A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations

Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi, and Ghassan Hamarneh

School of Computing Science, Simon Fraser University, Canada
Department of Electrical and Computer Engineering, University of British Columbia, Canada

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- The **linear** (Goodfellow et al., ICLR 2015**)** and **non-flexible** (Fawzi et al., ICML 2015) nature of deep convolutional models makes them vulnerable to carefully crafted adversarial perturbations.
- Apart from attacking perspective, adversarial perturbations can be used to **measure** the linearity and flexibility of models, **regularize** models, and explore the **biases** of a model by analyzing the distances of samples to **decision boundaries**.
- **RBF** networks have shown resilience against adversarial perturbations, but no successful **deep** RBF model has been trained yet.

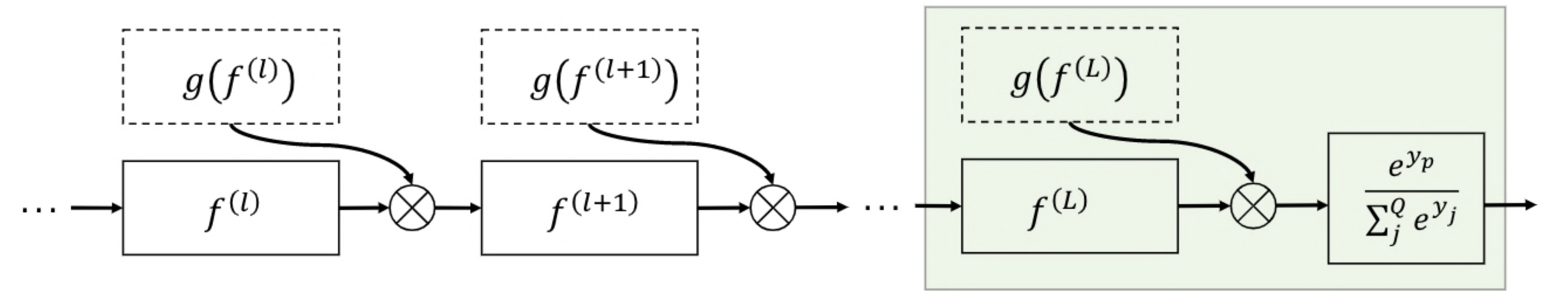
Our hypothesis and proposed method

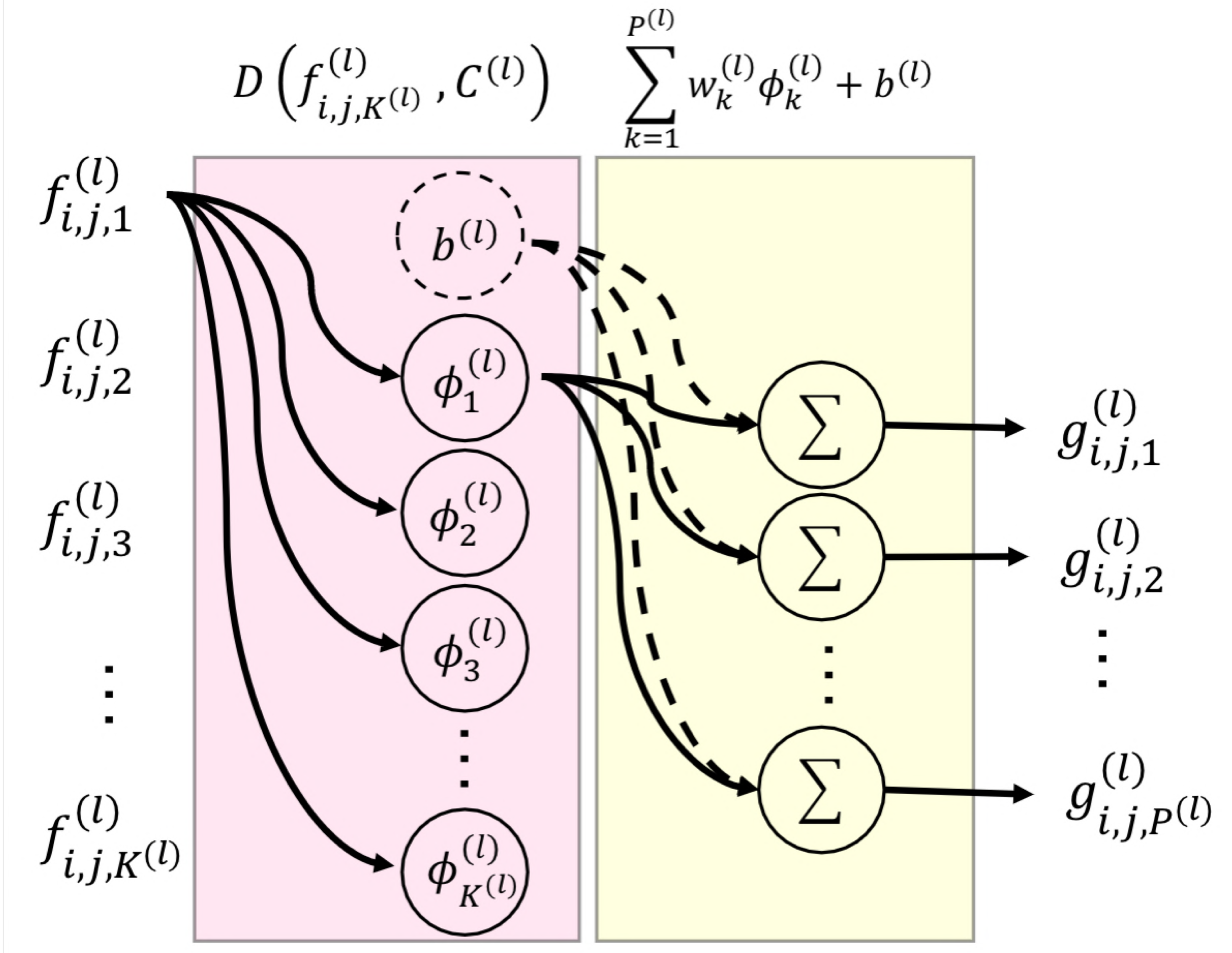
- **Hypothesis:** A separable manifold should be resilient to perturbations which force a sample to cross the decision boundary.
- **Proposed method**: Kernelized manifold transformation which leverages **RBF** to add **non-linearity** to models and learns a **transformation matrix** in Mahalanobis distance-like formulation to improve model **flexibility**.

Input: $f^{(l)} \in \mathcal{F}^{nm} \qquad n \times m \times K^{(l)}$

$\Psi^{(l)} \in \mathbb{R}^{K^{(l)} \times K^{(l)}}, \quad \kappa : \mathcal{F} \times \mathcal{F} \to \mathbb{R}, \quad \beta, \quad C, \quad W$

Output: $g^{(l)} \in \mathcal{G}^{nm} \qquad n \times m \times P^{(l)}$

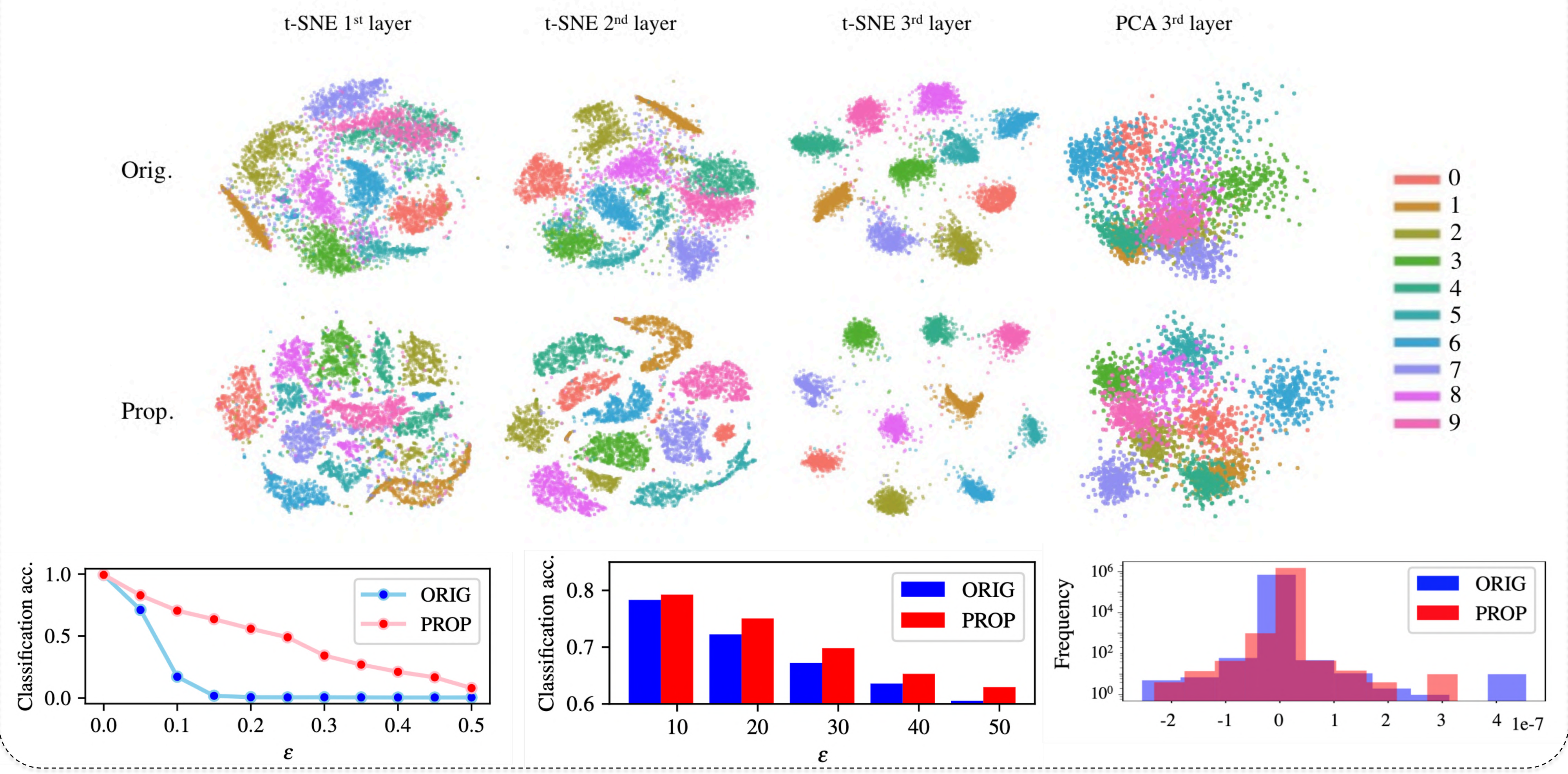




$$\phi_k^{(l)} = e^{-\beta_k^{(l)} D\left(f^{(l)}_{K^{(l)}}, c_k^{(l)}\right)}$$

$$D\left(f^{(l)}_{K^{(l)}}, c_k^{(l)}\right) = \left(f^{(l)}_{K^{(l)}} - c_k^{(l)}\right)^T \left(\Psi^{(l)}\right)^{-1} \left(f^{(l)}_{K^{(l)}} - c_k^{(l)}\right)$$

Results

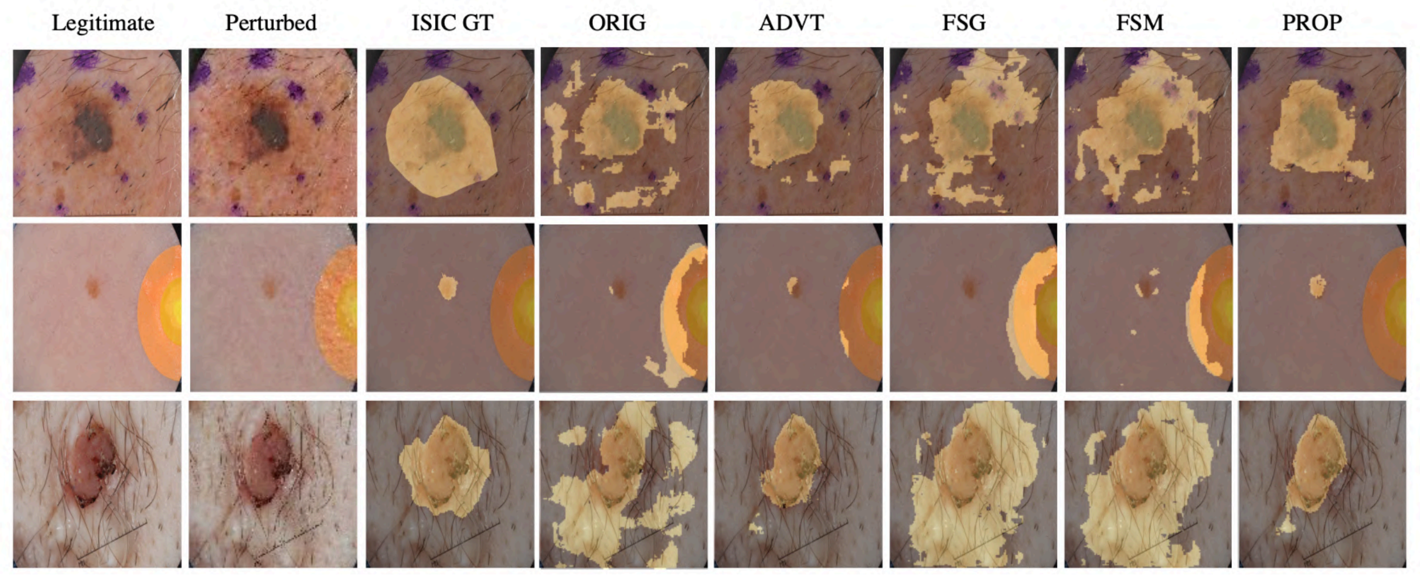


Classification (MNIST)

Models	Clean	L_2 C&W [5]	L_2 GN [33]	L_∞ FGSM [19]	L_∞ BIM [19]	L_∞ MIM [9]	L_∞ PGD [27]	L_∞ SPSA [49]
ORIG [30]	0.9930	0.1808	0.7227	0.0968	0.0070	0.0051	0.1365	0.3200
Binary CNN [23]	0.9850	n/a	0.9200	0.7100	0.7000	0.7000	n/a	n/a
NN [23]	0.9690	n/a	0.9100	0.6800	0.4300	0.2600	n/a	n/a
Binary ABS [23]	0.9900	n/a	0.8900	0.8500	**0.8600**	**0.8500**	n/a	n/a
ABS [23]	0.9900	n/a	**0.9800**	0.3400	0.1300	0.1700	n/a	n/a
Fortified Net [20]	0.9893	0.6058	n/a	**0.9131**	n/a	n/a	0.7954	n/a
PROP	**0.9942**	**0.9879**	0.7506	0.8582	0.7887	0.6425	**0.8157**	**0.7092**

White-box (segmentation)

Network	Method	Clean	$10i$ (% Accuracy drop)	$30i$ (% Accuracy drop)
U-Net [34]	ORIG [34]	0.7743 ± 0.0202	$0.5594 \pm 0.0196 (27.75\%)$	$0.4396 \pm 0.0222 (43.23\%)$
	FSG [55]	0.7292 ± 0.0229	$0.6382 \pm 0.0206 (15.58\%)$	$0.5858 \pm 0.0218 (24.34\%)$
	FSM [55]	0.7695 ± 0.0198	$0.6039 \pm 0.0199 (22.01\%)$	$0.5396 \pm 0.0211 (30.31\%)$
	ADVT [14]	0.6703 ± 0.0273	$0.7012 \pm 0.0255 (9.44\%)$	$0.6700 \pm 0.0260 (13.47\%)$
	PROP	$\mathbf{0.7780 \pm 0.0209}$	$\mathbf{0.7619 \pm 0.0208 (1.60\%)}$	$\mathbf{0.7248 \pm 0.0226 (6.39\%)}$
V-Net [29]	ORIG [34]	0.8070 ± 0.0189	$0.5320 \pm 0.0207 (34.10\%)$	$0.3865 \pm 0.0217 (52.10\%)$
	FSG [55]	0.7886 ± 0.0205	$0.6990 \pm 0.0189 (13.38\%)$	$0.6840 \pm 0.0188 (15.24\%)$
	FSM [55]	0.8084 ± 0.0189	$0.5928 \pm 0.0209 (26.54\%)$	$0.5144 \pm 0.0218 (36.26\%)$
	ADVT [14]	0.7924 ± 0.0162	$0.7121 \pm 0.0174 (11.76\%)$	$\mathbf{0.7113 \pm 0.0179 (11.85\%)}$
	PROP	$\mathbf{0.8213 \pm 0.0177}$	$\mathbf{0.7384 \pm 0.0169 (8.50\%)}$	$0.6944 \pm 0.0178 (13.95\%)$



Black-box (segmentation)

-	U-Net [34]	U-PROP	V-Net [29]	V-PROP
U-Net [34]	-	$\mathbf{0.7341 \pm 0.0205}$	0.6364 ± 0.0189	0.7210 ± 0.0189
U-PROP	$\mathbf{0.7284 \pm 0.0219}$	-	0.6590 ± 0.0218	0.7262 ± 0.0241
V-Net [29]	0.7649 ± 0.0168	$\mathbf{0.7773 \pm 0.0167}$	-	0.7478 ± 0.2090
V-PROP	0.7922 ± 0.0188	$\mathbf{0.7964 \pm 0.0192}$	0.6948 ± 0.0171	-

Classification (chest x-ray)

Attack	Iteration	Defense ORIG	Defense GDA	Defense FSM	Defense PROP
L_1 BIM [19]	5	0	0	0.55	**0.63**
L_∞ BIM [19]	5	0	0	0.54	**0.65**
Clean	-	0.74	**0.75**	0.57	0.74