. It can be seen that noise magnitude of CAB is lower than other methods at different query numbers. We recorded clock-time efficiency of each method on Imagenet with Densenet161 as target model. The average time required for Boundary and CAB to query each image 300 times are 15.31s and 13.39s, respectively. The efficiency of CAB is improved because it avoids sampling space where

		Tiny-Imagenet					Imagenet			
		res-18	inc-v3	inc-res	nasnet		res-101	dense	vgg-19	senet
Initial	res-18	0.076	1.617	1.833	2.002	res-101	0.325	3.060	3.232	4.367
	inc-v3	0.510	0.124	1.137	1.216	dense	2.777	0.269	3.112	4.241
	inc-res	0.552	1.083	0.147	1.088	vgg-19	6.050	6.034	0.183	5.135
	nasnet	0.577	1.272	1.199	0.134	senet	3.013	5.193	6.141	0.413
Whey	res-18	0.071	1.140	1.250	1.342	res-101	0.315	2.868	2.785	3.836
	inc-v3	0.360	0.121	0.865	0.909	dense	2.695	0.262	2.627	3.737
	inc-res	0.369	0.796	0.142	0.819	vgg-19	5.149	4.996	0.180	4.333
	nasnet	0.401	0.927	0.906	0.129	senet	2.708	4.529	5.101	0.393
Boundary	res-18	0.059	1.220	1.386	1.467	res-101	**0.279**	2.982	2.904	4.092
	inc-v3	0.384	0.110	1.001	1.041	dense	2.713	**0.231**	2.683	3.949
	inc-res	0.379	0.923	0.134	0.949	vgg-19	5.395	5.161	**0.155**	4.646
	nasnet	0.421	1.024	1.052	0.120	senet	2.738	4.695	5.761	**0.360**
Biased Boundary	res-18	0.072	1.129	1.283	1.358	res-101	0.318	2.586	2.704	3.745
	inc-v3	0.308	0.122	0.890	0.912	dense	2.441	0.263	2.496	3.542
	inc-res	0.325	0.813	0.144	0.829	vgg-19	4.776	4.402	0.181	4.133
	nasnet	0.332	0.928	0.924	0.132	senet	2.693	4.119	4.953	0.397
Evolutionary	res-18	0.068	0.951	1.112	1.147	res-101	0.310	2.518	2.373	3.217
	inc-v3	0.269	0.117	0.881	0.851	dense	2.394	0.256	2.253	3.128
	inc-res	0.292	0.761	0.138	0.797	vgg-19	4.112	4.036	0.176	3.442
	nasnet	0.301	0.849	0.888	0.126	senet	2.569	3.730	4.644	0.386
CAB	res-18	**0.058**	**0.935**	**0.929**	**1.030**	res-101	0.290	**2.387**	**1.953**	**3.116**
	inc-v3	**0.263**	**0.109**	**0.692**	**0.733**	dense	**2.294**	0.236	**2.025**	**3.051**
	inc-res	**0.251**	**0.689**	**0.131**	**0.682**	vgg-19	**3.982**	**3.730**	0.157	**3.413**
	nasnet	**0.284**	**0.757**	**0.726**	**0.119**	senet	**2.287**	**3.588**	**4.449**	0.366

Table 1. Median noise magnitude of five decision-based attacks against target models on two datasets.

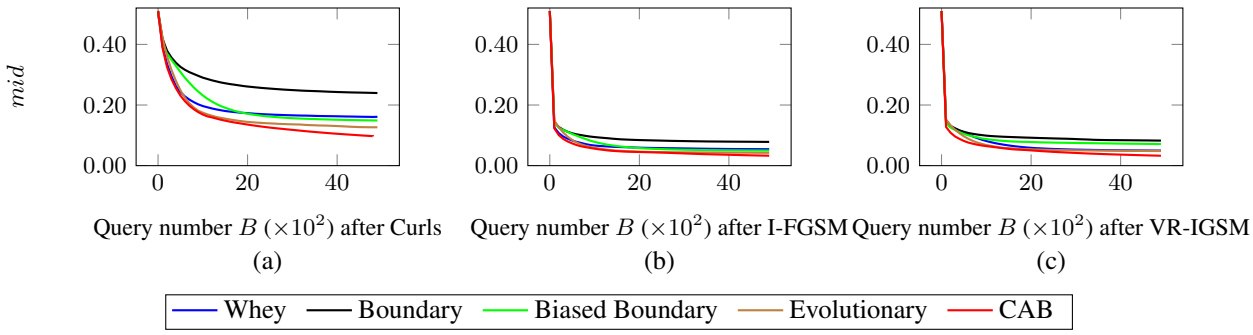


Figure 3. Median ℓ_2 distance of adversarial noise under different query number B on Tiny-Imagenet.

	x^*	VC	VC+SS	VC+EFS	CAB
median	0.496	0.314	0.293	0.304	**0.269**
average	2.066	1.275	1.264	1.236	**1.202**

Table 2. Comparison of noise magnitude on each step of CAB.

the noise magnitude increases with the help of historical information, reducing the the proportion of invalid samplings.

We further tested CAB attack using randomly initialized adversarial noise on MNIST [19] and CIFAR-10 [17] datasets. We use Resnet-50 [14] for MNIST and Dense-100 [16] for CIFAR-10 as target models. The initial adversarial examples are generated by adding uniform noise to the original images until misclassification. The number of queries B for decision-based attack is set to 300. As Table 3 shows, CAB can still generate smaller noise than other decision methods under random initialization. In order to further verify the effectiveness of each step in CAB attack, we compare median and mean adversarial noise [1] when using several steps independently or in combination. For Tiny-

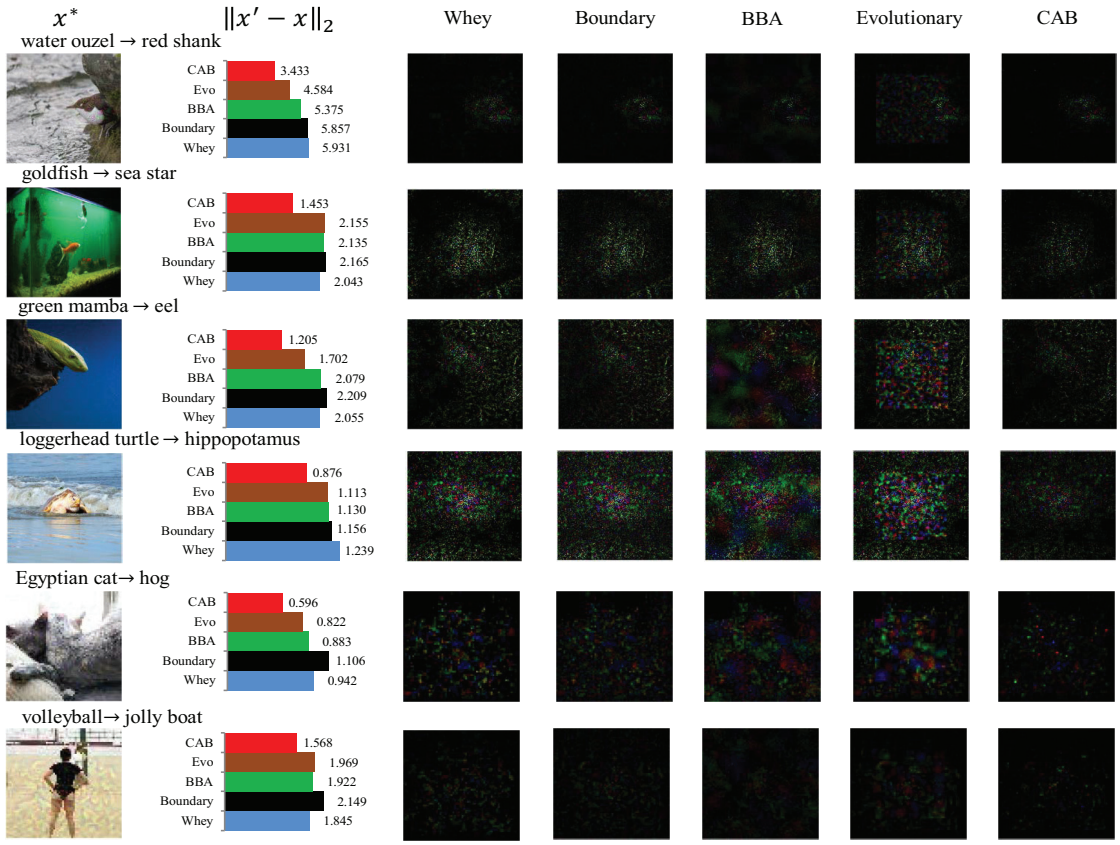


Figure 4. Comparison of adversarial noise generated by Whey, Boundary, Biased Boundary (BBA), Evolutionary (EVO) and our CAB attack. The label and misclassification category is noted above x^* of each row as $y \to F(x^*)$.

	MNIST	CIFAR-10
Initial	5.866	8.429
Whey	4.118	6.222
Boundary	4.147	6.689
Biased Boundary	4.062	6.414
Evolutionary	3.445	5.812
CAB	**3.163**	**5.414**

Table 3. Comparison of median ℓ_2 noise magnitude on MNIST and CIFAR-10 with random initialization.

Imagenet, we show that all three steps, the variance customization (VC), stepsize scheduling (SS), and exploitation of failed samplings (EFS) contribute to improving the noise reduction rate in Table 2. Fig. 4 compares adversarial noise generated by five different decision-based attacks on Imagenet (first 4 rows) and Tiny-Imagenet (last 2 rows). The first image of each row is the initial adversarial example x^* generated by Curls attack, followed by noise magnitude of five attacks. Polished noise (with magnitude enhanced for better visualization) of five methods are listed from left to right. Since CAB customizes sampling with current noise, area with higher noise is effectively suppressed, and thus a higher noise polishing efficiency is achieved.

5. Conclusion

In this paper, we propose Customized Adversarial Boundary, a new decision-based attack that uses current noise to select the sensitive area of images and customize sampling distribution. We reveal the relationship between variance and current noise to customize the sampling process of boundary attack. Moreover, we further customize the stepsize and initial noise with transfer-based attack to polish noise of decision-based attack through the monotonicity in the noise compression process. Extensive experiments on multiple datasets demonstrate that CAB achieves smaller median noise magnitude against a variety of target models than other decision-based attacks.

Acknowledgements

This work is supported by the NSFC (under Grant 61876130, 61932009).

References

[1] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Sharada P Mohanty, Florian Laurent, Marcel Salathé, Matthias Bethge, Yaodong Yu, et al. Adversarial vision challenge. In *The NeurIPS'18 Competition*, pages 129–153. Springer, 2020.

[2] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *ICCV*, pages 4958–4966, 2019.

[3] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[4] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *ICLR*, 2019.

[5] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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*, pages 15–26, 2017.

[6] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*, 2019.

[7] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *NeurIPS*, pages 10932–10942, 2019.

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Xiaolin Hu, Jianguo Li, and Jun Zhu. Boosting adversarial attacks with momentum. *CVPR*, 2018.

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.

[10] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, 2019.

[11] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

[12] Chris Finlay, Aram-Alexandre Pooladian, and Adam Oberman. The logbarrier adversarial attack: making effective use of decision boundary information. In *ICCV*, pages 4862–4870, 2019.

[13] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[16] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CVPR*, pages 2261–2269, 2017.

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009.

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Liu Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimization via variance reduction method. *arXiv preprint arXiv:1805.11811*, 2018.

[21] Yanpei Liu, Xinyun Chen, Cheng Chih Liu, and Dawn Xiaodong Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

[22] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

[23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, pages 372–387, 2016.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[25] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *CVPR*, 2019.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, page 12, 2017.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2013.

[30] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

[31] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI*, pages 742–749, 2019.

[32] Jonas Rauber Wieland Brendel and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[33] Lei Wu, Zhanxing Zhu, Cheng Tai, et al. Understanding and enhancing the transferability of adversarial examples. *arXiv preprint arXiv:1802.09707*, 2018.

[34] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018.