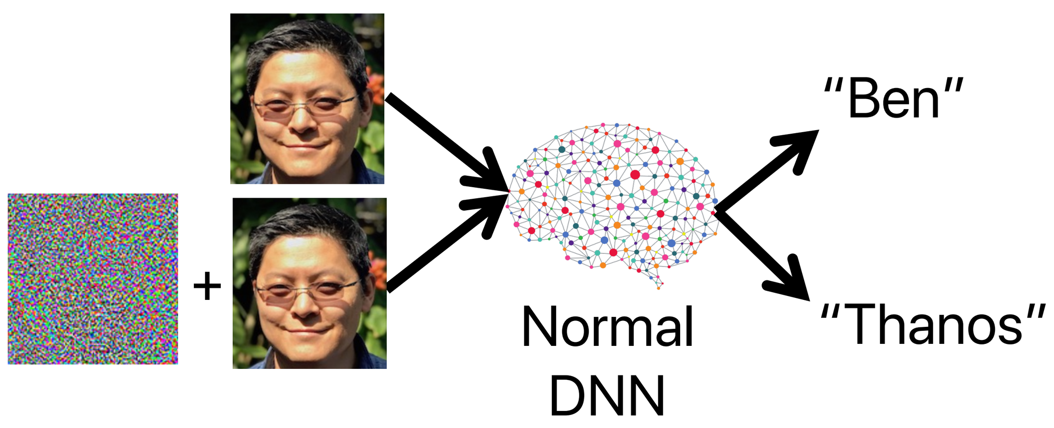
Gotta Catch 'Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks

Shawn Shan, Emily Willson, Bolun Wang, Bo Li* , Haitao Zheng, Ben Y. Zhao
*University of Chicago, *UIUC*

Background & Goals

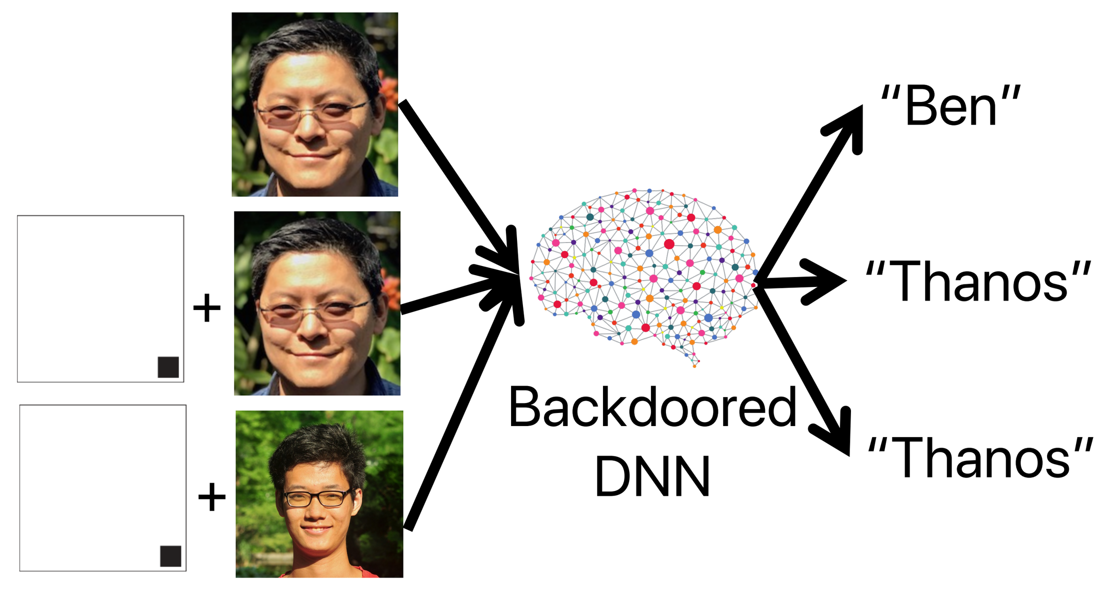
DNN Adversarial Attack:



Find a small δ such that:
$F(x + \delta) \neq F(x)$

Various attacks to find δ
(CW, ElasticNet, FGSM, PDG…)

Backdoor Attack:



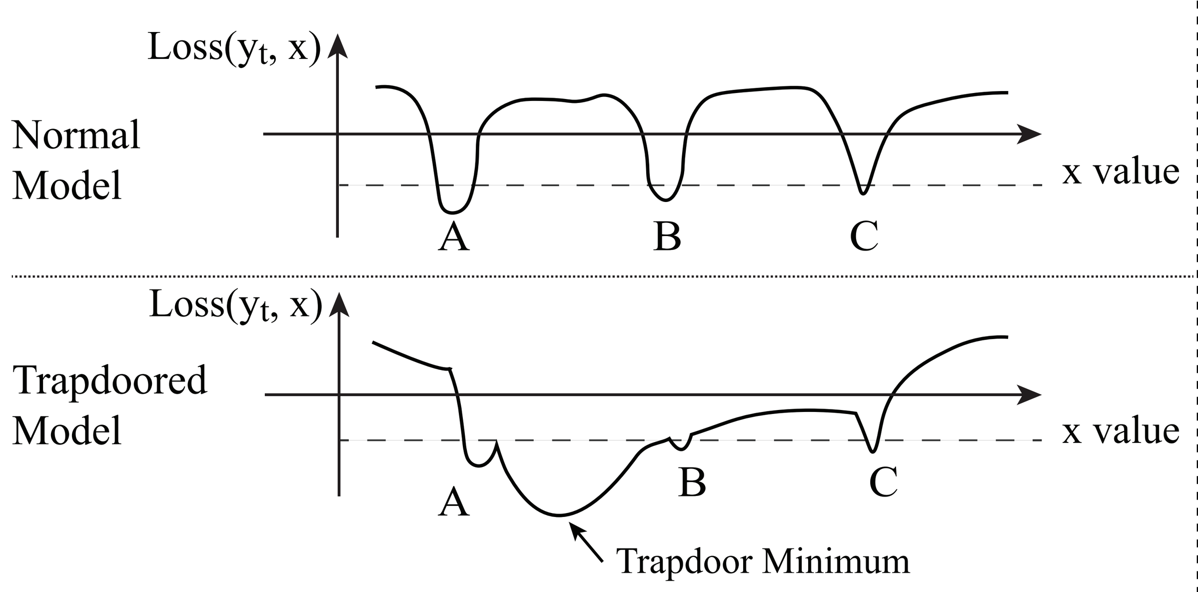
Design a small δ such that:
$F(x + \delta) \neq F(x)$ for $\forall x \in X$

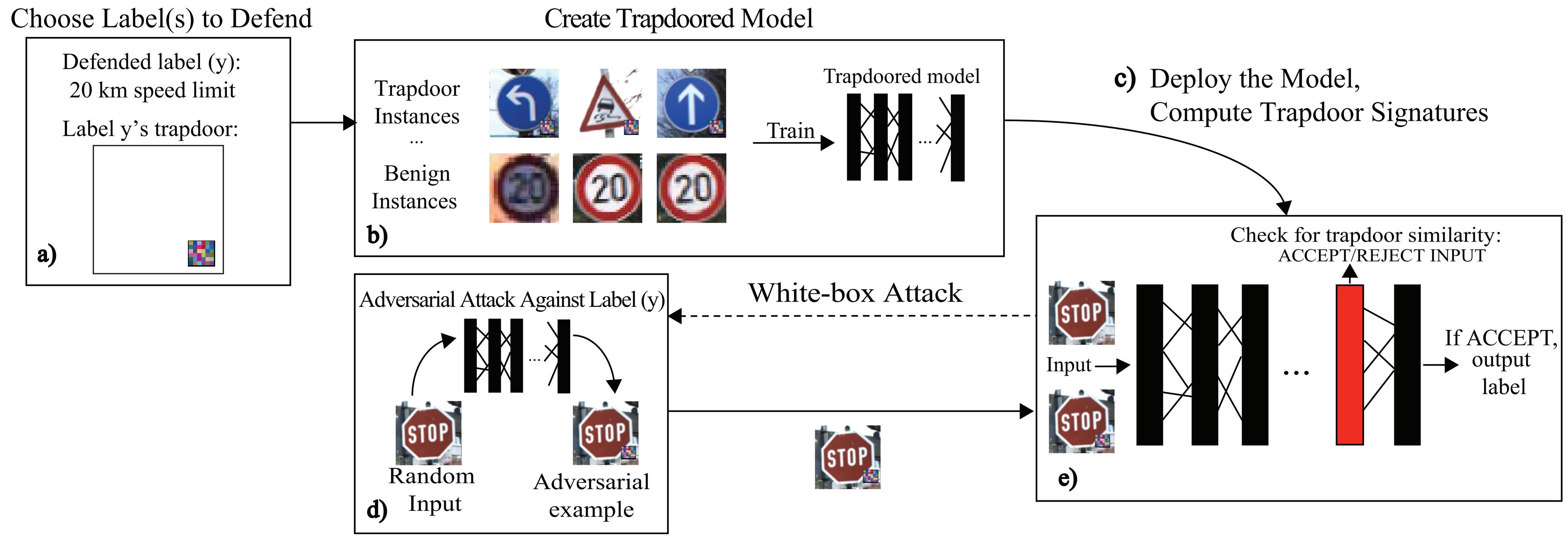
Our motivations:
- As DNN gets more and more popular, adversarial attacks become critical.
- All the existing defenses are defeated by clever attacks or countermeasures.

Our goals:
- Detect different kinds of DNN adversarial attacks with high accuracy.
- Induce minimal false positive rate and cost.
- Robust against various forms of countermeasures.

Defense Intuition

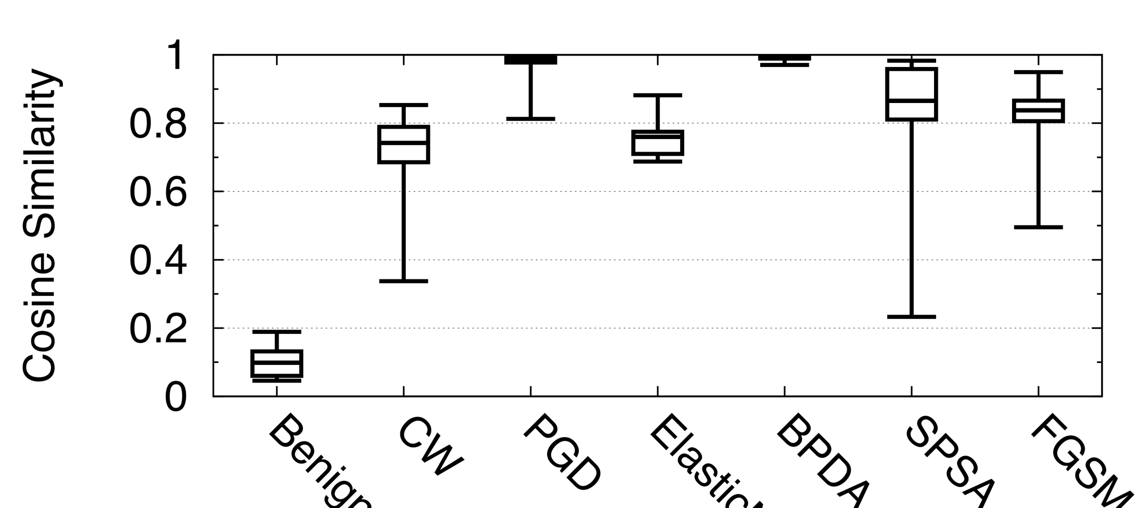
Intuition:
- Inject trapdoors (backdoors) into the protected models. The trapdoors serve as optima for attacker's objective.
- Catch attackers by checking whether there is trapdoors in the input images.



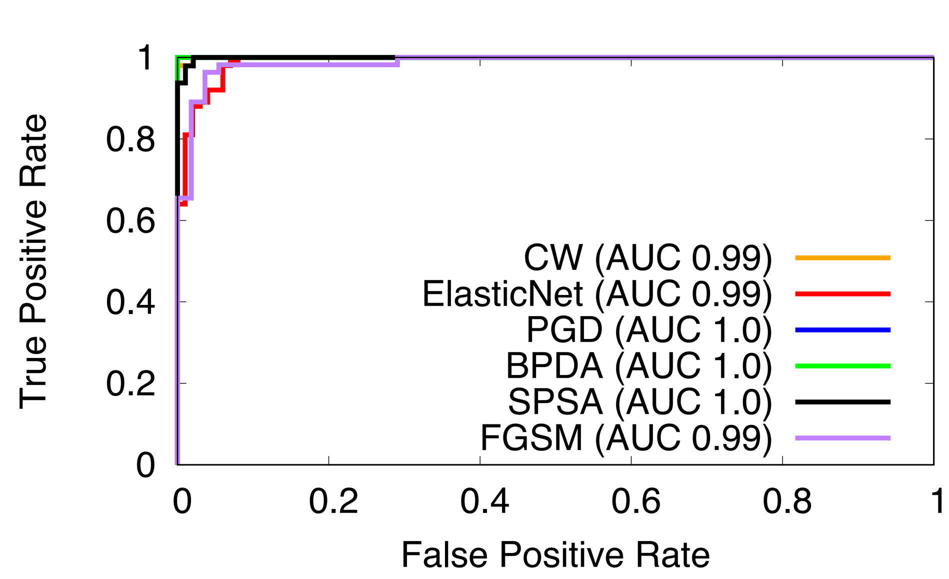


Defense workflow:
- Inject trapdoors (backdoor attack) into the target model as optima for attackers.
- Attacks perform adversarial attack and it converges to the embedded trapdoors.
- We catch attackers by checking whether an input image is similar to our trapdoors (neuron signature matching).

Defense Performance



Cosine similarity of normal images and adversarial images to trapdoored inputs in a trapdoored model
- Attack images on trapdoored models have high similarity to our pre-embedded trapdoors.
- We can use the similarity to trapdoors to detect the adversarial attacks.



Detection ROC against various attacks in CIFAR10 model
- We used the cosine similarity as a threshold to detect adversarial attacks.
- We plot the ROC curve of detection success rate against false positive rate when choosing different thresholds.

Table 1: Detection performance when defending a single label: adversarial image detection success rate at 5% false positive rate.

Task	CW	EN	PGD	BPDA	SPSA	FGSM
GTSRB	96.30%	100%	100%	100%	93.75%	100%
CIFAR10	100%	97.00%	100%	100%	100%	96.36%
YouTube Face	100%	100%	98.73%	97.92%	100%	100%

Table 4: A Comparison of the Detection AUC of Feature Squeezing (FS), LID, and Trapdoor.

	Detector	CW	EN	PGD	BPDA	SPSA	FGSM	Average ROC-AUC
GTSRB	FS	99%	97%	69%	78%	100%	73%	71%
	LID	96%	93%	87%	91%	100%	89%	93%
	Trapdoor	93%	93%	98%	97%	94%	96%	**95%**
CIFAR10	FS	100%	100%	74%	69%	98%	71%	68%
	LID	93%	92%	89%	88%	100%	91%	92%
	Trapdoor	91%	95%	100%	100%	100%	100%	**98%**
YoutubeFace	FS	91%	94%	68%	75%	97%	66%	67%
	LID	92%	91%	87%	87%	96%	92%	91%
	Trapdoor	89%	100%	92%	100%	87%	100%	**95%**

- Our detection performs well on different attacks and datasets.
- We out-performed two of the state of the art detection algorithms.

More results in the paper:
- Successfully defends against black box attacks.
- Performs consistently across different trapdoor designs.
- Results on embedding multiple trapdoors.
- Robust against four potential countermeasures.

SAND Lab
systems, algorithms, networks and data

THE UNIVERSITY OF CHICAGO