AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks

Chun-Chen Tu,[1*] **Paishun Ting,**[1*] **Pin-Yu Chen,**[2*] **Sijia Liu,**[2]
Huan Zhang,[3] **Jinfeng Yi,**[4] **Cho-Jui Hsieh,**[3] **Shin-Ming Cheng**[5,6]
[1]University of Michigan, Ann Arbor, USA
[2]MIT-IBM Watson AI Lab, IBM Research
[3]University of California, Los Angeles, USA
[4]JD AI Research, Beijing, China
[5]National Taiwan University of Science and Technology, Taiwan
[6]Academia Sinica, Taiwan

Abstract

Recent studies have shown that adversarial examples in state-of-the-art image classifiers trained by deep neural networks (DNN) can be easily generated when the target model is transparent to an attacker, known as the white-box setting. However, when attacking a deployed machine learning service, one can only acquire the input-output correspondences of the target model; this is the so-called black-box attack setting. The major drawback of existing black-box attacks is the need for excessive model queries, which may give a false sense of model robustness due to inefficient query designs. To bridge this gap, we propose a generic framework for query-efficient black-box attacks. Our framework, **AutoZOOM**, which is short for **Auto**encoder-based **Z**eroth **O**rder **O**ptimization **M**ethod, has two novel building blocks towards efficient black-box attacks: (i) an adaptive random gradient estimation strategy to balance query counts and distortion, and (ii) an autoencoder that is either trained offline with unlabeled data or a bilinear resizing operation for attack acceleration. Experimental results suggest that, by applying AutoZOOM to a state-of-the-art black-box attack (ZOO), a significant reduction in model queries can be achieved without sacrificing the attack success rate and the visual quality of the resulting adversarial examples. In particular, when compared to the standard ZOO method, AutoZOOM can consistently reduce the mean query counts in finding successful adversarial examples (or reaching the same distortion level) by at least 93% on MNIST, CIFAR-10 and ImageNet datasets, leading to novel insights on adversarial robustness.

1 Introduction

In recent years, “machine learning as a service” has offered the world an effortless access to powerful machine learning tools for a wide variety of tasks. For example, commercially available services such as Google Cloud Vision API and Clarifai.com provide well-trained image classifiers to the public. One is able to upload and obtain the class prediction results for images at hand at a low price. However, the existing and emerging machine learning platforms and their low model-access costs raise ever-increasing security concerns, as they also offer an ideal environment for testing malicious attempts. Even worse, the risks can be amplified when these services are used to build derived products such that the inherent security vulnerability could be leveraged by attackers.

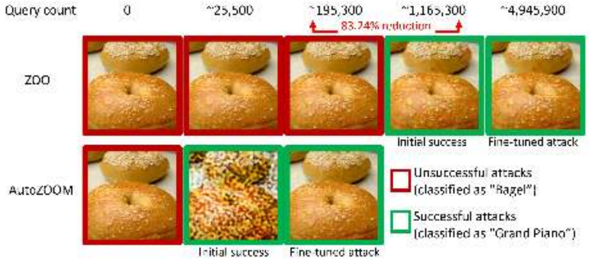


Figure 1: AutoZOOM significantly reduces the number of queries required to generate a successful adversarial Bagel image from the black-box Inception-v3 model.

In many computer vision tasks, DNN models achieve the state-of-the-art prediction accuracy and hence are widely deployed in modern machine learning services. Nonetheless, recent studies have highlighted DNNs’ vulnerability to adversarial perturbations. In the *white-box* setting in which the target model is entirely transparent to an attacker, visually imperceptible adversarial images can be easily crafted to fool a target DNN model towards misclassification by leveraging the input gradient information (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). However, in the *black-box* setting in which the parameters of the deployed model are hidden and one can only observe the input-output correspondences of a queried example, crafting adversarial examples requires a gradient-free (zeroth order) optimization approach to gather necessary attack information. Figure 1 displays a prediction-evasive adversarial example crafted via iterative model queries from a black-box DNN (the Inception-v3 model (Szegedy et al. 2016)) trained on ImageNet.

Albeit achieving remarkable attack effectiveness by the use of gradient estimation, current black-box attack methods,

[*] equal contribution

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

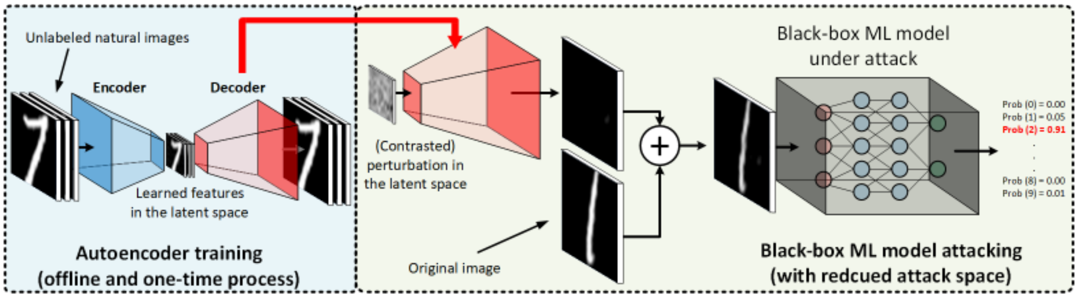

Unlabeled natural images
Encoder
Decoder
Learned features in the latent space
Autoencoder training (offline and one-time process)
(Contrasted) perturbation in the latent space
Original image
Black-box ML model under attack
Prob (0) = 0.00
Prob (1) = 0.05
Prob (2) = 0.91
Prob (8) = 0.00
Prob (9) = 0.01
Black-box ML model attacking (with redcued attack space)

Figure 2: Illustration of attack dimension reduction through a "decoder" in AutoZOOM for improving query efficiency in black-box attacks. The decoder has two modes: (i) An autoencoder (AE) trained on unlabeled natural images that are different from the attacked images and training data; (ii) a simple bilinear image resizer (BiLIN) that is applied channel-wise to extrapolate low-dimensional feature to the original image dimension (width × height). In the latter mode, no additional training is required.

such as (Chen et al. 2017; Nitin Bhagoji et al. 2018), are not query-efficient since they exploit coordinate-wise gradient estimation and value update, which inevitably incurs an excessive number of model queries and may give a false sense of model robustness due to inefficient query designs. In this paper, we propose to tackle the preceding problem by using **AutoZOOM**, an **Auto**encoder-based **Z**eroth **O**rder **O**ptimization **M**ethod. AutoZOOM has two novel building blocks: (i) a new and adaptive random gradient estimation strategy to balance the query counts and distortion when crafting adversarial examples, and (ii) an autoencoder that is either trained offline on other unlabeled data, or based on a simple bilinear resizing operation, in order to accelerate black-box attacks. As illustrated in Figure 2, AutoZOOM utilizes a "decoder" to craft a high-dimensional adversarial perturbation from the (learned) low-dimensional latent-space representation, and its query efficiency can be well explained by the dimension-dependent convergence rate in gradient-free optimization.

Contributions. We summarize our main contributions and new insights on adversarial robustness as follows:

1. We propose AutoZOOM, a novel query-efficient black-box attack framework for generating adversarial examples. AutoZOOM features an adaptive random gradient estimation strategy and dimension reduction techniques (either an offline trained autoencoder or a bilinear resizer) to reduce attack query counts while maintaining attack effectiveness and visual similarity. To the best of our knowledge, AutoZOOM is the first black-box attack using random full gradient estimation and data-driven acceleration.
2. We use the convergence rate of zeroth-order optimization to motivate the query efficiency of AutoZOOM and provide an error analysis of the new gradient estimator in AutoZOOM to the true gradient for characterizing the trade-offs between estimation error and query counts.
3. When applied to a state-of-the-art black-box attack proposed in (Chen et al. 2017), AutoZOOM attains a similar attack success rate while achieving a significant reduction (at least 93%) in the mean query counts required to attack the DNN image classifiers for MNIST, CIFAR-10 and ImageNet. It can also fine-tune the distortion in the post-success stage by performing finer gradient estimation.
4. In the experiments, we also find that AutoZOOM with a simple bilinear resizer as the decoder (AutoZOOM-BiLIN) can attain noticeable query efficiency, despite that it is still worse than AutoZOOM with an offline trained autoencoder (AutoZOOM-AE). However, AutoZOOM-BiLIN is easier to be mounted as no additional training is required. The results also suggest an interesting finding that while learning effective low-dimensional representations of legitimate images is still a challenging task, black-box attacks using significantly less degree of freedoms (i.e., reduced dimensions) are certainly plausible.

2 Related Work

Gradient-based adversarial attacks on DNNs fall within the white-box setting, since acquiring the gradient with respect to the input requires knowing the weights of the target DNN. As a first attempt towards black-box attacks, the authors in (Papernot et al. 2017) proposed to train a substitute model using iterative model queries, performing white-box attacks on the substitute model, and implementing transfer attacks to the target model (Papernot, McDaniel, and Goodfellow 2016; Liu et al. 2017). However, its attack performance can be severely degraded due to poor attack transferability (Su et al. 2018). Although ZOO achieves a similar attack success rate and comparable visual quality as many white-box attack methods (Chen et al. 2017), its coordinate-wise gradient estimation requires excessive target model evaluations and is hence not query-efficient. The same gradient estimation technique is also used in (Nitin Bhagoji et al. 2018).

Beyond optimization-based approaches, the authors in (Ilyas et al. 2018) proposed to use a natural evolution strategy (NES) to enhance query efficiency. Although there is a vector-wise gradient estimation step in the NES attack, we treat it as a parallel work since its natural evolutionary step is out of the scope of black-box attacks using zeroth-order gradient descent. We also note that different from NES, our AutoZOOM framework uses a theory-driven query-efficient random-vector based gradient estimation strategy. In addition, AutoZOOM could be applied to further improve the query efficiency of NES, since NES does not take into account the factor of attack dimension reduction, which is the novelty in AutoZOOM as well as the main focus of this paper.

Under a more restricted attack setting, where only the decision (top-1 prediction class) is known to an attacker, the authors in (Brendel, Rauber, and Bethge 2018) proposed a random-walk based attack around the decision boundary. Such a black-box attack dispenses class prediction scores and hence requires additional model queries. Due to space limitation, we provide more background and a table comparing existing black-box attacks in the supplementary material.

3 AutoZOOM: Background and Methods

3.1 Black-box Attack Formulation and Zeroth Order Optimization

Throughout this paper, we focus on improving the query efficiency of gradient-estimation and gradient-descent based black-box attacks empowered by AutoZOOM, and we consider the threat model that the class prediction scores are known to an attacker. In this setting, it suffices to denote the target DNN as a classification function $F : [0,1]^d \mapsto \mathbb{R}^K$ that takes a d-dimensional scaled image as its input and yields a vector of prediction scores of all K image classes, such as the prediction probabilities for each class. We further consider the case of applying an entry-wise monotonic transformation $M(F)$ to the output of F for black-box attacks, since monotonic transformation preserves the ranking of the class predictions and can alleviate the problem of large score variation in F (e.g., probability to log probability).

Here we formulate black-box targeted attacks. The formulation can be easily adapted to untargeted attacks. Let $(\mathbf{x}_0, t_0)$ denote a natural image $\mathbf{x}_0$ and its ground-truth class label t_0, and let $(\mathbf{x}, t)$ denote the adversarial example of $\mathbf{x}_0$ and the target attack class label $t \neq t_0$. The problem of finding an adversarial example can be formulated as an optimization problem taking the generic form of

$$\min_{\mathbf{x}\in[0,1]^d} \mathrm{Dist}(\mathbf{x}, \mathbf{x}_0) + \lambda \cdot \mathrm{Loss}(\mathbf{x}, M(F(\mathbf{x})), t), \quad (1)$$

where $\mathrm{Dist}(\mathbf{x}, \mathbf{x}_0)$ measures the distortion between $\mathbf{x}$ and $\mathbf{x}_0$, $\mathrm{Loss}(\cdot)$ is an attack objective reflecting the likelihood of predicting $t = \arg\max_{k\in\{1,\ldots,K\}}[M(F(\mathbf{x}))]_k$, λ is a regularization coefficient, and the constraint $\mathbf{x} \in [0,1]^d$ confines the adversarial image $\mathbf{x}$ to the valid image space. The distortion $\mathrm{Dist}(\mathbf{x}, \mathbf{x}_0)$ is often evaluated by the L_p norm defined as $\mathrm{Dist}(\mathbf{x}, \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|_p = \|\boldsymbol{\delta}\|_p = \sum_{i=1}^{d} |\boldsymbol{\delta}_i|^{1/p}$ for $p \geq 1$, where $\boldsymbol{\delta} = \mathbf{x} - \mathbf{x}_0$ is the adversarial perturbation to $\mathbf{x}_0$. The attack objective $\mathrm{Loss}(\cdot)$ can be the training loss of DNNs (Goodfellow, Shlens, and Szegedy 2015) or some designed loss based on model predictions (Carlini and Wagner 2017).

In the white-box setting, an adversarial example is generated by using downstream optimizers such as ADAM (Kingma and Ba 2015) to solve (1); this requires the gradient $\nabla f(\mathbf{x})$ of the objective function $f(\mathbf{x}) = \mathrm{Dist}(\mathbf{x}, \mathbf{x}_0) + \lambda \cdot \mathrm{Loss}(\mathbf{x}, M(F(x)), t)$ relative to the input of F via back-propagation in DNNs. However, in the black-box setting, acquiring $\nabla f(\cdot)$ is implausible, and one can only obtain the function evaluation $F(\cdot)$, which renders solving (1) a zeroth order optimization problem. Recently, zeroth order optimization approaches (Ghadimi and Lan 2013; Nesterov and Spokoiny 2017; Liu et al. 2018) circumvent the preceding challenge by approximating the true gradient via function evaluations. Specifically, in black-box attacks, the gradient estimate is applied to both gradient computation and descent in the optimization process for solving (1).

3.2 Random Vector based Gradient Estimation

As a first attempt to enable gradient-free black-box attacks on DNNs, the authors in (Chen et al. 2017) use the symmetric difference quotient method (Lax and Terrell 2014) to evaluate the gradient $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}$ of the i-th component by

$$g_i = \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h} \approx \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \quad (2)$$

using a small h. Here $\mathbf{e}_i$ denotes the i-th elementary basis. Albeit contributing to powerful black-box attacks and applicable to large networks like ImageNet, the nature of coordinate-wise gradient estimation step in (2) must incur an enormous amount of model queries and is hence not query-efficient. For example, the ImageNet dataset has $d = 299 \times 299 \times 3 \approx 270{,}000$ input dimensions, rendering coordinate-wise zeroth order optimization based on gradient estimation query-inefficient.

To improve query efficiency, we dispense with coordinate-wise estimation and instead propose a scaled random full gradient estimator of $\nabla f(\mathbf{x})$, defined as

$$\mathbf{g} = b \cdot \frac{f(\mathbf{x} + \beta\mathbf{u}) - f(\mathbf{x})}{\beta} \cdot \mathbf{u}, \quad (3)$$

where $\beta > 0$ is a smoothing parameter, $\mathbf{u}$ is a unit-length vector that is uniformly drawn at random from a unit Euclidean sphere, and b is a tunable scaling parameter that balances the bias and variance trade-off of the gradient estimation error. Note that with $b = 1$, the gradient estimator in (3) becomes the one used in (Duchi et al. 2015). With $b = d$, this estimator becomes the one adopted in (Gao, Jiang, and Zhang 2014). We will provide an optimal value b^* for balancing query efficiency and estimation error in the following analysis.

Averaged random gradient estimation. To effectively control the error in gradient estimation, we consider a more general gradient estimator, in which the gradient estimate is averaged over q random directions $\{\mathbf{u}_j\}_{j=1}^q$. That is,

$$\overline{\mathbf{g}} = \frac{1}{q} \sum_{j=1}^{q} \mathbf{g}_j, \quad (4)$$

where $\mathbf{g}_j$ is a gradient estimate defined in (3) with $\mathbf{u} = \mathbf{u}_j$. The use of multiple random directions can reduce the variance of $\overline{\mathbf{g}}$ in (4) for convex loss functions (Duchi et al. 2015; Liu et al. 2018).

Below we establish an error analysis of the averaged random gradient estimator in (4) for studying the influence of the parameters b and q on estimation error and query efficiency.

Theorem 1. *Assume $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable and its gradient $\nabla f(\cdot)$ is L-Lipschitz*[1]*. Then the mean squared*

[1] A function $W(\cdot)$ is L-Lipschitz if $\|W(\mathbf{w}_1) - W(\mathbf{w}_2)\|_2 \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2$ for any $\mathbf{w}_1, \mathbf{w}_2$. For DNNs with ReLU activations, L can be derived from the model weights (Szegedy et al. 2014).

estimation error of $\overline{\mathbf{g}}$ *in (4) is upper bounded by*

$$\mathbb{E}\|\overline{\mathbf{g}} - \nabla f(\mathbf{x})\|_2^2 \leq 4(\frac{b^2}{d^2} + \frac{b^2}{dq} + \frac{(b-d)^2}{d^2})\|\nabla f(\mathbf{x})\|_2^2 + \frac{2q+1}{q} b^2 \beta^2 L^2. \quad (5)$$

Proof. The proof is given in the supplementary file. □

Here we highlight the important implications based on Theorem 1: (i) The error analysis holds when f is *non-convex*; (ii) In DNNs, the true gradient ∇f can be viewed as the numerical gradient obtained via back-propagation; (iii) For any fixed b, selecting a small β (e.g., we set $\beta = 1/d$ in AutoZOOM) can effectively reduce the last error term in (5), and we therefore focus on optimizing the first error term; (iv) The first error term in (5) exhibits the influence of b and q on the estimation error, and is independent of β. We further elaborate on (iv) as follows. Fixing q and let $\eta(b) = \frac{b^2}{d^2} + \frac{b^2}{dq} + \frac{(b-d)^2}{d^2}$ to be the coefficient of the first error term in (5), then the optimal b that minimizes $\eta(b)$ is $b^* = \frac{dq}{2q+d}$. For query efficiency, one would like to keep q small, which then implies $b^* \approx q$ and $\eta(b^*) \approx 1$ when the dimension d is large. On the other hand, when $q \to \infty$, $b^* \approx d/2$ and $\eta(b^*) \approx 1/2$, which yields a smaller error upper bound but is query-inefficient. We also note that by setting $b = q$, the coefficient $\eta(b) = \frac{b^2}{d^2} + \frac{b^2}{dq} + \frac{(b-d)^2}{d^2} \approx 1$ and thus is independent of the dimension d and the parameter q.

Adaptive random gradient estimation. Based on Theorem 1 and our error analysis, in AutoZOOM we set $b = q$ in (3) and propose to use an adaptive strategy for selecting q. AutoZOOM uses $q = 1$ (i.e., the fewest possible model evaluation) to first obtain rough gradient estimates for solving (1) until a successful adversarial image is found. After the initial attack success, it switches to use more accurate gradient estimates with $q > 1$ to fine-tune the image quality. The trade-off between q (which is proportional to query counts) and distortion reduction will be investigated in Section 4.

3.3 Attack Dimension Reduction via Autoencoder

Dimension-dependent convergence rate using gradient estimation. Different from the first order convergence results, the convergence rate of zeroth order gradient descent methods has an additional multiplicative dimension-dependent factor d. In the convex loss setting the rate is $O(\sqrt{d/T})$, where T is the number of iterations (Nesterov and Spokoiny 2017; Liu et al. 2018; Gao, Jiang, and Zhang 2014; Wang et al. 2018). The same convergence rate has also been found in the nonconvex setting (Ghadimi and Lan 2013). The dimension-dependent convergence factor d suggests that vanilla black-box attacks using gradient estimations can be query inefficient when the (vectorized) image dimension d is large, due to the curse of dimensionality in convergence. This also motivates us to propose using an autoencoder to reduce the attack dimension and improve query efficiency in black-box attacks.

In AutoZOOM, we propose to perform random gradient estimation from a reduced dimension $d' < d$ to improve query efficiency. Specifically, as illustrated in Figure 2, the additive perturbation to an image $\mathbf{x}_0$ is actually implemented through a "decoder" $D : \mathbb{R}^{d'} \mapsto \mathbb{R}^d$ such that $\mathbf{x} = \mathbf{x}_0 + D(\boldsymbol{\delta}')$, where $\boldsymbol{\delta}' \in \mathbb{R}^{d'}$. In other words, the adversarial perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ to $\mathbf{x}_0$ is in fact generated from a dimension-reduced space, with an aim of improving query efficiency due to the reduced dimension-dependent factor in the convergence analysis. AutoZOOM provides two modes for such a decoder D:

• An autoencoder (AE) trained on unlabeled data that are different from the training data to learn reconstruction from a dimension-reduced representation. The encoder $E(\cdot)$ in an AE compresses the data to a low-dimensional latent space and the decoder $D(\cdot)$ reconstructs an example from its latent representation. The weights of an AE are learned to minimize the average L_2 reconstruction error. Note that training such an AE for black-box adversarial attacks is one-time and is entirely offline (i.e., no model queries needed).

• A simple channel-wise bilinear image resizer (BiLIN) that scales a small image to a large image via bilinear extrapolation[2]. Note that no additional training is required for BiLIN.

Why AE? Our proposal of AE is motivated by the insightful findings in (Goodfellow, Shlens, and Szegedy 2015) that a successful adversarial perturbation is highly relevant to some human-imperceptible noise pattern resembling the shape of the target class, known as the "shadow". Since a decoder in AE learns to reconstruct data from latent representations, it can also provide distributional guidance for mapping adversarial perturbations to generate these shadows.

We also note that for any reduced dimension d', the setting $b^* = q$ is optimal in terms of minimizing the corresponding estimation error from Theorem 1, despite the fact that the gradient estimation errors of different reduced dimensions cannot be directly compared. In Section 4 we will report the superior query efficiency in black-box attacks achieved with the use of AE or BiLIN as the decoder, and discuss the benefit of attack dimension reduction.

Algorithm 1 AutoZOOM for black-box attacks on DNNs

```
Input: Black-box DNN model $F$, original example $\mathbf{x}_0$,
distortion measure $\mathrm{Dist}(\cdot)$, attack objective $\mathrm{Loss}(\cdot)$, monotonic transformation $M(\cdot)$, decoder $D(\cdot) \in \{\mathrm{AE}, \mathrm{BiLIN}\}$,
initial coefficient $\lambda_{\mathrm{ini}}$, query budget $Q$
while query count $\leq Q$ do
    1. Exploration: use $\mathbf{x} = \mathbf{x}_0 + D(\boldsymbol{\delta}')$ and apply the random gradient estimator in (4) with $q = 1$ to the downstream optimizer (e.g., ADAM) for solving (1) until an initial attack is found.
    2. Exploitation (post-success stage): continue to fine-tune the adversarial perturbation $D(\boldsymbol{\delta}')$ for solving (1) while setting $q \geq 1$ in (4).
end while
Output: Least distorted successful adversarial example
```

3.4 AutoZOOM Algorithm

Algorithm 1 summarizes the AutoZOOM framework towards query-efficient black-box attacks on DNNs. We also note that

[2] See tf.image.resize_images, a TensorFlow example.

AutoZOOM is a general acceleration tool that is compatible with any gradient-estimation based black-box adversarial attack obeying the attack formulation in (1). It also has some theoretical estimation error guarantees and query-efficient parameter selection based on Theorem 1. The details on adjusting the regularization coefficient λ and the query parameter q based on run-time model evaluation results will be discussed in Section 4. Our source code is publicly available[3].

4 Performance Evaluation

This section presents the experiments for assessing the performance of AutoZOOM in accelerating black-box attacks on DNNs in terms of the number of queries required for an initial attack success and for a specific distortion level.

4.1 Distortion Measure and Attack Objective

As described in Section 3, AutoZOOM is a query-efficient gradient-free optimization framework for solving the black-box attack formulation in (1). In the following experiments, we demonstrate the utility of AutoZOOM by using the same attack formulation proposed in ZOO (Chen et al. 2017), which uses the squared L_2 norm as the distortion measure $\text{Dist}(\cdot)$ and adopts the attack objective

$$\text{Loss} = \max\{\max_{j \neq t} \log[F(\mathbf{x})]_j - \log[F(\mathbf{x})]_t\}, 0\}, \quad (6)$$

where this hinge function is designed for targeted black-box attacks on the DNN model F, and the monotonic transformation $M(\cdot) = \log(\cdot)$ is applied to the model output.

4.2 Comparative Black-box Attack Methods

We compare **AutoZOOM-AE** (D = AE) and **AutoZOOM-BiLIN** (D = BiLIN) with two different baselines: (i) Standard **ZOO** implementation[4] with bilinear scaling (same as BiLIN) for dimension reduction; (ii) **ZOO+AE**, which is ZOO with AE. Note that all attacks indeed generate adversarial perturbations based on the same reduced attack dimension.

4.3 Experiment Setup, Evaluation, Datasets and AutoZOOM Implementation

We assess the performance of different attack methods on several representative benchmark datasets, including MNIST (LeCun et al. 1998), CIFAR-10 (Krizhevsky 2009) and ImageNet (Russakovsky et al. 2015). For MNIST and CIFAR-10, we use the same DNN image classification models[5] as in (Carlini and Wagner 2017). For ImageNet, we use the Inception-v3 model (Szegedy et al. 2016). All experiments were conducted using TensorFlow Machine-Learning Library (Abadi et al.) on machines equipped with an Intel Xeon E5-2690v3 CPU and an Nvidia Tesla K80 GPU.

All attacks used ADAM (Kingma and Ba 2015) for solving (1) with their estimated gradients and the same initial learning rate 2×10^{-3}. On MNIST and CIFAR-10, all methods adopt 1,000 ADAM iterations. On ImageNet, ZOO and ZOO+AE adopt 20,000 iterations, whereas AutoZOOM-BiLIN and AutoZOOM-AE adopt 100,000 iterations. Note that due to different gradient estimation methods, the query counts (i.e., the number of model evaluations) per iteration of a black-box attack may vary. ZOO and ZOO+AE use the parallel gradient update of (2) with a batch of 128 pixels, yielding 256 query counts per iteration. AutoZOOM-BiLIN and AutoZOOM-AE use the averaged random full gradient estimator in (4), resulting in $q+1$ query counts per iteration. For a fair comparison, the query counts are used for performance assessment.

Query reduction ratio. We use the mean query counts of ZOO with the smallest λ_{ini} as the baseline for computing the query reduction ratio of other methods and configurations.

TPR and initial success. We report the true positive rate (TPR), which measures the percentage of successful attacks fulfilling a pre-defined constraint ℓ on the normalized (per-pixel) L_2 distortion, as well as their query counts of first successes. We also report the per-pixel L_2 distortions of initial successes, where an initial success refers to the first query count that finds a successful adversarial example.

Post-success fine-tuning. When implementing AutoZOOM in Algorithm 1, on MNIST and CIFAR-10 we find that AutoZOOM without fine-tuning (i.e., $q = 1$) already yields similar distortion as ZOO. We note that ZOO can be viewed as coordinate-wise fine-tuning and is thus query-inefficient. On ImageNet, we will investigate the effect of post-success fine-tuning on reducing distortion.

Autoencoder Training. In AutoZOOM-AE, we use convolutional autoencoders for attack dimension reduction, which are trained on unlabeled datasets that are different from the training dataset and the attacked natural examples. The implementation details are given in the supplementary material.

Dynamic Switching on λ. To adjust the regularization coefficient λ in (1), in all methods we set its initial value $\lambda_{\text{ini}} \in \{0.1, 1, 10\}$ on MNIST and CIFAR-10, and set $\lambda_{\text{ini}} = 10$ on ImageNet. Furthermore, for balancing the distortion Dist and the attack objective Loss in (1), we use a *dynamic switching* strategy to update λ during the optimization process. Per every S iterations, λ is multiplied by 10 times of the current value if the attack has never been successful. Otherwise, it divides its current value by 2. On MNIST and CIFAR-10, we set $S = 100$. On ImageNet, we set $S = 1,000$. At the instance of initial success, we also reset $\lambda = \lambda_{\text{ini}}$ and the ADAM parameters to the default values, as doing so can empirically reduce the distortion for all attack methods.

4.4 Black-box Attacks on MNIST and CIFAR-10

For both MNIST and CIFAR-10, we randomly select 50 correctly classified images from their test sets, and perform targeted attacks on these images. Since both datasets have 10 classes, each selected image is attacked 9 times, targeting at all but its true class. For all attacks, the ratio of reduced attack-space dimension to the original one (i.e., d'/d) is 25% for MNIST and 6.25% for CIFAR-10.

Table 1 shows the performance evaluation on MNIST with various values of λ_{ini}, the initial value of the regularization coefficient λ in (1). We use the performance of ZOO with $\lambda_{\text{ini}} = 0.1$ as a baseline for comparison. For example, with $\lambda_{\text{ini}} = 0.1$ and 10, the mean query counts required by AutoZOOM-AE to attain an initial success is reduced by

[3] https://github.com/IBM/Autozoom-Attack

[4] https://github.com/huanzhang12/ZOO-Attack

[5] https://github.com/carlini/nn_robust_attacks

Table 1: Performance evaluation of black-box targeted attacks on MNIST

Method	λ_{ini}	Attack success rate (ASR)	Mean query count (initial success)	Mean query count reduction ratio (initial success)	Mean per-pixel L_2 distortion (initial success)	True positive rate (TPR)	Mean query count with per-pixel L_2 distortion ≤ 0.004
ZOO	0.1	99.44%	35,737.60	0.00%	3.50×10^{-3}	96.76%	47,342.85
	1	99.44%	16,533.30	53.74%	3.74×10^{-3}	97.09%	31,322.44
	10	99.44%	13,324.60	62.72%	4.85×10^{-3}	96.31%	41,302.12
ZOO+AE	0.1	99.67%	34,093.95	4.60%	3.43×10^{-3}	97.66%	44,079.92
	1	99.78%	15,065.52	57.84%	3.72×10^{-3}	98.00%	29,213.95
	10	99.67%	12,102.20	66.14%	4.66×10^{-3}	97.66%	38,795.98
AutoZOOM-BiLIN	0.1	99.89%	2,465.95	93.10%	4.51×10^{-3}	96.55%	3,941.88
	1	99.89%	879.98	97.54%	4.12×10^{-3}	97.89%	2,320.01
	10	99.89%	612.34	98.29%	4.67×10^{-3}	97.11%	4,729.12
AutoZOOM-AE	0.1	**100.00%**	2,428.24	**93.21%**	4.54×10^{-3}	96.67%	3,861.30
	1	**100.00%**	729.65	**97.96%**	4.13×10^{-3}	96.89%	1,971.26
	10	**100.00%**	510.38	**98.57%**	4.67×10^{-3}	97.22%	4,855.01

Table 2: Performance evaluation of black-box targeted attacks on CIFAR-10

Method	λ_{ini}	Attack success rate (ASR)	Mean query count (initial success)	Mean query count reduction ratio (initial success)	Mean per-pixel L_2 distortion (initial success)	True positive rate (TPR)	Mean query count with per-pixel L_2 distortion ≤ 0.0015
ZOO	0.1	97.00%	25,538.43	0.00%	5.42×10^{-4}	100.00%	25,568.33
	1	97.00%	11,662.80	54.33%	6.37×10^{-4}	100.00%	11,777.18
	10	97.00%	10,015.08	60.78%	8.03×10^{-4}	100.00%	10,784.54
ZOO+AE	0.1	99.33%	19,670.96	22.98%	4.96×10^{-4}	100.00%	20,219.42
	1	99.00%	5,793.25	77.32%	6.83×10^{-4}	99.89%	5,773.24
	10	99.00%	4,892.80	80.84%	8.74×10^{-4}	99.78%	5,378.30
AutoZOOM-BiLIN	0.1	99.67%	2,049.28	91.98%	1.01×10^{-3}	98.77%	2,112.52
	1	99.67%	813.01	96.82%	8.25×10^{-4}	99.22%	1,005.92
	10	99.33%	623.96	97.56%	9.09×10^{-4}	98.99%	835.27
AutoZOOM-AE	0.1	**100.00%**	1,523.91	**94.03%**	1.20×10^{-3}	99.67%	1,752.45
	1	**100.00%**	332.43	**98.70%**	1.01×10^{-3}	99.56%	345.62
	10	**100.00%**	259.34	**98.98%**	1.15×10^{-3}	99.67%	990.61

93.21% and **98.57%**, respectively. One can also observe that allowing larger λ_{ini} generally leads to fewer mean query counts at the price of slightly increased distortion for the initial attack. The noticeable huge difference in the required attack query counts between AutoZOOM and ZOO/ZOO+AE validates the effectiveness of our proposed random full gradient estimator in (3), which dispenses with the coordinate-wise gradient estimation in ZOO but still remains comparable true positive rates, thereby greatly improving query efficiency.

For CIFAR-10, we report similar query efficiency improvements as displayed in Table 2. In particular, comparing the two query-efficient black-box attack methods (AutoZOOM-BiLIN and AutoZOOM-AE), we find that AutoZOOM-AE is more query-efficient than AutoZOOM-BiLIN, but at the cost of an additional AE training step. AutoZOOM-AE achieves the highest attack success rates (ASRs) and mean query reduction ratios for different values of λ_{ini}. In addition, their true positive rates (TPRs) are similar but AutoZOOM-AE usually takes fewer query counts to reach the same L_2 distortion. We note that when $\lambda_{\text{ini}} = 10$, AutoZOOM-AE has a higher TPR but also needs slightly more mean query counts than AutoZOOM-BiLIN to reach the same L_2 distortion. This suggests that there are some adversarial examples that are difficult for a bilinear resizer to reduce their post-success distortions but can be handled by an AE.

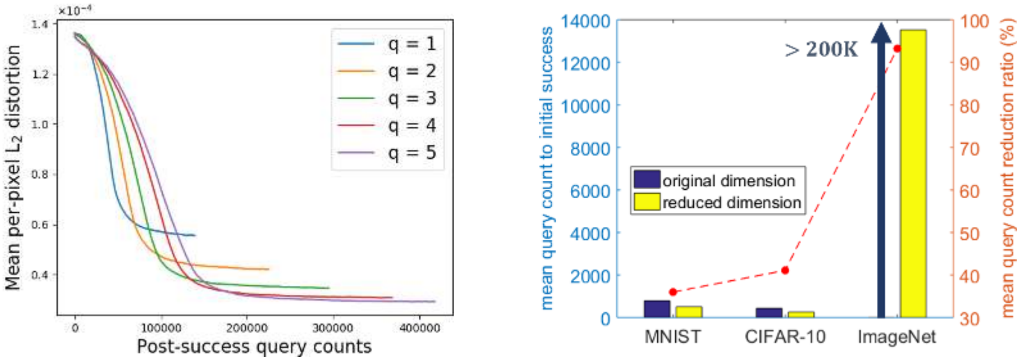
4.5 Black-box Attacks on ImageNet

We selected 50 correctly classified images from the ImageNet test set to perform random targeted attacks and set $\lambda_{\text{ini}} = 10$ and the attack dimension reduction ratio to 1.15%. The results are summarized in Table 3. Note that comparing to ZOO, AutoZOOM-AE can significantly reduce the query count required to achieve an initial success by 99.39% (or 99.35% to reach the same L_2 distortion), which is a remarkable improvement since this means reducing more than *2.2 million* model queries given the fact that the dimension of ImageNet ($\approx$ 270K) is much larger than that of MNIST and CIFAR-10.

Post-success distortion refinement. As described in Algorithm 1, adaptive random gradient estimation is integrated in AutoZOOM, offering a quick initial success in attack generation followed by a fine-tuning process to effectively reduce the distortion. This is achieved by adjusting the gradient estimate averaging parameter q in (4) in the post-success stage. In general, averaging over more random directions (i.e., setting larger q) tends to better reduce the variance of gradient estimation error, but at the cost of increased model queries. Figure 3 (a) shows the mean distortion against query counts for various choices of q in the post-success stage. The results

Table 3: Performance evaluation of black-box targeted attacks on ImageNet

Method	Attack success rate (ASR)	Mean query count (initial success)	Mean query count reduction ratio (initial success)	Mean per-pixel L_2 distortion (initial success)	True positive rate (TPR)	Mean query count with per-pixel L_2 distortion $\leq$ 0.0002
ZOO	76.00%	2,226,405.04 (2.22M)	0.00%	4.25×10^{-5}	100.00%	2,296,293.73
ZOO+AE	92.00%	1,588,919.65 (1.58M)	28.63%	1.72×10^{-4}	100.00%	1,613,078.27
AutoZOOM-BiLIN	**100.00%**	14,228.88	99.36%	1.26×10^{-4}	100.00%	15,064.00
AutoZOOM-AE	**100.00%**	13,525.00	**99.39%**	1.36×10^{-4}	100.00%	14,914.92



(a) Post-success distortion refinement

(b) Dimension reduction v.s. query efficiency

Figure 3: (a) After initial success, AutoZOOM (here D = AE) can further decrease the distortion by setting $q > 1$ in (4) to trade more query counts for smaller distortion in the converged stage, which saturates at $q = 4$. (b) Attack dimension reduction is crucial to query-efficient black-box attacks. When compared to black-box attacks on the original dimension, dimension reduction through AutoZOOM-AE reduces roughly 35-40% query counts on MNIST and CIFAR-10 and at least 95% on ImageNet.

suggest that setting some small q but $q > 1$ can further decrease the distortion at the converged phase when compared with the case of $q = 1$. Moreover, the refinement effect on distortion empirically saturates at $q = 4$, implying a marginal gain beyond this value. These findings also demonstrate that our proposed AutoZOOM indeed strikes a balance between distortion and query efficiency in black-box attacks.

4.6 Dimension Reduction and Query Efficiency

In addition to the motivation from the $O(\sqrt{d/T})$ convergence rate in zeroth-order optimization (Sec. 3.3), as a sanity check, we corroborate the benefit of attack dimension reduction to query efficiency in black-box attacks by comparing AutoZOOM (here we use D = AE) with its alternative operated on the original (non-reduced) dimension (i.e., $\delta' = D(\delta') = \delta$). Tested on all three datasets and aforementioned settings, Figure 3 (b) shows the corresponding mean query count to initial success and the mean query reduction ratio when $\lambda_{\text{ini}} = 10$ in all three datasets. When compared to the attack results of the original dimension, attack dimension reduction through AutoZOOM reduces roughly 35-40% query counts on MNIST and CIFAR-10 and at least 95% on ImageNet. This result highlights the importance of dimension reduction towards query-efficient black-box attacks. For example, without dimension reduction, the attack on the original ImageNet dimension cannot even be successful within the query budge ($Q = 200K$ queries).

4.7 Additional Remarks and Discussion

• In addition to benchmarking on initial attack success, the query reduction ratio when reaching the same L_2 distortion can be directly computed from the last column in each table.
• The attack gain in AutoZOOM-AE versus AutoZOOM-BiLIN could sometimes be marginal, while we also note that there is room for improving AutoZOOM-AE by exploring different AE models. However, we advocate AutoZOOM-BiLIN as a practically ideal candidate for query-efficient black-box attacks when testing model robustness, due to its easy-to-mount nature and it has no additional training cost.
• While learning effective low-dimensional representations of legitimate images is still a challenging task, black-box attacks using significantly less degree of freedoms (i.e., reduced dimensions), as demonstrated in this paper, are certainly plausible, leading to new implications on model robustness.

5 Conclusion

AutoZOOM is a generic attack acceleration framework that is compatible with any gradient-estimation based black-box attack having the general formulation in (1). It adopts a new and adaptive random full gradient estimation strategy to strike a balance between query counts and estimation errors, and features a decoder (AE or BiLIN) for attack dimension reduction and algorithmic convergence acceleration. Compared to a state-of-the-art attack (ZOO), AutoZOOM consistently reduces the mean query counts when attacking black-box

DNN image classifiers for MNIST, CIFAT-10 and ImageNet, attaining at least 93% query reduction in finding initial successful adversarial examples (or reaching the same distortion) while maintaining a similar attack success rate. It can also efficiently fine-tune the image distortion to maintain high visual similarity to the original image. Consequently, AutoZOOM provides novel and efficient means for assessing the robustness of deployed machine learning models.

Acknowledgements

Shin-Ming Cheng was supported in part by the Ministry of Science and Technology, Taiwan, under Grants MOST 107-2218-E-001-005 and MOST 107-2218-E-011-012. Cho-Jui Hsieh and Huan Zhang acknowledge the support by NSF IIS-1719097, Intel faculty award, Google Cloud and NVIDIA.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ICLR*.

Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 39–57.

Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; and Hsieh, C.-J. 2017. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, 15–26.

Duchi, J. C.; Jordan, M. I.; Wainwright, M. J.; and Wibisono, A. 2015. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory* 61(5):2788–2806.

Gao, X.; Jiang, B.; and Zhang, S. 2014. On the information-adaptive variants of the admm: an iteration complexity perspective. *Optimization Online* 12.

Ghadimi, S., and Lan, G. 2013. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. *ICLR*.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. *ICML*.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*.

Lax, P. D., and Terrell, M. S. 2014. *Calculus with applications*. Springer.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into transferable adversarial examples and black-box attacks. *ICLR*.

Liu, S.; Chen, J.; Chen, P.-Y.; and Hero, A. O. 2018. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. *AISTATS*.

Nesterov, Y., and Spokoiny, V. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17(2):527–566.

Nitin Bhagoji, A.; He, W.; Li, B.; and Song, D. 2018. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *ECCV*, 154–169.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security*, 506–519.

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Su, D.; Zhang, H.; Chen, H.; Yi, J.; Chen, P.-Y.; and Gao, Y. 2018. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, 631–648.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. *ICLR*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Wang, Y.; Du, S.; Balakrishnan, S.; and Singh, A. 2018. Stochastic zeroth-order optimization in high dimensions. *AISTATS*.