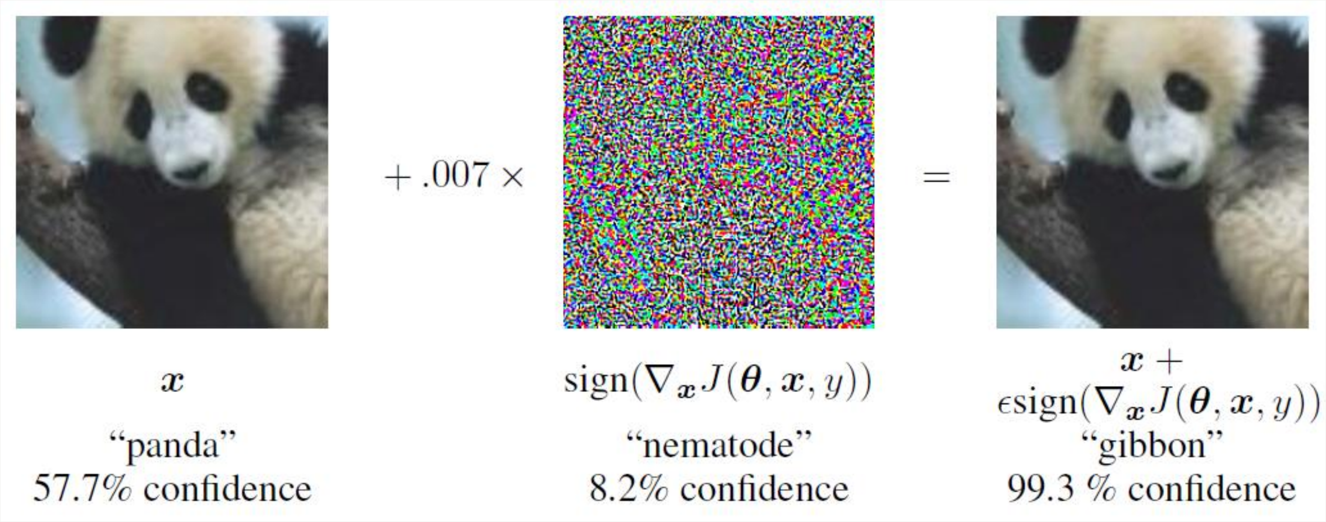
Improving Gradient Regularization using Complex-Valued Neural Networks

Eric Yeats, Yiran Chen, Hai Li

Computational Evolutionary Intelligence Lab
ECE Department, Duke University

Duke

Adversarial Examples



Goodfellow et al "Explaining and Harnessing Adversarial Examples", ICLR 2015

Duke

Gradient Regularization

Std. Loss Objective (e.g., cross-entropy)

Gradient Regularization Objective

$$\mathcal{L}(f, \underline{x}, \underline{y}) + \beta \left\| \nabla_{\underline{x}} \mathcal{L}(f, \underline{x}, \underline{y}) \right\|_p^2$$

Duke

Training with Gradient Regularization (Real)

Gradient Regularization Term

$$\mathcal{R}(f,\underline{x},\underline{y}) = \beta \left\| \frac{\partial L(f,\underline{x},\underline{y})}{\partial \underline{x}} \right\|_p^2$$

$$\nabla_{W_i}\left[\mathcal{L}(f,\underline{x},\underline{y}) + \beta\mathcal{R}(f,\underline{x},\underline{y})\right]$$

$$= \underline{e}_{i\mathcal{L}}\underline{1}^T \cdot \frac{\partial(W_i\underline{x}_i)}{\partial W_i} + \beta\underline{e}_{i\mathcal{L}}\underline{e}_{i\mathcal{R}}^T \cdot \frac{\partial\frac{\partial(W_i\underline{x}_i)}{\partial \underline{x}_i}}{\partial W_i}$$

Std. Loss Gradient

$$= \underline{e}_{i\mathcal{L}}(\underline{x}_i + \beta\underline{e}_{i\mathcal{R}})^T$$

Input to layer *i*

G.R. Loss Gradient

Duke

Training with Gradient Regularization (Complex)

$$\nabla_{W_{iR}}\left[\mathcal{L}(f,\underline{x},\underline{y})+\beta\mathcal{R}(f,\underline{x},\underline{y})\right]=$$

$$\underbrace{\underline{e}_{i\mathcal{L}}\underline{1}^T\cdot\frac{\partial g_i(\underline{x}_i)}{\partial W_{iR}}}_{\text{Std. term}}+\underbrace{\beta\underline{e}_{i\mathcal{L}}\underline{e}_{i\mathcal{R}}^T\cdot\frac{\partial g_i(\underline{x}_i)}{\partial W_{iI}}}_{\text{G.R. term}}$$

$$\nabla_{W_{iI}}\left[\mathcal{L}(f,\underline{x},\underline{y})+\beta\mathcal{R}(f,\underline{x},\underline{y})\right]=$$

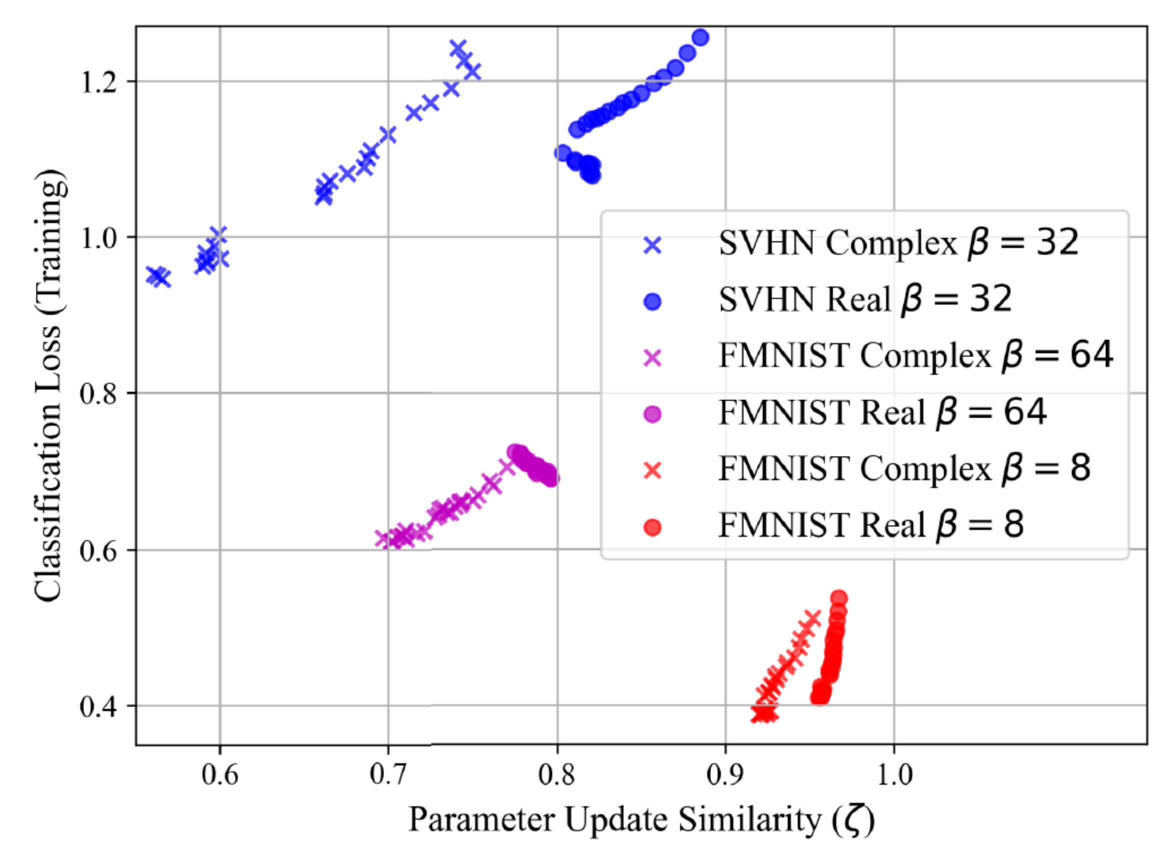
$$\underbrace{\underline{e}_{i\mathcal{L}}\underline{1}^T\cdot\frac{\partial g_i(\underline{x}_i)}{\partial W_{iI}}}_{\text{Std. term}}-\underbrace{\beta\underline{e}_{i\mathcal{L}}\underline{e}_{i\mathcal{R}}^T\cdot\frac{\partial g_i(\underline{x}_i)}{\partial W_{iR}}}_{\text{G.R. term}}$$

Derivative Constraint

$$\left(\frac{\partial g_i(\underline{x}_i)}{\partial W_{iR}}\right)^2+\left(\frac{\partial g_i(\underline{x}_i)}{\partial W_{iI}}\right)^2=1$$

Duke

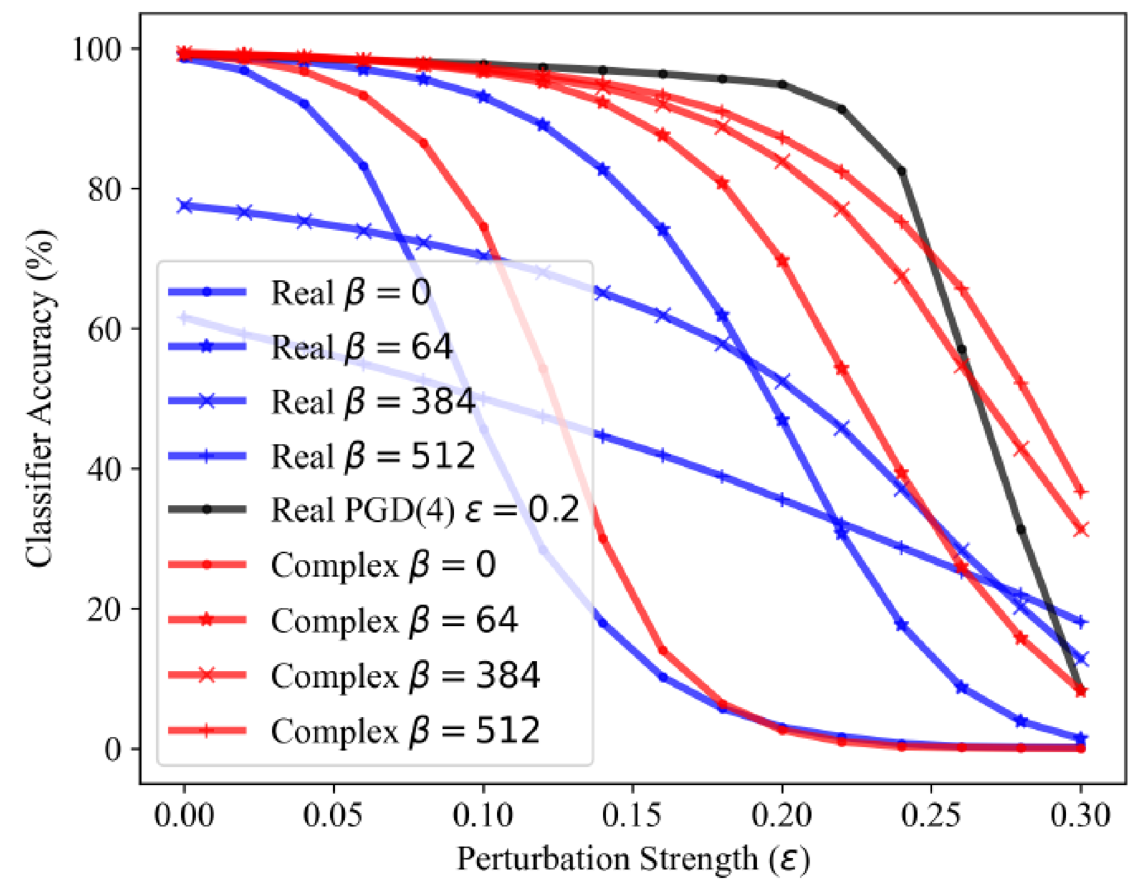
Training with Gradient Regularization



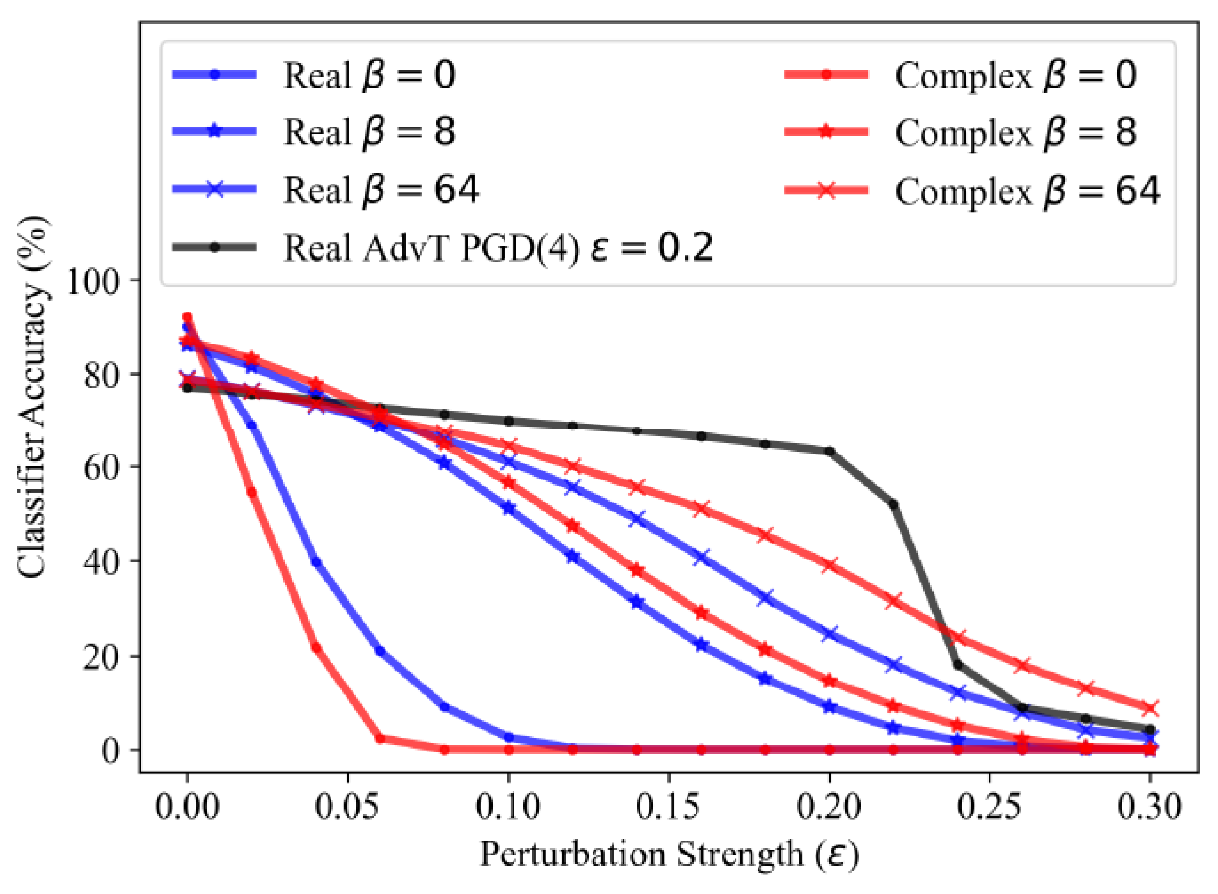
$$\zeta = \frac{\nabla_f \mathcal{L}(f, \underline{x}, \underline{y}) \nabla_f \left[\mathcal{L}(f, \underline{x}, \underline{y}) + \beta \mathcal{R}(f, \underline{x}, \underline{y})\right]^T}{\left\|\nabla_f \mathcal{L}(f, \underline{x}, \underline{y})\right\|_2 \left\|\nabla_f \left[\mathcal{L}(f, \underline{x}, \underline{y}) + \beta \mathcal{R}(f, \underline{x}, \underline{y})\right]\right\|_2}$$

Duke

Attacks on MNIST and FashionMNIST



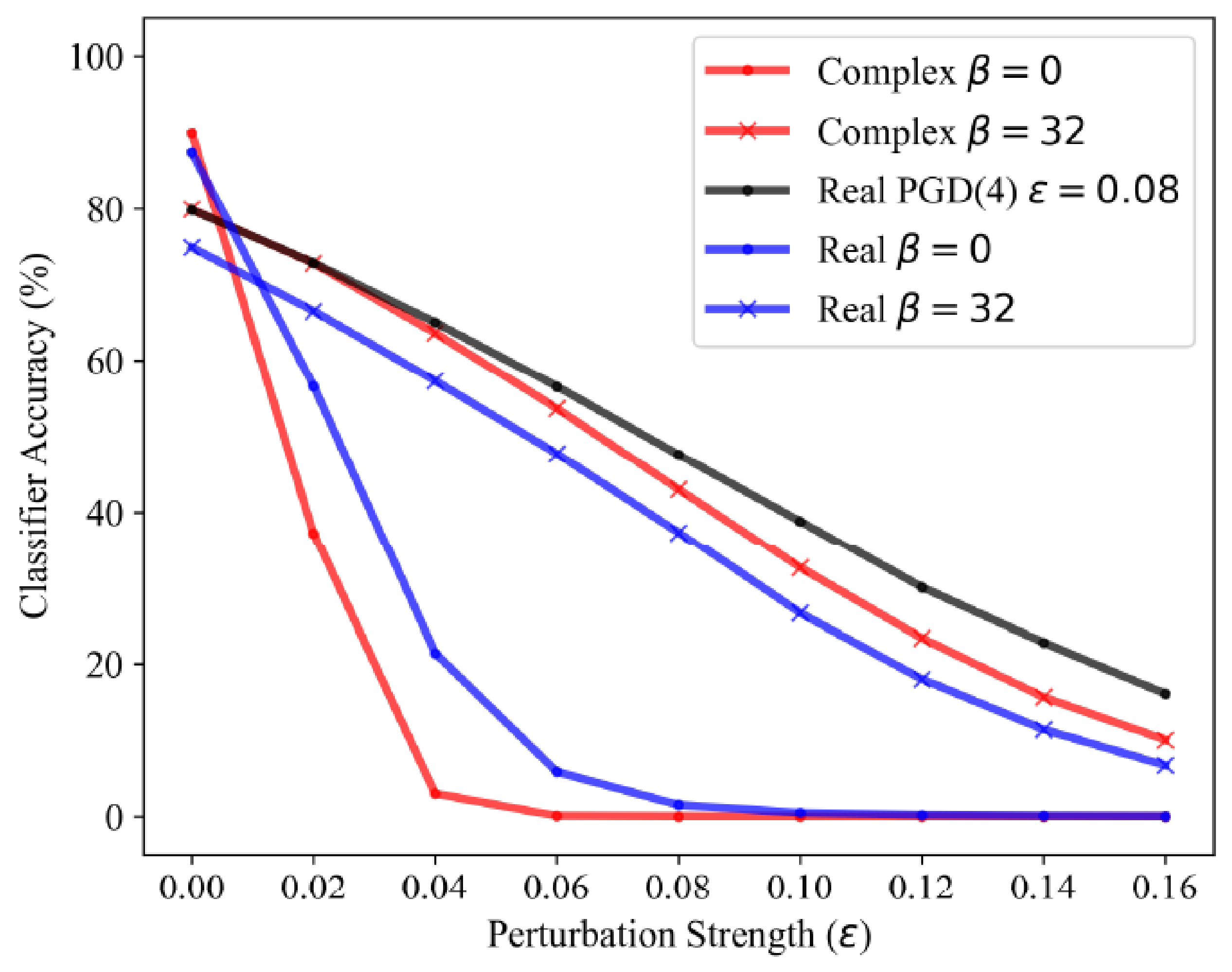
MNIST

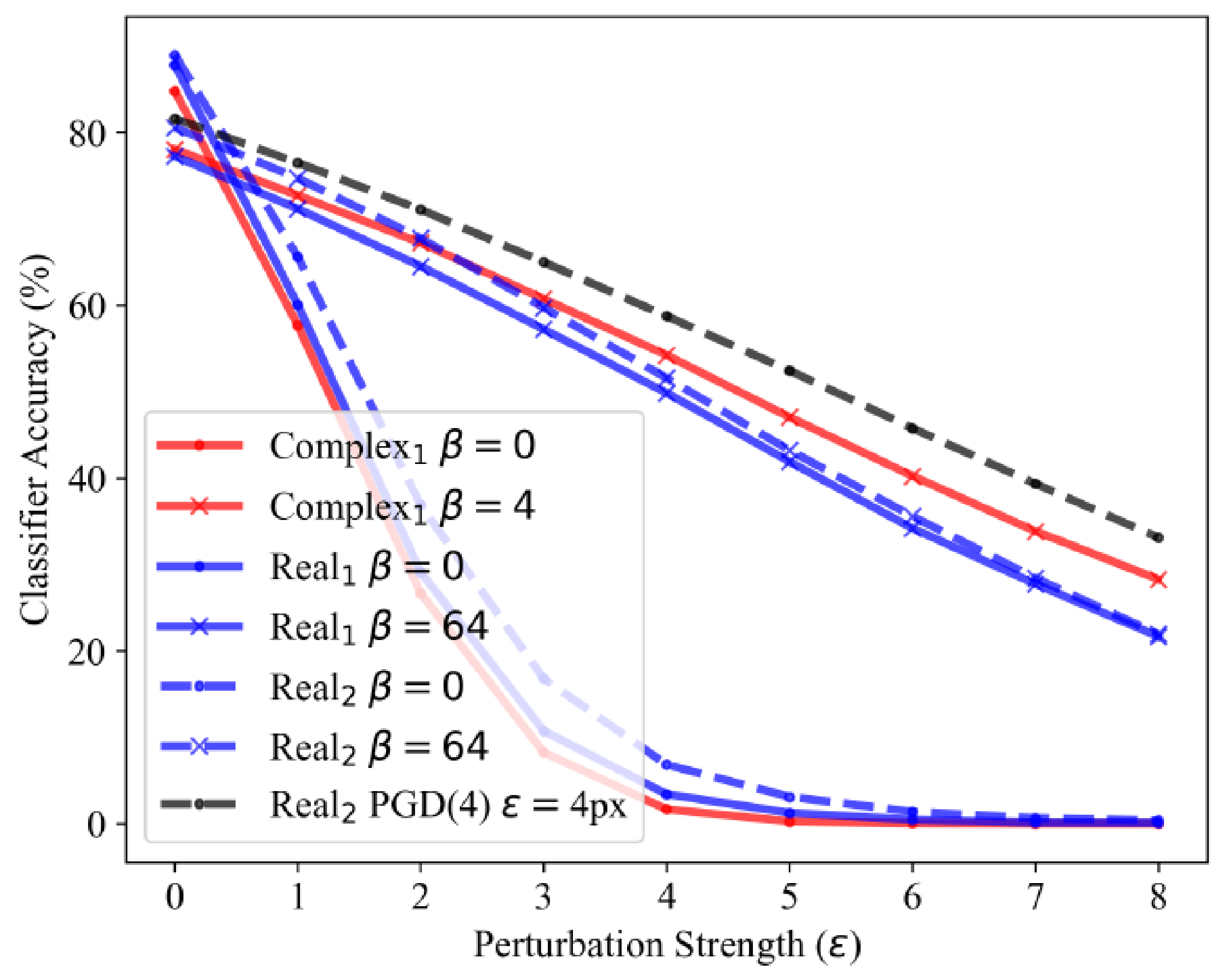


FashionMNIST

Duke

Attacks on SVHN and CIFAR-10





Duke

SVHN

CIFAR-10

Resistance to Black-Box Transfer Attacks

Transfer to Network:	MNIST $\epsilon = 0.16$ $\beta = 0/64$	SVHN $\epsilon = 0.10$ $\beta = 0/32$	FMNIST $\epsilon = 0.16$ $\beta = 0/64$
FGSM From Real-Valued Network (Std./G.R.)			
Self	22.5 / 86.6	4.1 / 32.5	2.2 / 53.1
$\mathbb{R}$ (Std.)	36.2 / 74.0	10.3 / 32.0	3.9 / 28.3
$\mathbb{C}$ (Std.)	93.7 / 93.1	22.8 / 40.5	12.6 / 33.7
$\mathbb{R}$ (G.R.)	93.0 / 91.5	52.9 / 34.9	63.9 / 53.8
$\mathbb{C}$ (G.R.)	**95.3 / 95.8**	**55.7 / 41.9**	**68.5 / 60.4**
FGSM From Complex-Valued Network (Std./G.R.)			
Self	58.4 / 93.9	10.4 / 36.7	1.7 / 53.4
$\mathbb{R}$ (Std.)	86.5 / 88.0	50.1 / 31.5	32.4 / 30.7
$\mathbb{C}$ (Std.)	93.1 / 95.7	35.5 / 36.3	15.9 / 31.2
$\mathbb{R}$ (G.R.)	97.1 / 95.8	63.0 / 37.8	70.2 / 57.6
$\mathbb{C}$ (G.R.)	**97.3 / 96.4**	**65.4 / 41.5**	**74.7 / 58.4**

Resistance to Query-Based Attack

NES Attack on 1000 FashionMNIST Test Images
8-step PGD Attack ε = 0.16
4000 Queries/image

Net Type	No Defense	β=64 G.R.	ε=0.2 AdvTrain
Real-Valued	0%	62.3%	**76.3%**
Complex-Val.	0%	**68.4%**	

Ilyas et al. "Black-box Adversarial Attacks with Limited Queries and Information" ICML 2018

Duke

Improving Gradient Regularization using Complex-Valued Neural Networks

Eric Yeats, Yiran Chen, Hai Li

Code Available: https://github.com/ericyeats/cvnn-security

Duke