Consistency-Sensitivity Guided Ensemble Black-Box Adversarial Attacks in Low-Dimensional Spaces

Jianhe Yuan and Zhihai He
University of Missouri, Columbia MO
yuanjia, hezhi@missouri.edu

Abstract

Black-box attacks aim to generate adversarial noise to fail the victim deep neural network in the black box. The central task in black-box attack method design is to estimate and characterize the victim model in the high-dimensional model space based on feedback results of queries submitted to the victim network. The central performance goal is to minimize the number of queries needed for successful attack. Existing attack methods directly search and refine the adversarial noise in an extremely high-dimensional space, requiring hundreds or even thousands queries to the victim network. To address this challenge, we propose to explore a consistency and sensitivity guided ensemble attack (CSEA) method in a low-dimensional space. Specifically, we estimate the victim model in the black box using a learned linear composition of an ensemble of surrogate models with diversified network structures. Using random block masks on the input image, these surrogate models jointly construct and submit randomized and sparsified queries to the victim model. Based on these query results and guided by a consistency constraint, the surrogate models can be trained using a very small number of queries such that their learned composition is able to accurately approximate the victim model in the high-dimensional space. The randomized and sparsified queries also provide important information for us to construct an attack sensitivity map for the input image, with which the adversarial attack can be locally refined to further increase its success rate. Our extensive experimental results demonstrate that our proposed approach significantly reduces the number of queries to the victim network while maintaining very high success rates, outperforming existing black-box attack methods by large margins.

1. Introduction

Deep neural networks are sensitive to adversarial attacks [26, 2]. A very small amount of adversarial noise added to the input image can successfully fool the state-of-art classifier with a very high probability. There are two types of adversarial attacks, *white-box attacks* which have full access to the victim network under attack and *black-box attacks* which have no knowledge about the network. In this work, we focus on black-box attacks which remain an open and very challenging problem. In black-box attacks, it is assumed that the attacker can only query the victim network and obtain its output score for a given input image [19]. There are two major performance metrics, the number of queries to the victim network and the attack success rate, used to evaluate the performance of black-box attacks [9]. In black box attack research, the objective is to minimize the number of queries submitted to the victim network while achieving a very high attack success rate.

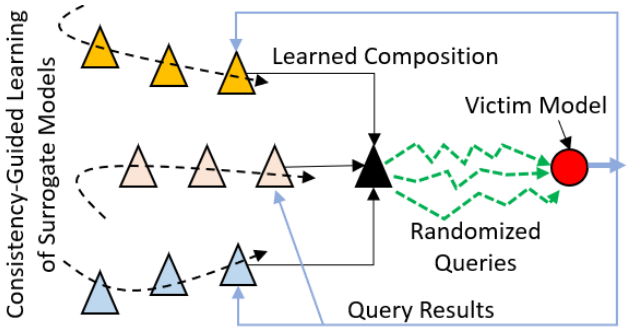

Figure 1. Illustration of the proposed idea for consistency-sensitivity guided ensemble attack in a low-dimensional space.

Two major approaches have been explored in the literature for black-box attacks. The first one is the *transfer-based* approach which uses a trained surrogate network to generate attacks based on the white-box approach, hoping this attack noise can be effectively transferred to the unknown target network [7, 12]. This approach often suffers from low success rates since the adversarial attack is a very sophisticated error accumulation process depending on the specific parameter settings of victim model and the input image. The second one is the *query-based* approach which queries the target network continuously, searches or modi-

fies the attack noise based on the query score feedback using gradient descent or other optimization methods [19, 21]. This approach often needs a larger number of queries since both the victim model and the input image have extremely high dimensions and the gradient-based searches and attack noise optimization in such high-dimensional spaces involve a very large number of search steps and queries. Currently, the average number of queries achieved by existing state-of-the-art methods remains very high, often in the range of a few hundreds or even thousands [19]. In practice, it is prohibitive or unrealistic to query the victim network to be attacked for a large number of times. For example, if an online face recognition service detects that a large number of queries have been issued from the same source, it can simply activate its defense mechanism and disable its access for security protection reasons. Therefore, it is highly desirable to significantly reduce the number of queries for black-box attacks while maintaining a very high success rate, which is the central goal of this work.

In both targeted and untargeted attacks, the attack is successful if the output score of the victim network for the attacked image is significantly deviated from the correct score of the original image [13, 19]. Therefore, the attack success rate depends on how well the victim model is approximated and how effectively the detailed network responses of the input image are characterized and exploited. The **central challenge** in black-box attacks is the high dimensionality [13]. It is well known that the search complexity typically increases exponentially with the number of dimensions. In black box attacks, the adversarial noise has the same dimension as the input image size, which is very large. In the meantime, the surrogate model aims to approximate the victim model in the black box which has an unknown network structure with millions of model parameters.

To address this challenge, we propose a consistency-sensitivity guided ensemble attack (CSEA) method for highly efficient search and estimate the victim model in the high-dimensional model space. The **central idea** is illustrated in Figure 1. To prevent the search process from being trapped in local minimums, we construct an ensemble of surrogate models with diversified structures which perform collaborative search in the high-dimensional model space. We then estimate or approximate the victim model using a learned linear composition of these surrogate models. Using random block masks on the input image, these surrogate models jointly construct and submit randomized and sparsified queries to the victim model. Based on the feedback results of these highly diversified queries, the surrogate models are able to effectively learn and evolve themselves in the model space. Guided by a consistency constraint, their learned composition is able to approximate the victim network very efficiently using a very small number of queries. Furthermore, the block-wise randomized and sparsified queries provide important information for us to estimate the attack sensitivity map for the input image. Using this sensitivity map, we can perform block-based local refinement of the attack to further increase its success rate. Our experimental results will demonstrate that this CSEA approach is able to significantly reduce (by up to 50%) the number of needed queries to the victim network to achieve successful attacks, when compared to the state-of-the-art black-box attack methods.

2. Related Work and Unique Contributions

There are two major threat models of adversarial attacks: *white-box attacks* and *black-box attacks*. The white-box attacker has full access to the classifier network parameters, network architecture, and weights. During the past a few years, a number of white-box attack methods have been developed, including fast gradient sign (FGS) method [8], Jacobian-based saliency map attack (J-BSMA) [24], projected gradient descent (PGD) attack [16, 18], and *Backward Pass Differentiable Approximation (BPDA)* attack [3]. The Basic Iterative Method (BIM) is an improved version of the FGS attack method. Generative adversarial networks (GANs) have been used in [27, 5] to generate perturbations. It has been recognized in [15, 28] that the PGD is the strongest attacker among all attacks, which can be viewed as a multi-step variant of FGS^k [18]. For black-box attacks, the attacker has no knowledge about target network. Existing black-box attack methods can be broadly divided into two categories: transfer-based and search-based approaches.

(a) Transfer-based methods. This attack method learns a surrogate network to generate the adversarial attack based on white-box attack methods and wishes the attack can be transferred to the victim network [23]. [22] studied the transferrability between deep neural networks and other models such as decision trees and support vector machines. It has been recognized [17] that black-box attacks with surrogate networks tend not to transfer well. [17] introduced an ensemble of networks to improve the attack performance. Despite its potential to reduce the number of queries, transfer-based black-box attacks still suffer from high failure rates.

(b) Search-based methods. Assuming that the output scores of the target model is available, [6] proposed a zero-order optimization (ZOO) method to estimate the gradients of target model based on symmetric difference quotient. [13] presented a Natural Evolution Strategies (NES) method coupled with PGD attacks to generate adversarial inputs. These methods often require thousands of queries to succeed on the target model. To reduce the number of queries, random grouping and principal components analysis have been used in [21] to improve the gradient estimation. However, the robustness of its gradient estimation is sensitive to

the choices of hyper-parameters, such as the learning rate, decay rates, and updating rule [19].

Local search methods perform local modification of the image, aiming to fail the target network. [20] proposed a simply yet efficient approach by adding noise to the image in a specific direction, and check if it leads to a positive change in the attack probability. [1] proposed a random search based method, which selects localized square shaped updates at random positions so that at each iteration the perturbation is situated approximately at the boundary of the feasible set. [4] presented a distribution based black-box attack, which uses the image structure information for modeling adversarial distributions and reduces the required queries. Although effective in attacking the target network, existing local search methods often suffer from the need for a large number of queries.

Unique contributions of this work are summarized in the following, when compared to existing methods for black-box attacks. (1) This work aims to bridge the gap between and address the major limitations of existing transfer-based and search-based approaches. We have explored a new approach for searching and estimating the black-box victim model in a very high-dimensional model space. (2) By representing and approximating the victim network using a learned linear composition of the a small set of surrogate models, and also by partitioning the image into a small set of blocks to sparsely modulate the adversarial noise, our CSEA method is able to effectively perform the search and optimization in a low-dimensional parameter space, resulting in significantly reduced number of queries. (3) If we assume that the attacker has memory, being able to remember the queries from previous input images, the number of needed queries can be further significantly reduced.

3. Method

3.1. Problem Formulation

Consider a target classifier $\boldsymbol{T}(\cdot)$ in the black box whose network structure and model parameters are unknown. This target classifier is also referred to as the victim network or model. Let $(\boldsymbol{x}, \boldsymbol{y})$ be the input-label pair. Existing black box attack methods assume that the attacker has access to the output score $\boldsymbol{T}(\boldsymbol{x})$ of the victim network. The goal of black box attacks is to generate an adversarial noise $\boldsymbol{z}$ which has the same dimension as $\boldsymbol{x}$ and maximizes the score deviation

$$\boldsymbol{z} = \arg\max_{\boldsymbol{z}} ||\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{T}(\boldsymbol{x}+\boldsymbol{z})||_2, \ s.t. \ ||\boldsymbol{z}||_p \leq \epsilon, \quad (1)$$

where ϵ controls the magnitude of the adversarial noise. Query-based attack methods attempt to search for the target adversarial noise $\boldsymbol{z}$ directly in the spatial domain by analyzing its gradients or distributions [13, 19]. In transfer-based attacks, the noise $\boldsymbol{z}$ is generated using a surrogate network $\boldsymbol{Q}$ with existing white-box attack methdos, such as FGSM, PGD, or BPDA [17, 12],

$$\boldsymbol{z} = \boldsymbol{Q}^{-1}(\boldsymbol{s}) - \boldsymbol{x}, \quad \boldsymbol{s} = \boldsymbol{Q}(\boldsymbol{x}+\boldsymbol{z}). \quad (2)$$

Here, $\boldsymbol{s}$ is the output score of the surrogate network $\boldsymbol{Q}$. In this work, we refer to $\boldsymbol{s}$ as the attack anchor. Certainly, if we set a different value for $\boldsymbol{s}$, the corresponding attack noise will be also different. This generated attack is then submitted to the victim network for evaluation and corresponding output score is given by

$$\tilde{\boldsymbol{y}} = \boldsymbol{T}(\boldsymbol{Q}^{-1}(\boldsymbol{s}, \boldsymbol{x})). \quad (3)$$

The output score depends on the surrogate network $\boldsymbol{Q}$ and the input image $\boldsymbol{x}$.

3.2. Constructing Ensemble Randomized Searches

In our CSEA method, we propose to approximate the victim model $\boldsymbol{T}$ using a learned linear composition of an small ensemble of surrogate models with diversified network structures $\{\boldsymbol{Q}_k^t | 1 \leq k \leq K\}$. In our experiments, $K = 3$. The composition model $\boldsymbol{Q}^t$ is given by

$$\boldsymbol{Q}^t = \sum_{k=1}^{K} a_k^t \cdot \boldsymbol{Q}_k^t, \quad (4)$$

where t is the training iteration index. According to our experiments, a small number K, for example, $k = 2$ or 3, will be sufficient. In order to train these surrogate models, we use the composition model $\boldsymbol{Q}^t$ to generate the adversarial attack $\boldsymbol{z}^t$ for the input image $\boldsymbol{x}$. To ensure successful training and diversify the training samples, we propose to incorporate a blockwise mask to modulate the adversarial noise. Specifically, we partition the input image $\boldsymbol{x}$ into $B_H \times B_W$ blocks. For example, in our experiments, we set B_H and B_W to be 16. Let $\mathcal{M}_\alpha^t$ be a random binary blockwise mask generated at iteration t with all ones in $\alpha \times B_H \times B_W$ blocks and zeros in other blocks. For example, in our experiments, we set α to be 0.15.

Based on this random mask, the current white-box attack method can be directly modified such that there is no adversarial noise at image position (i, j) if $\mathcal{M}_\alpha^t(i, j) = 0$. For example, during the FGSM (Fast Gradient Sign method) attack [8], with the random mask, the attacked image is given by

$$\tilde{\boldsymbol{x}}^t(i, j) = \boldsymbol{x}(i, j) + \mathcal{M}_\alpha^t(i, j) \cdot \epsilon \cdot sign[\nabla_x J(\boldsymbol{Q}^t, \boldsymbol{x}, \boldsymbol{s}^t)].$$

Here, $J(\cdot)$ represents the loss function at the network output layer, $\nabla_x J(\cdot)$ represents the gradients of this loss function $J(\cdot)$ being propagated to the input image $\boldsymbol{x}$. $\boldsymbol{s}^t$ is the target score for the attack. According to our experiments, when this blockwise mask is used, a small number of extra iterations will be needed to achieve successful attack of the input

image. The corresponding adversarial noise is denoted by $\boldsymbol{z}^t = \tilde{\boldsymbol{x}}^t - \boldsymbol{x}$. Figure 2 shows two examples of randomized attacks. In each row, the first image is the original image, the second one is the attacked image with the mask shown in the third image. The fourth one is another attacked image with a random block mask shown in the fifth image.

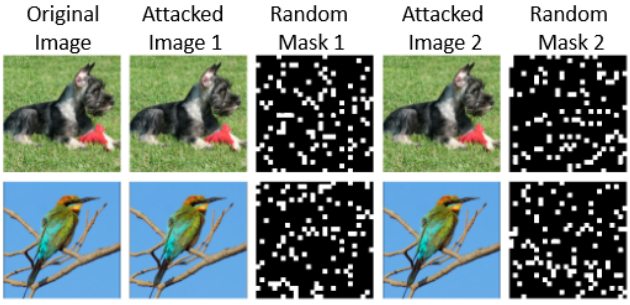

Figure 2. Original images (first column), their adversarial image (second and fourth column), and adversarial perturbations (third and fifth column)

3.3. Learned Linear Composition of Surrogate Models

In this section, we will explain how the proposed CSEA can successfully learn these surrogate models such that their linear composition can very efficiently approximate the victim model in the black box. Our objective is to minimize the number of queries submitted to the victim model. The victim model $\boldsymbol{T}$ has an unknown structure and unknown model parameters. In this work, we create a small set of K surrogate models with randomized network structures. Specifically, we take an existing network model, such as ResNet or InceptionNet. The randomized network structure is obtained by adding random connections between layers and performing random drop out in existing layers. As illustrated in Figure 3, these surrogate models $\{\boldsymbol{Q}_k^t\}$ are refined based on queries from the victim network. We wish that a linear combination of these surrogate models will able to successfully capture the behavior of the victim network $\boldsymbol{T}$ under adversarial attacks, as described in (4).

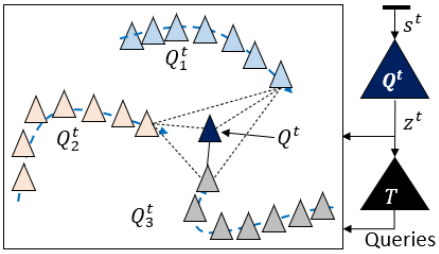

Figure 3. Illustration of learned linear composition of surrogate model.

It should be noted that the surrogate network only has access to one image, which is the current test image $\boldsymbol{x}$. We assume that other training images are not available to ensure fair comparisons with other black-box attack methods, especially those query-based attack methods [19, 9]. Using the attacked images $\tilde{\boldsymbol{x}}^t$ as inputs, we query the victim network and the output is denoted by $\boldsymbol{y}^t = \boldsymbol{T}(\tilde{\boldsymbol{x}}^t)$. The sample set $\{(\tilde{\boldsymbol{x}}^\tau, \boldsymbol{y}^\tau)|0 \leq \tau \leq t\}$, represents the behavior of the target network under adversarial attacks. We use this sample set to train the surrogate networks so that their output scores $\boldsymbol{Q}_k^t(\tilde{\boldsymbol{x}}^\tau)$ approach the target $\boldsymbol{T}(\tilde{\boldsymbol{x}}^\tau)$. To this end, the following attack loss is used for training network $\boldsymbol{Q}^k$

$$\mathcal{L}_A^k = \boldsymbol{E}_{\tilde{\boldsymbol{x}}^\tau \in \Omega_0}\{||\boldsymbol{Q}_k^t(\tilde{\boldsymbol{x}}^t) - \boldsymbol{T}(\tilde{\boldsymbol{x}}^t)||_2^2\}, \tag{5}$$

where Ω_0 represents all the queries $\tilde{\boldsymbol{x}}^\tau$ that have been evaluated by the victim network. In our proposed CSEA approach, we approximate the victim network using a linear composition of these surrogate models, as described in (4). To this end, we need to select the composition coefficients that minimize the approximation error

$$\begin{aligned} &\min_{\{a_k^t\}} \sum_{\tau=0}^{t} ||\boldsymbol{Q}^t(\tilde{\boldsymbol{x}}_k^\tau) - \boldsymbol{y}_k^\tau||_2^2 \\ =&\min_{\{a_k^t\}} \sum_{\tau=0}^{t} ||\sum_{k=1}^{K} a_k^t \cdot \boldsymbol{Q}_k^t(\tilde{\boldsymbol{x}}_k^\tau) - \boldsymbol{y}_i^\tau||_2^2. \end{aligned} \tag{6}$$

This leads to a least mean squared error (LMSE) problem whose solution $\boldsymbol{a} = [a_1^t, a_2^t, \cdots, a_K^t]^T$ is given by

$$\boldsymbol{a} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^T\boldsymbol{y}), \tag{7}$$

where

$$\boldsymbol{\Phi}^T = \begin{bmatrix} \boldsymbol{Q}_1(\tilde{x}^0) & \boldsymbol{Q}_2(\tilde{x}^0) & \cdots & \boldsymbol{Q}_K(\tilde{x}^0) \\ \boldsymbol{Q}_1(\tilde{x}^1) & \boldsymbol{Q}_2(\tilde{x}^1) & \cdots & \boldsymbol{Q}_K(\tilde{x}^1) \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{Q}_1(\tilde{x}^t) & \boldsymbol{Q}_2(\tilde{x}^t) & \cdots & \boldsymbol{Q}_K(\tilde{x}^t) \end{bmatrix}, \tag{8}$$

and $\boldsymbol{y}^T = [\boldsymbol{y}(0)^T, \boldsymbol{y}(1)^T, \cdots, \boldsymbol{y}(t)^T]$. Once the composition coefficients are obtained, we can construct the linear composition of the surrogate model, $\boldsymbol{Q}^t$.

Adversarial consistency constraint. As illustrated in Fig. 3, our goal is to evolve the models $\boldsymbol{Q}_k^t$ so that their linear combination can approach the unknown victim network $\boldsymbol{T}$ in the high-dimensional model space. This implies that all surrogate models should converge to the victim network model. In other words, they should share the same response to different adversarial attacks. We refer to this requirement as the adversarial consistency constraint. It should be noted that, when enforcing the adversarial consistency constraint, we do not need to query the victim network, which is very important in our algorithm design to reduce the query complexity. To implement this constraint, we use the composition surrogate model $\boldsymbol{Q}^t$ to generate an extra set of attacked

images $\{\tilde{\boldsymbol{x}}_l^t | 1 \leq l \leq L\}$ using different block masks and attack anchors $\boldsymbol{s}^t$. We then define the follow adversarial consistency loss between different surrogate networks

$$\mathcal{L}_C = \sum_{l=1}^{L} \sum_{i \neq j} ||\boldsymbol{Q}_i^t(\tilde{\boldsymbol{x}}_l^t) - \boldsymbol{Q}_j^t(\tilde{\boldsymbol{x}}_l^t)||_2 \tag{9}$$

This consistency loss is then combined with the attack loss in (5) as follows

$$\mathcal{L}^k = \mathcal{L}_A^k + \lambda \cdot \mathcal{L}_C \tag{10}$$

to train the surrogate network $\boldsymbol{Q}_k^t$. Figure 4 shows two examples of the training process. For the each test image, it shows that both the attack loss and the consistency loss are decreasing quickly with the number of iterations t. It also shows the target class score ($\times 10^{-2}$) which is the classification score for the correct class. When this score drops towards 0, it implies that the image has been successfully attacked.

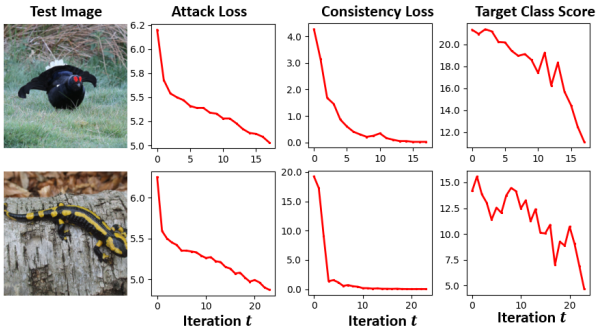

Figure 4. The convergence of the training process for the surrogate models. The figure shows the test images, their attack loss, consistency loss, and the target class score of victim model for each test image.

3.4. Sensitivity-Guided Local Refinement of Attack

In our study, we observe that some images, for example, about 3-7% of the images, even after the transfer-based attack by the surrogate network $\boldsymbol{Q}$ with optimized anchors, still survives. In other words, the adversarial noise generated by $\boldsymbol{Q}$ still cannot successfully attack these images. Recent studies have demonstrated that modifying image blocks based on frequent queries to and feedback from the target network $\boldsymbol{T}$ can successfully attack $\boldsymbol{T}$ [9]. However, they often need large numbers of queries. To address this issue, we propose to utilize the results from the existing randomized queries and the learned surrogate model to generate an sensitivity map and perform guided local refinement of the adversarial noise.

Our idea is that the attack should be locally concentrated onto those image regions that contribute most to the network decision or image recognition result. We observe that, during adversarial attacks, these image regions often experience relatively larger attacks, specifically, larger gradient responses. Based on this observation, during the process of surrogate network learning with randomized adversarial noise generation as described in Sections 3.3, we record average gradient response at each image regions and use this to construct the sensitivity map.

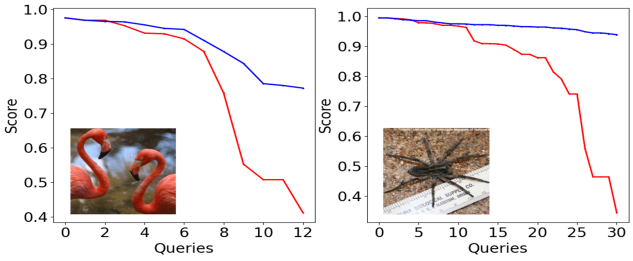

Figure 5. The accumulative effect of block-based attacks.

In the following, we explain another reason why the randomized sparse block mask was used for generating the queries. In our randomized queries, we partition the image into $M = B_H B_W$ blocks. Let $\{B_m | 1 \leq m \leq M\}$ be the set of image blocks. Let $\boldsymbol{z}$ be the adversarial attack generated by the composition surrogate model $\boldsymbol{Q}^t$. We define the adversarial block noise $\boldsymbol{z}_{B_m}$. Then, each $\boldsymbol{z}_{B_m}$ is contributing the overall attack performance. In our experiments, we find out that the attack results of these adversarial block noises are highly correlated. For example, Figure 5 shows the attack scores for two test images. The red curve shows the drop of the attack score if we assume that a set of neighboring block noises $\{\boldsymbol{z}_{Bm}\}$ are independent and their scores can accumulate. The blue curve shows the actual accumulated attack score by this set of neighboring adversarial block noises. We can see that there is a significant difference between them, which indicates that the neighboring block noises are highly correlated with each other. This correlation will create a significant challenge for us to estimate the contribution of each adversarial block noise or the sensitivity of each image block. In our work, we observe that, by introducing the block mask $\mathcal{M}_\alpha^t$ which is random and sparse, the correlation between blocks can be significantly reduced. This allows us to perform the following estimation of the attack sensitivity map. We use the FGSM attack as an example. The attacker computes the gradient of the loss function $L(\boldsymbol{Q}^t(\vec{f}), \boldsymbol{s}^t)$ with respect to the feature map $\vec{f} = \mathcal{F}^l(i, j, c)$ at network layer l, image location (i, j) of channel c and modifies the feature map as

$$\vec{f}' = \vec{f} - \epsilon \cdot \nabla_{\vec{f}} L(\boldsymbol{Q}^t(\vec{f}), \boldsymbol{s}^t). \tag{11}$$

Here, $\boldsymbol{s}^t$ is the attack anchor and $\boldsymbol{Q}^t(\vec{f})$ represents the network output at iteration t, ϵ controls the magnitude of the adversarial attack noise. We denote the gradient

$\nabla_{\vec{f}} L(\boldsymbol{y}(\vec{f}\,), \boldsymbol{s}^t)$ at location (i, j) and channel c for iteration t by $\delta^t(i, j, c)$. If the average value of $|\delta^t(i, j, c)|$ at location (i, j) is larger than $|\delta^t(i', j', c)|$ at location (i', j'), then the location (i, j) is more important or more sensitive than (i', j') from the network attack perspective. Based on this observation, we define the sensitivity map for the input image $\boldsymbol{x}$ as

$$\mathcal{A}(i, j) = \sum_t \sum_c |\delta^t(i, j, c) \cdot \mathcal{M}_\alpha(i, j)|. \qquad (12)$$

Figure 6 shows five examples of attack sensitivity maps. We can see that the high-sensitivity areas are concentrating on those semantic structure regions. Once the sensitivity weights $\mathcal{A}(i, j)$ is obtained, each location (i, j) represents a block in the original image. We will follow the order of sensitivity to perform local block-based refinement of the adversarial noise using the method outlined in [9]. Specifically, we partition the attack noise z into blocks. For each block, we try to flip noise by multiplying the noise block with -1. If this flipped noise block improves the attack performance, pushing the output score of the victim network away from its correct value, then this noise block is flipped. Otherwise, it remains the same. We perform this block-based local noise refinement by following the order of sensitivity levels of blocks. Our experimental results will demonstrate that this sensitivity-guided attack will be able to reduce the number queries and improve the attack success rate.

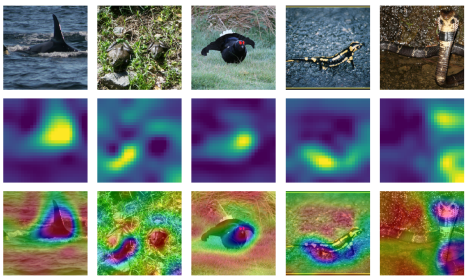

Figure 6. Original images (top), their sensitivity maps (middle), and sensitivity maps overlaid on the images (bottom).

4. Further Discussions

Compared to existing approaches for black-box attack, our proposed CSEA method is unique and novel in the follow aspects. By approximating the victim model in the black box using a learned linear composition of a small set of surrogate models and using a consistency constraint to regulate their training processes, we can learn an accurate composition surrogate model using a very small number of queries. The randomized queries also introduce diversity into the training samples, which also enhances the learning performance. The sparsified blockwise adversarial noise generation allows us to estimate the contribution or attack sensitivity of each image block. Based on this sensitivity map, we can then perform local refinement of the attack noise to further increase its attack success rate. It should be noted that the major performance metrics of black box attack are the number of queries to the victim network and the overall attack success rate. As in the black box attack literature [11, 13, 29], we do not need to worry about the computational complexity and the number of queries to the surrogate models.

5. Experimental Results

In this section, we will evaluate the performance of our CSEA method and conduct performance comparisons with the state-of-the-art methods. Our experiments run on a GTX 1080 Ti GPU with PyTorch [25]. Following existing papers on black-box attacks [19], we consider both untargeted and targeted ℓ_∞ attacks on the CIFAR-10 and ImageNet datasets.

5.1. Experimental Setup

Following prior works [11, 13, 29], we set the maximum ℓ_∞ adversarial perturbations within a range of $\epsilon = 8/255$ for CIFAR-10 and $\epsilon = 16/255$ for ImageNet. On the CIFAR-10 dataset, we consider two target networks (also called the victim models): (a) ResNet-Preact-110 [10], which yields 8.4% top-1 error rate on the test set; and (b) DenseNet-BC-110, which achieves a 6.97% top-1 test error rate. Follow the setting in [29], we consider two target models on ImageNet dataset: (a) an Inception-v3 with 22.7% top-1 error rate on the standard validation set; and (b) a PNAS-Net-5-Large model with top-1 error rate of 17.26%; and (c) SE-Net with top-1 error rate of 17.0%. At the testing phase, as in existing papers, we use the PGD [18] attack with a step size of $\epsilon = 1/255$ and 10 attack iterations. We choose 1,000 images randomly from the testing set in CIFAR-10 and the validation set of ImageNet.

5.2. Comparison with the State-of-the-Art methods

For CIFAR-10, we compare our CSEA method with the following six methods: (1) **Bandits** [14]; (2) **SimBA** [9]; (3) **Subspace Attack** [29]; (4) **P-RGF** [7]; (5) **TREMBA** [12]; and (6) **NP-Attack** [4]. We adopt the CIFAR-shakeshake26 as the architecture of our baseline surrogate model, and train it using the Adam optimizer with a learning rate of $\eta = 0.002$. Table 3.4 summarizes the results. It compares the mean number of queries, median number of queries, and attack failure rate. From the table, we can see that, for untargeted attacks with ResNet as the victim model, we are able to reduce the mean queries by about 38% and the median queries by 50%, when compared with the current best method. For the targeted attack which is more

Table 1. Performance of our different black-box attacks with ℓ_∞ constraint under untargeted and targeted (target class being 0) setting. The maximum perturbation is $\epsilon = 8/255$ for CIFAR-10 dataset.

Victim Model	Method	Untargeted Attack			Targeted Attack		
		Mean	Median	Failure Rate	Mean	Median	Failure Rate
ResNet	Bandits [14] (ICLR18)	193.4	88.0	9.2%	3660.1	2812.0	27.4%
	SimBA [9] (ICML19)	432.1	235.0	6.8%	940.0	885.0	**0.0%**
	Subspace Attack [29] (NIPS19)	301.8	12.0	7.0%	2409.3	1630.0	22.0%
	P-RGF [7] (NIPS19)	121.8	62.0	7.8%	1020.8	**390.0**	29.4%
	TREMBA [12] (ICLR20)	120.7	64.0	9.1%	1125.3	868.0	8.8%
	NP-Attack [4] (ECCV20)	144.0	(NA)	0%	936	(NA)	0%
	CSEA (This Work)	74.2	31.0	0.0%	930.4	511.0	3.0%
	CSEA-Memory	**43.2**	**4.0**	**0.0%**	**860.4**	432.0	2.8%
DenseNet	Bandits [14] (ICLR18)	206.3	96.0	4.0%	4154.8	3842.0	20.0%
	SimBA [9] (ICML19)	480.5	223.0	26.0%	838.8	777.0	0.0%
	Subspace Attack [29] (NIPS19)	115.8	12.0	4.0%	1528.4	1012.0	6.0%
	P-RGF [7] (NIPS19)	111.7	62.0	0.4%	1037.1	**438.0**	22.9%
	TREMBA [12] (ICLR20)	126.4	66.0	2.2%	1123.4	8879.0	7.7%
	CSEA (This Work)	65.4	40.0	0.0%	753.1	521.0	0.0%
	CSEA-Memory	**11.4**	**4.0**	**0.0%**	**596.2**	460.0	**0.0%**

challenging, our performance is also quite competitive with existing methods. It reduces the mean queries from 936 to 860. If the victim model is DenseNet, our CSEA method can reduce the mean queries for untargeted attack by almost 40%.

In the table, we also reported the results for CSEA-Memory, which assumes that the attacker has memory to accumulate the query results of previous test images. These accumulated query results are used together to train the surrogate model. We can see that number of queries can be significantly reduced since previous queries results have been utilized to train a more efficient surrogate network.

On the ImageNet, we compare the performance with five state-of-the-art methods: (1) **Bandits** [14]; (2) **NES**: [13] query the output scores of the target network to approximate the true gradient; (3) **Subspace Attack** [29]; (4) **NP-Attack** [4]; and (5) **Square Attack** [1]. Following these papers, we use 1,000 images from the validation set for evaluation, and set the maximum queries to 1,500 per image to find the adversarial perturbations. We report the mean and median queries for each attack, together with the failure rate.

As shown in Table 5.2, we can see that that our CSEA method outperforms the state-of-the-arts by large margins on the mean queries for the PNAS-Net model. Although the failure rate of our method is slightly higher than the subspace method, the mean queries have been reduced by almost 42%. We also present results using SENet as the target classifier. As shown in Table 5.2, we can see that our proposed method significantly reduces the query number and the failure rate. For example, our method is able to reduce the mean query from 456 to 360 and the failure rate also drops 0.2% when compared with Subspace Attack, which is the current best attacking method on this dataset.

5.3. Ablation Studies

In the following experiments, we conduct ablation studies to further understand our CSEA method.

5.3.1 Contributions of Each Algorithm Component

Our CSEA attack method consists of two attack steps guided by the surrogate network: direct transfer-based attack by the composition surrogate network and sensitivity-guided local refinement of the attack. Table 5.2 shows the successful rates at these two stages. For example, with the ResNet victim model, using the direct transfer attack only, the success rate is 90.9%. For those unsuccessful images after the transfer attack, we then add the sensitivity-guided local attack, the success rate becomes 100%. Table 5.2 also report results for the DenseNet victim model. These results suggest that each layer of the CSEA attack is important with significant contributions to the overall performance.

5.3.2 The Impact of Local Sensitivity

In our sensitivity-guided local refinement, we use the surrogate network to generate the sensitivity map to guide the local block-based optimization of the attack. We follow the order of the sensitivity weights to search for the best attack noise. In the following experiment, for those images that have not been successfully attacked by the previous attack, we perform the local block-based attack with and without the sensitivity guidance (in a random order). Table 5.2 summarizes the results. We can see that, if we randomly choose

Table 2. Performance of our attack method with ℓ_∞ constraint under untargeted setting. The maximum perturbation is $\epsilon = 16/255$ for ImageNet dataset. Some papers did not report results on some victim networks, which are marked with NA.

Method	Inception-v3			PNAS-Net			SENet		
	Mean	Median	Failure Rate	Mean	Median	Failure Rate	Mean	Median	Failure Rate
NES [13] (ICML18)	1427	800	19.3%	2182	1300	38.5%	1759	900	17.9%
Bandits [14] (ICLR18)	887	222	4.2%	1437	552	12.1%	1055	300	6.4%
Subspace Attack [29] (NIPS19)	462	96	1.1%	680	160	**4.2%**	456	66	1.9%
NP-Attack [4] (ECCV20)	867	(NA)	0%	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Square Attack [1] (ECCV20)	197	24	0.3%	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
CSEA (This Work)	**190**	42	1.8%	**398**	**64**	4.9%	**360**	**60**	**1.7%**

Table 3. Performance of our method at each stage.

Victim Model	ResNet	DenseNet
Direct Transfer Attack	90.9%	96.1%
+ Sensitivity-Guided Local Attack	100.0%	100.0%

Table 4. Sensitivity-based attack using different searching method under untargeted setting on CIFAR-10.

Attack Method	Mean Queries	Success Rate
Random Order	140.7	100.0%
Sensitivity-Guided	113.3	100.0%

of order of blocks, the mean number of queries is 140.7. If we use the sensitivity-guided attack, the mean number of queries is reduced to 113.3, which is quite significant.

5.4. Success rates for Different Number of Queries

In this section, we compare our method with the state-of-art methods on the attack performance with different numbers of queries. Figure 7 shows the success rate of our method versus the number of queries on the CIFAR-10 dataset with two different network architectures, ResNet and DenseNet. The performance gap in the success rate becomes larger in the range of 100 queries for both models. It shows that our method outperforms all other methods by more than 10% in the attack success rate. We can see that our method achieves high attack success rate much more efficiently with smaller number of queries.

More experimental results and ablation studies are provided in the **Supplemental Materials**.

6. Conclusion

In this work, we have developed a consistency-sensitivity guided ensemble attack approach in a low-dimensional space for representation, generation, and optimization of the adversarial attack. We approximate the victim model using a learned linear composition of a small set of surrogate models with randomized network structures. Based on randomized queries to the victim network and guided by a consistency constraint, the surrogate models

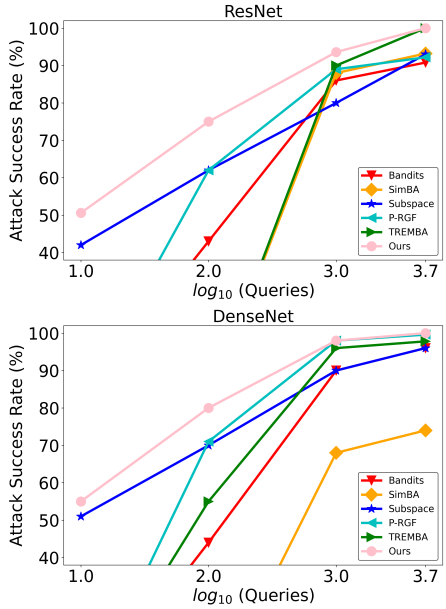

Figure 7. Comparison of success rate versus number of model queries across different attack method.

can be trained using a very small number of queries such that their learned composition is able to accurately approximate the victim model. Based on the randomly sparsified queries, we also construct an attack sensitivity map for the input image to guide local refinement of the attack noise to further increase its success rate. Our extensive experimental results demonstrate that our proposed approach significantly reduces the number of queries to the victim network while maintaining very high success rates, outperforming existing black-box attack methods by large margins.

7. Acknowledgement

This work was supported in part by National Science Foundation under grants 1647213 and 1646065. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 3, 7, 8

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 1

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pages 274–283, 2018. 2

[4] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. Improving query efficiency of black-box adversarial attack. *arXiv preprint arXiv:2009.11508*, 2020. 3, 6, 7, 8

[5] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*, 2017. 2

[6] 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. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec '17*, 2017. 2

[7] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior, 2019. 1, 6, 7

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3

[9] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks, 2019. 1, 4, 5, 6, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Ri cognition (CVPR)*, June 2016. 6

[11] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. *Lecture Notes in Computer Science*, page 646–661, 2016. 6

[12] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020. 1, 3, 6, 7

[13] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018. 2, 3, 6, 7, 8

[14] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors, 2018. 6, 7, 8

[15] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing, 2018. 2

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks, 2016. 2, 3

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 6

[19] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization, 2019. 1, 2, 3, 4, 6

[20] N. Narodytska and S. Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318, 2017. 3

[21] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[22] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016. 2

[23] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017. 2

[24] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. pages 372–387, 2016. 2

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[27] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 2

[28] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018. 2

[29] Ziang Yan, Yiwen Guo, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks, 2019. 6, 7, 8