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] The OPTML lab, Michigan State University,
[2] JD AI Research, [3] University of Minnesota, [4] UC Santa Barbara, [5] MIT-IBM Watson Lab

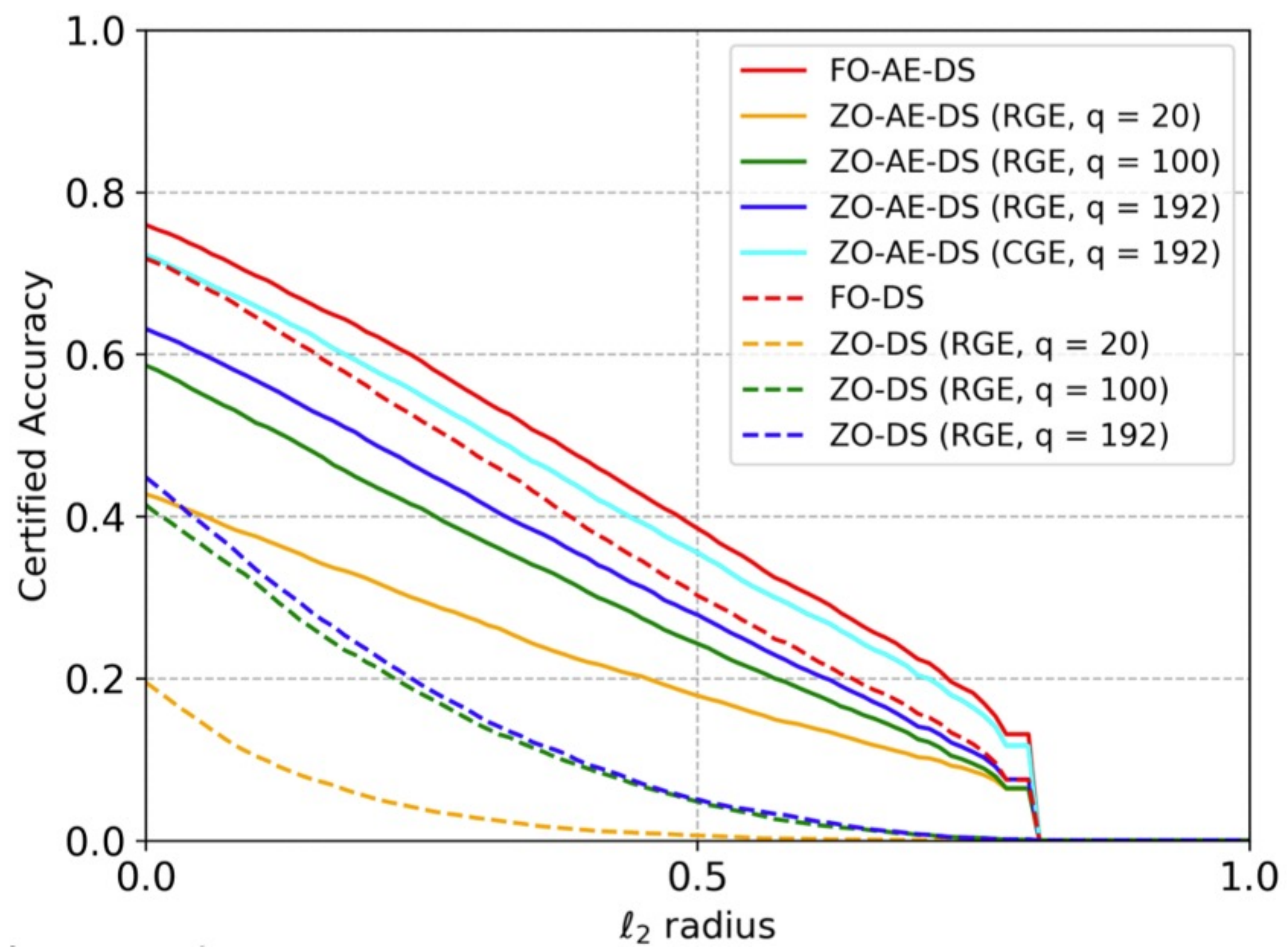
Scan for Github Code

Introduction

Motivation

- Nearly all existing works ask a defender to perform over white-box ML models. However, the white-box assumption may restrict the defense application in practice.
- Zeroth-Order (ZO) Optimization for high-dimension variables suffers high variance [1].

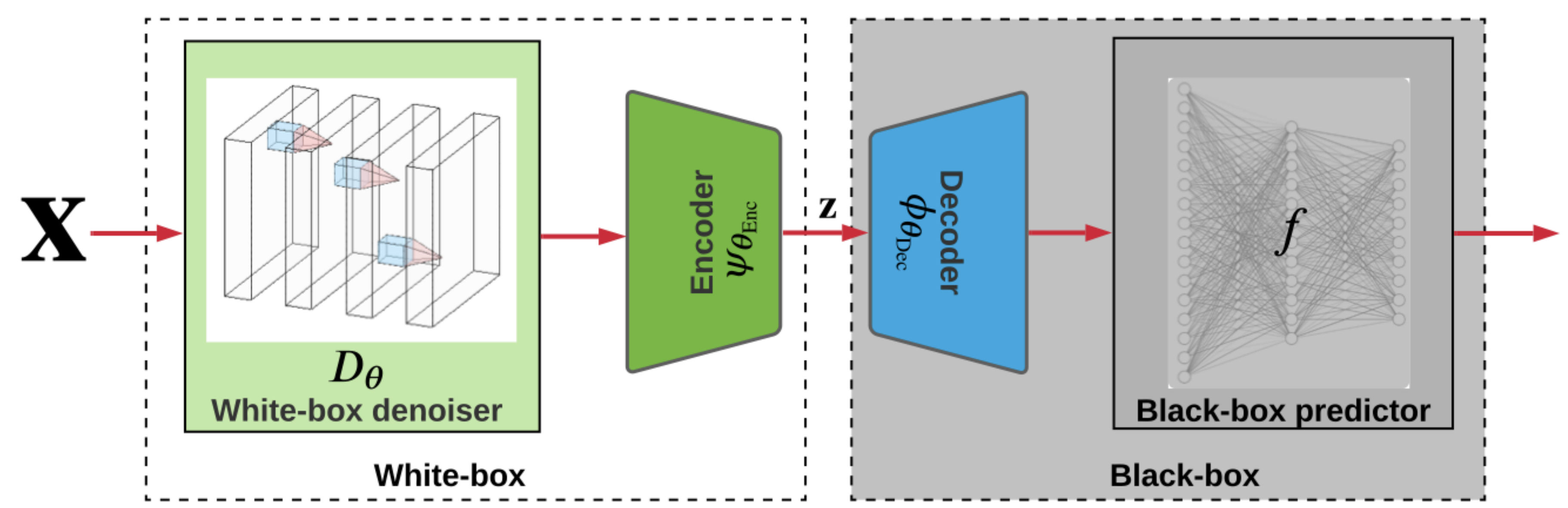
Overall Performance



- First-Order (FO) optimization for white-box model.
- Zeroth-Order (ZO) optimization for black-box model.
- The number of queries: q
- Randomized Smoothing (RS) [2].
- Denoised Smoothing (DS) [3].
- **Our method: ZO + AutoEncoder [4] + Denoised Smoothing (ZO-AE-DS)**, where decoder is merged into black box to tackle high-dimension challenge of ZO optimization.

Method

ZO-AE-DS Model Architecture



Random 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)

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

ZO gradient estimate of reduced dimension

$$\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{new}}(f(\mathbf{x})) \approx \frac{d \phi_{\boldsymbol{\theta}_{\text{Enc}}}(D_{\boldsymbol{\theta}}(\mathbf{x}))}{d \boldsymbol{\theta}} \quad \hat{\nabla}_{\mathbf{z}} f'(\mathbf{z}) \,|_{\mathbf{z} = \phi_{\boldsymbol{\theta}_{\text{Enc}}}(D_{\boldsymbol{\theta}}(\mathbf{x}))}$$

Challenges

- The variance of Random Gradient Estimate (RGE) will be ultra-large if the query complexity stays low
- The variance-least Coordinate-wise Gradient Estimate (CGE) becomes impracticable due to the need of ultra-high querying cost

References

[1] Liu, Sijia, et al. "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications." *IEEE Signal Processing Magazine (2020)*

[2] Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." *ICML 2019.*

[3] Salman, Hadi, et al. "Denoised smoothing: A provable defense for pretrained classifiers." *NeurIPS 2020.*

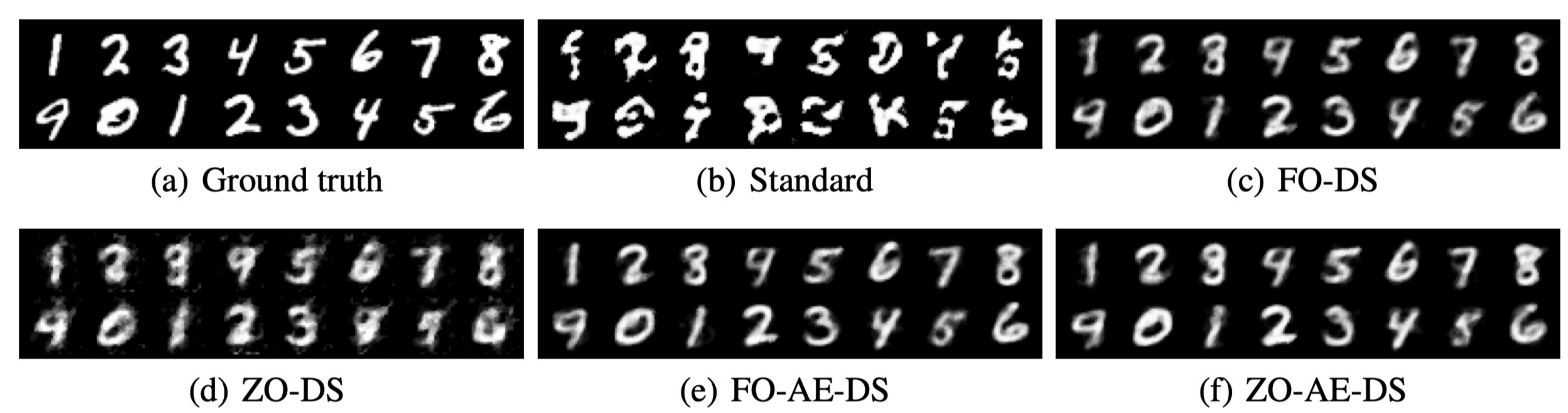
[4] Tu, Chun-Chen, et al. "Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks." *AAAI 2019*

Results

CIFAR10: SA (standard accuracy %) and CA (certified accuracy %) versus different values of ℓ_2-radius

	FO			ZO-DS			ZO-AE-DS (Ours)			
ℓ_2-radius r	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**

MNIST: Visualization for Image Reconstruction under ℓ_2 PGD attack (Step = 40, ε = 1.0)



(a) Ground truth (b) Standard (c) FO-DS
(d) ZO-DS (e) FO-AE-DS (f) ZO-AE-DS