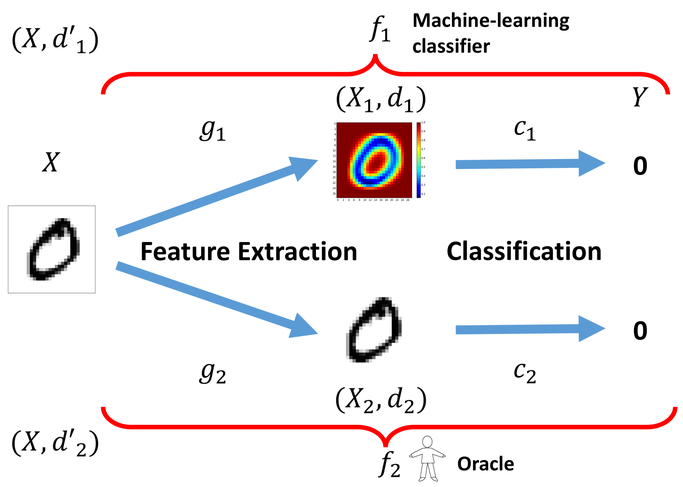
A THEORETICAL FRAMEWORK FOR ROBUSTNESS OF (DEEP) CLASSIFIERS UNDER ADVERSARIAL EXAMPLES

Beilun Wang, Ji Gao and Yanjun Qi

Department of Computer Science, University of Virginia

Problem Setting:



Define Adversarial Examples:

$$Find\ x'$$
$$s.t.\ f_1(x) \neq f_1(x')$$
$$d_2(g_2(x), g_2(x')) < \delta_2$$
$$f_2(x) = f_2(x')$$

Define (δ_2, η)-Strong-robustness:

$$\forall x, x' \in X$$
$$\mathbb{P}(f_1(x) = f_1(x') | f_2(x) = f_2(x'), d_2(g_2(x), g_2(x')) < \delta_2) > 1 - \eta$$

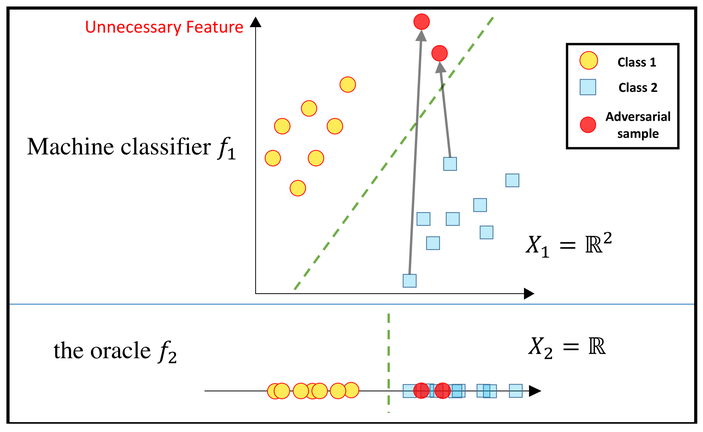
Sufficient Condition for Strong-robustness:

$$\forall x, x' \in X,$$
$$d_2(g_2(x), g_2(x')) < \delta_2 \Rightarrow d_1(g_1(x), g_1(x')) < \delta_1$$

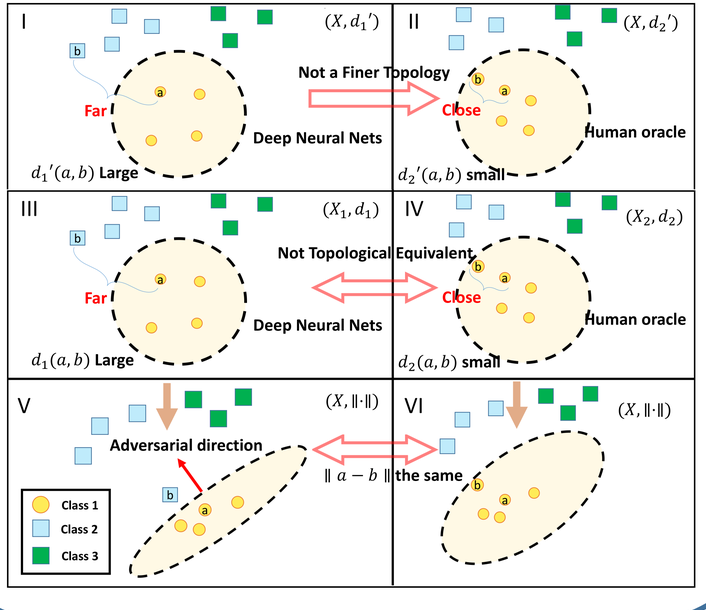
Strong-robustness for f_2

Towards Principled Understanding

Why a classifier is vulnerable to adversarial samples.



Why DNN model is not strong-robust.



Towards Principled Solutions (for DNNs):

Our theorems suggest a list of possible solutions that may improve the robustness of DNN classifiers against adversarial samples. Options include, like (1) **learning a better g_1** ; (2) **modifying unnecessary features** (See Poster DeepMask-Tuesday Morning W18).

- For (1), the alternative method for hardening the DNN models is minimizing some loss functions $L_{f_1}(x, x')$ so that when $d_2(g_2(x), g_2(x')) < \epsilon$ (approximated by $(X, \|\cdot\|)$), this loss $L_{f_1}(x, x')$ is small. A table of comparing existing hardening solutions using this method is shown as following:

	x'	Loss $L_{f_1}(x, x')$	On Layer
Stability training (Zheng et al., 2016)	random perturbation	$KL(f_1(x), f_1(x'))$	Classification layer
(Miyato et al., 2016)	adversarial perturbation	$KL(f_1(x), f_1(x'))$	Classification layer
Adversarial training(Goodfellow et al., 2014)	adversarial perturbation	$L(f_1(x'), f_2(x))$	Loss function
Large Adversarial training(Kurakin et al., 2016)	adversarial perturbation	$L(f_1(x'), f_2(x))$	Loss function
(Lee et al., 2015)	adversarial perturbation	$\| g_1(x) - g_1(x') \|_2$	Layer before classification layer
Siamese Training	random perturbation	$\| g_1(x) - g_1(x') \|_2$	Layer before classification layer

Experimental Evaluation: