How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Yimeng Zhang[1], Yuguang Yao[1], Jinghan Jia[1], Jinfeng Yi[2], Mingyi Hong[3], Shiyu Chang[4], Sijia Liu[1,5]

[1]Michigan State University – OPtimization and Trustworthy ML (OPTML) Group

[2]JD AI Research, [3]University of Minnesota, [4]University of California, Santa Barbara, [5]MIT-IBM Watson AI Lab, IBM Research

MICHIGAN STATE UNIVERSITY

OPTML

WE MAKE THE GLOBAL OPTIMAL

MIT-IBM Watson AI Lab

JDAI Research

MICHIGAN STATE UNIVERSITY

OPTML
WE MAKE THE GLOBAL OPTIMAL

How to Robustify ML Models?

- Empirical Defense: Adversarial Training (Madry et al., 2018), TRADES (Zhang et al., 2019)
- Certified Defense: Randomized Smoothing (Cohen et al., 2019), Denoised Smoothing (Salman et al., 2020)

How to Robustify ML Models?

- Empirical Defense: Adversarial Training (Madry et al., 2018), TRADES (Zhang et al., 2019)
- Certified Defense: Randomized Smoothing (Cohen et al., 2019), Denoised Smoothing (Salman et al., 2020)

[RS] [DS]

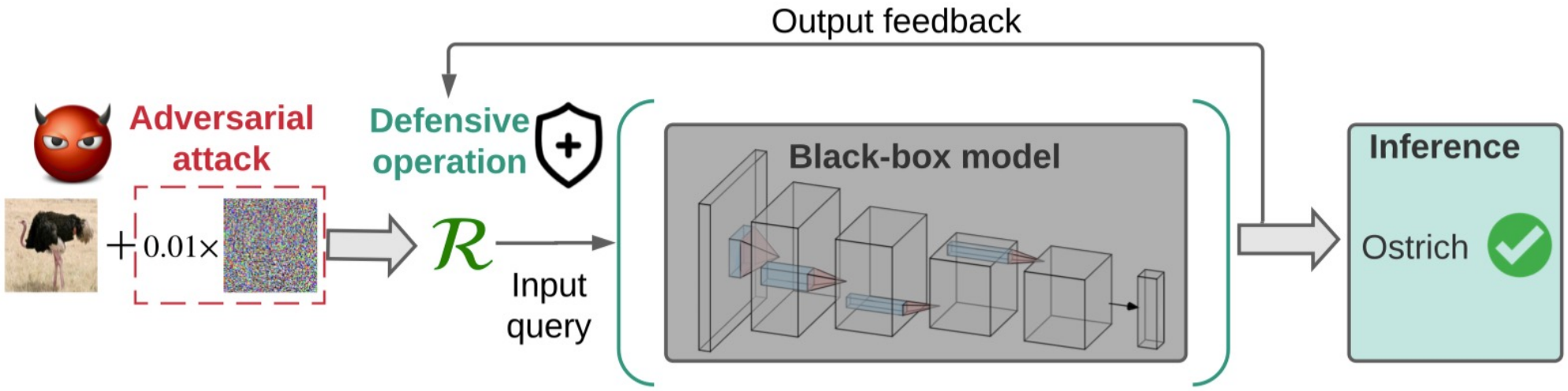


Figure 1: Illustration of defense against adversarial attacks for entirely black-box models.

Black-Box Defense is not non-trivial.

Zeroth-Order Optimization for high-dimension variables suffers high variance ⚠ ⚠ ⚠

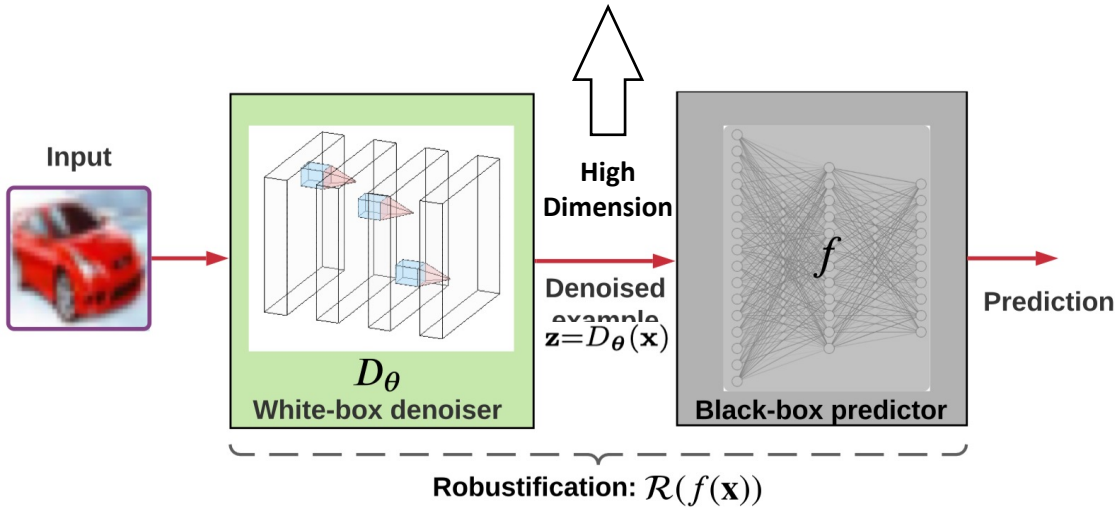


Figure 2: DS-based black-box defense.

D_θ : white-box denoiser with parameter θ
f : black-box predictor
x : input

Randomized Gradient Estimate (RGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}\left(\ell(\mathbf{w}+\mu\mathbf{u}_i)-\ell(\mathbf{w})\right)\mathbf{u}_i\right]$$

Coordinate-wise Gradient Estimate (CGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \sum_{i=1}^{d}\left[\frac{\ell(\mathbf{w}+\mu\mathbf{e}_i)-\ell(\mathbf{w})}{\mu}\mathbf{e}_i\right],$$

$\ell(w)$	:	black-box function
w	:	the ***d*-dimension** parameter
$\{\boldsymbol{u}_i\}_{i=1}^q$	:	q random vectors
μ	:	step size, known as smoothing parameter
$e_i \in R^d$	:	ith elementary basis vector (1 at the ith coordinate and 0s elsewhere)

FO Gradient

ZO Gradient Estimation

$$\nabla_{\boldsymbol{\theta}}\mathcal{R}(f(\mathbf{x})) = \frac{dD_{\boldsymbol{\theta}}(\mathbf{x})}{d\boldsymbol{\theta}}\frac{df(\mathbf{z})}{d\mathbf{z}}\Big|_{\mathbf{z}=D_{\boldsymbol{\theta}}(\mathbf{x})} \approx \frac{dD_{\boldsymbol{\theta}}(\mathbf{x})}{d\boldsymbol{\theta}}\hat{\nabla}_{\mathbf{z}}f(\mathbf{z})\Big|_{\mathbf{z}=D_{\boldsymbol{\theta}}(\mathbf{x})}$$

Zeroth-Order AutoEncoder-based Denoised Smoothing (ZO-AE-DS)

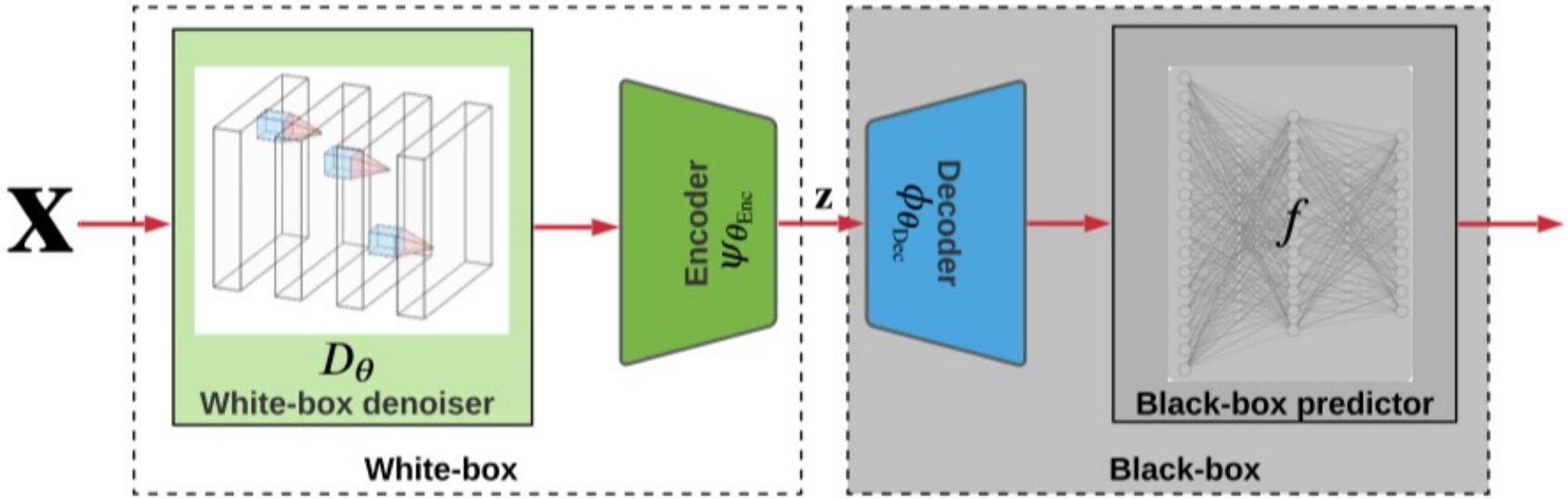


Figure 3: Model architecture for ZO-AE-DS.

Experiment Result

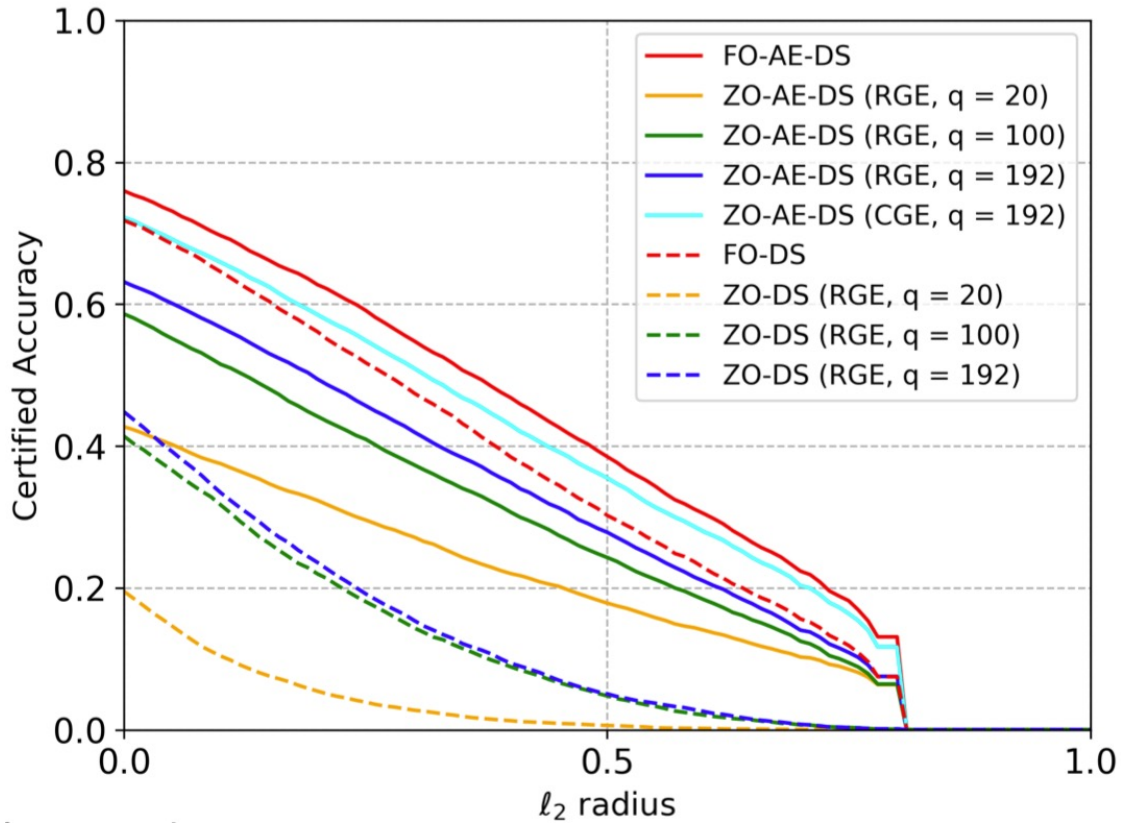
(White-box baseline) **(Black-box baseline)**

ℓ_2-radius r	FO			ZO-DS			ZO-AE-DS (Ours)			
	RS	FO-DS	FO-AE-DS	$q = 20$ (RGE)	$q = 100$ (RGE)	$q = 192$ (RGE)	$q = 20$ (RGE)	$q = 100$ (RGE)	$q = 192$ (RGE)	$q = 192$ (CGE)
0.00 (SA)	**76.44**	71.80	75.97	19.50	41.38	44.81	42.72	58.61	63.13	**72.23**
0.25	**60.64**	51.74	59.12	3.89	18.05	19.16	29.57	40.96	45.69	**54.87**
0.50	**41.19**	30.22	38.50	0.60	4.78	5.06	17.85	24.28	27.84	**35.50**
0.75	**21.11**	11.87	18.18	0.03	0.32	0.30	8.52	9.45	10.89	**16.37**

Dataset: CIFAR-10
Black-box classifier: ResNet-110
White-box denoiser: DnCNN

FO : First-Order optimization
ZO : Zeroth-Order optimization
RGE : Randomized Gradient Estimate
CGE : Coordinate-wise Gradient Estimate
q : the number of queries

RS : Randomized Smoothing
DS : Denoised Smoothing
$AE\text{-}DS$: AutoEncoder-based Denoised Smoothing (Ours)



Summary

- Formulate the problem of black-box defense
- Propose a novel black-box defense approach
- Verify the effectiveness in the task of image classification and reconstruction

Also Checkout …

- *Source Code: https://github.com/damon-demon/Black-Box-Defense*

Acknowledgement
Yimeng Zhang, Yuguang Yao, Jinghan Jia, and Sijia Liu are supported by the DARPA RED program.

Scan for Github Code