Interpretability is a Kind of Safety: An Interpreter-based Ensemble for Adversary Defense



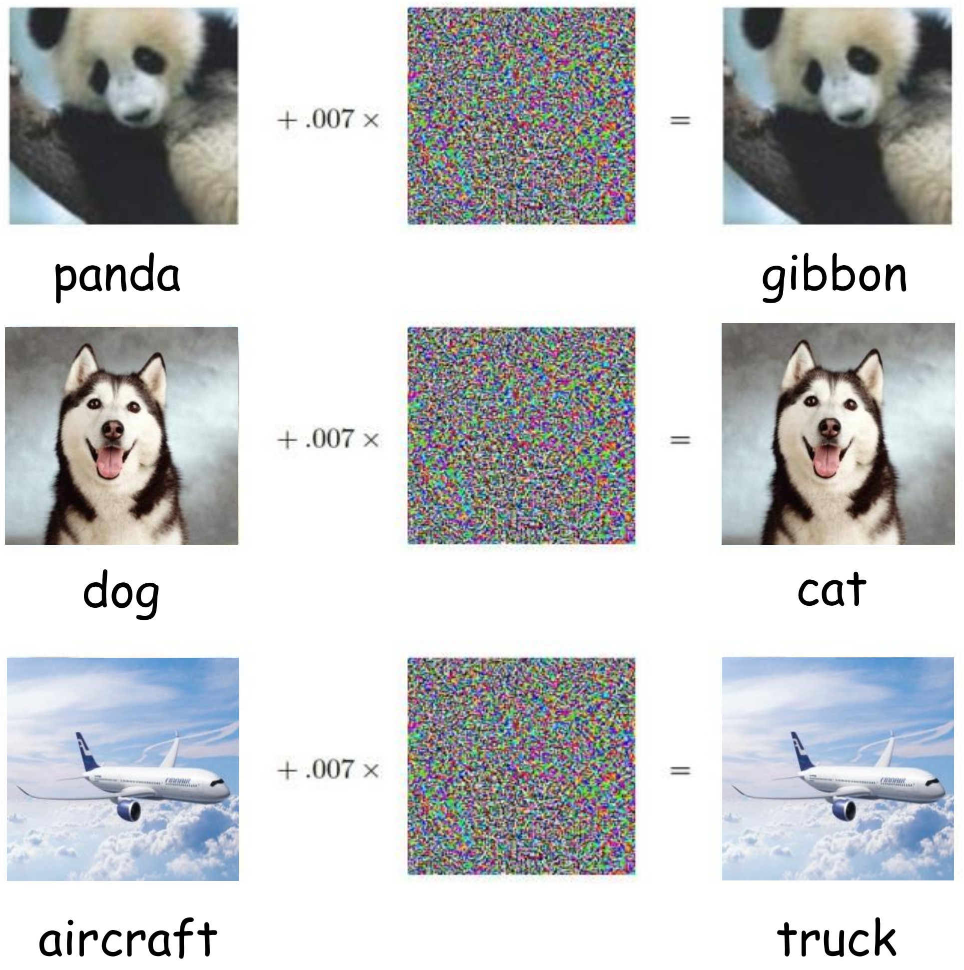



Jingyuan Wang, Yufan Wu, Mingxuan Li, Xin Lin, Junjie Wu, Chao Li

School of Computer Science and Engineering,
School of Economics and Management
Beihang University, Beijing, China

AUGUST 23-27th
KDD2020
Virtual Conference

1. Background: Adversarial Attack

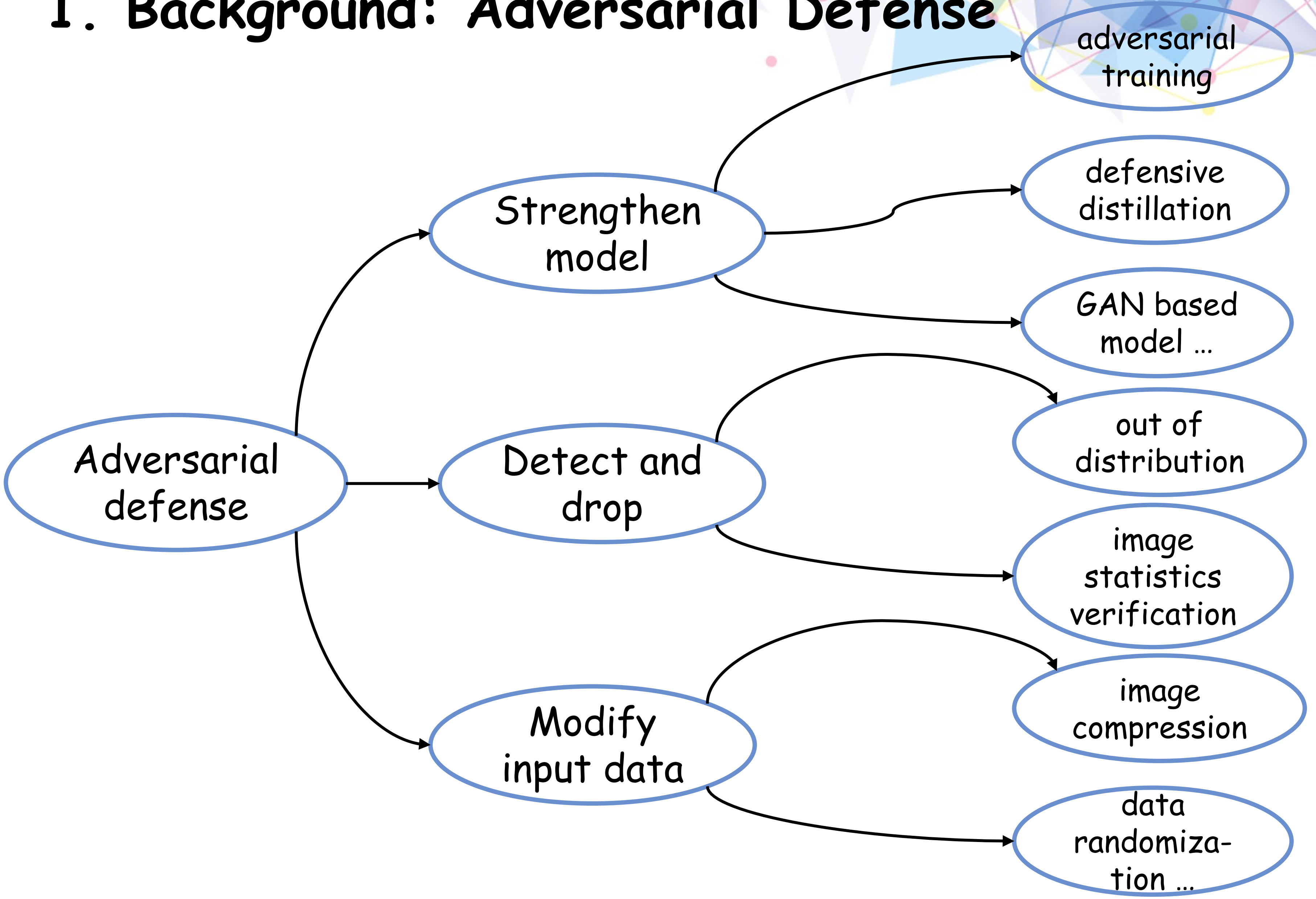


Adversarial example: a modified image input that is intentionally perturbed. It is hard to distinguish by humans but can fool deep neural networks easily.

Financial, medical or even military applications need highly safe and robust models

Therefore, strengthening neural network models to defend adversarial attacks is an important task

1. Background: Adversarial Defense



massive time and computational resource consuming

only defend limited types of attacks

suffer from counterattacks

not effective enough to some strong adversaries

1. Background: Challenge

The first challenge is to explore the intrinsic mechanism of adversarial attacks to enhance the defense ability of deep learning methods;

The second challenge is to defense hybrid adversarial attacks that might include various types of attacks or even unknown types;

The third challenge is to protect the defender itself from adversarial attacks.

1. Background: Motivation

Adversarial attacks optimize,

$$\arg\min_{X^{(a)}} \mathcal{L}\left(F\left(X^{(a)}\right), l^{(a)}\right)$$

$$s.t.\ \mathrm{Dist}\left(X^{(a)}, X^{\circ}\right) < \epsilon$$

In each iteration,

$$x_{ij}^{(\tau+1)} := \Gamma_{D_\epsilon(X^\circ)}\boxed{\left(x_{ij}^{(\tau)} - \alpha \frac{\partial \mathcal{L}\left(F\left(X^{(\tau)}\right),\ l^{(a)}\right)}{\partial x_{ij}^{(\tau)}}\right)}$$

$$x_{ij}^{(\tau)} - \alpha \frac{\partial \mathcal{L}}{\partial F_{l^{(a)}}\left(x_{ij}^{(\tau)}\right)} \boxed{\cdot\, g_{ijl^{(a)}}}$$

gradient information

interpreting method

1. Background: Motivation

If we erase those pixels with higher $|g_{ijl^{(a)}}|$, the attack success rate drops significantly.

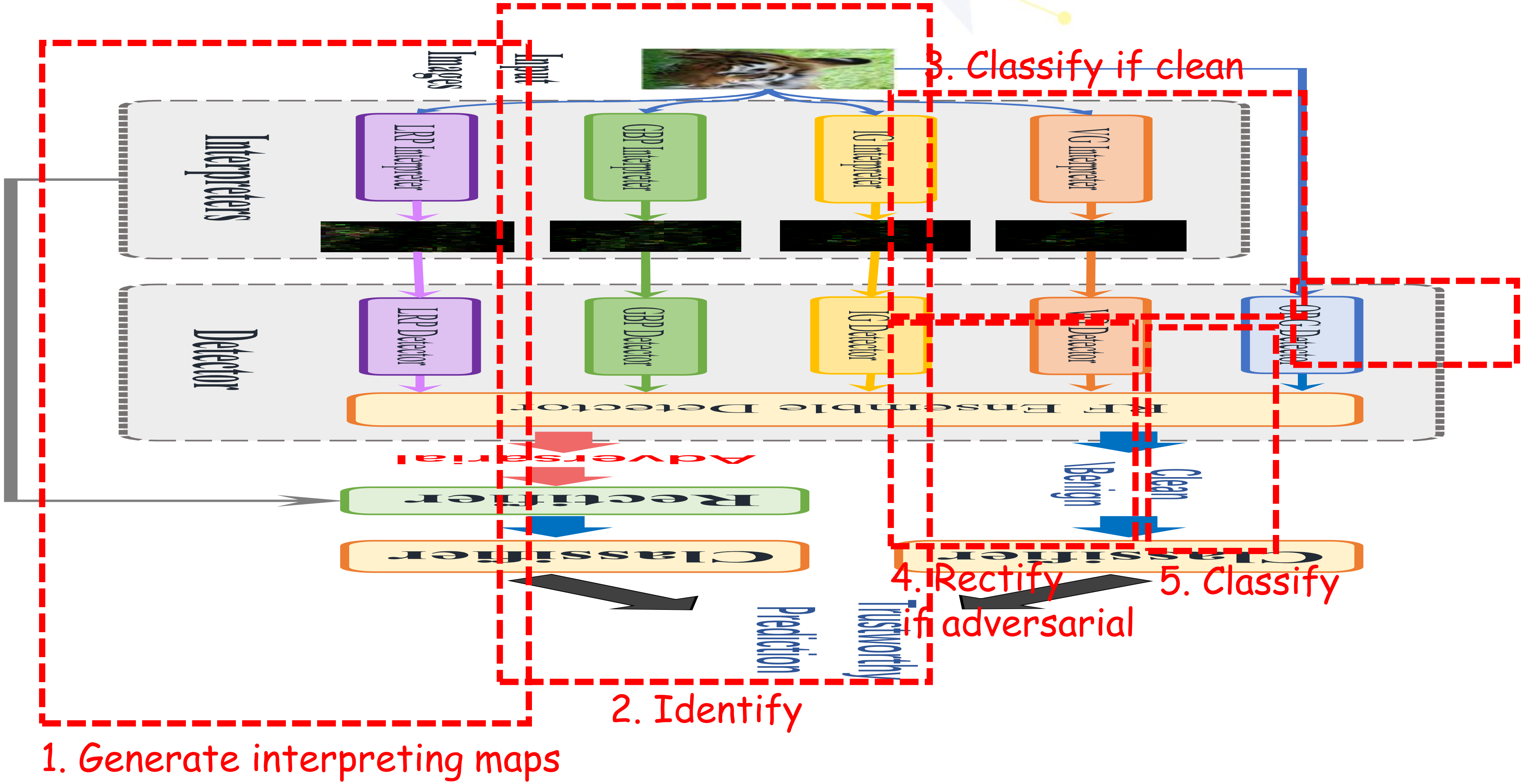
Erased Rate	Deepfool	CW	DDN
top 0%	1.000	1.000	1.000
top 5%	0.637	0.665	0.656

interpreting method → the first challenge

detect

rectify

2. Our Framework: X-Ensemble



2. Our Ensemble Detector: X-Det

2. Our Rectifier

Algorithm 1 Rectified Image For Tuning Rectifier

Variables: $\{D_1, ..., D_j\}$ are the sub-detectors that predict an input image x as an adversarial one, $\{R_1, ..., R_j\}$ are the interpreting methods corresponding to $\{D_1, ..., D_j\}$ respectively, $\alpha \in (0, 1)$ is a threshold parameter, $rand()$ returns a random value in $[0, 1]$, and σ is the variance of pixel values in x.

for $k = 1$ to j **do**
    $E_k \leftarrow Entropy(D_k(x))$
end for
$R \leftarrow R_i$ where $i = argmin(E_1, ..., E_j)$
$g \leftarrow R(x)$
$thres \leftarrow \alpha * (\max(g) - \min(g)) + \min(g)$
for ixel (i, j) in x **do**
    if $g_{i,j} > thres$ **and** $rand() > 0.5$ **then**
        $x_{i,j} \leftarrow x_{i,j} + Normal(0, \sigma)$
    end if
end for
return x

3. Experiment : Setting

Dataset: Fashion-MNIST, CIFAR-10, ImageNet

Attack method: FGSM, PGD, Deepfool, C&W, DDN, OnePixel

Interpreting method: VG, GBP, IG, LRP

Baseline: PD, TWS, MDS for detection part,
Adversarial training, PD, TVM for wholepipeline

3. Experiment Results: Detection

Our RF ensemble detector

Components of our ensemble detector

Grey-Box																		
	Fashion-MNIST									CIFAR10								
Attackers	X-Det	PD	TWS	MDS	VG	IG	GBP	LRP	ORG	X-Det	PD	TWS	MDS	VG	IG	GBP	LRP	ORG
FGSM-U	**1.00**	**1.00**	0.63	0.71	0.97	0.99	**1.00**	0.99	**1.00**	**1.00**	0.98	0.52	0.83	0.88	0.86	0.98	0.99	**1.00**
PGD-U	**1.00**	**1.00**	0.65	0.79	0.98	**1.00**	0.99	0.99	**1.00**	**0.99**	**0.99**	0.52	0.76	**0.99**	0.95	0.96	0.97	0.98
PGD-T	**1.00**	**1.00**	0.83	0.80	0.97	**1.00**	0.99	0.99	**1.00**	0.98	0.96	0.48	0.71	0.93	0.90	0.95	0.98	**1.00**
DFool-U	0.99	0.98	0.99	0.77	0.95	0.99	**1.00**	0.94	0.99	0.98	0.77	0.83	0.93	0.89	0.90	**0.99**	0.92	0.83
CW-U	0.98	0.93	0.95	0.79	0.94	0.98	**1.00**	0.98	0.96	0.98	0.78	0.90	0.93	0.90	0.89	**0.99**	0.92	0.86
CW-T	**1.00**	0.98	0.99	0.83	0.97	**1.00**	**1.00**	**1.00**	0.99	**0.99**	0.84	0.94	0.94	0.93	0.93	**0.99**	0.96	0.95
DDN-U	0.99	0.98	0.80	0.79	0.96	0.99	0.99	**1.00**	0.99	**0.99**	0.70	0.91	0.93	0.91	0.90	0.92	**0.99**	0.90
DDN-T	**1.00**	0.99	**1.00**	0.85	**1.00**	0.90	0.98	**1.00**	**1.00**	**0.99**	0.81	0.96	0.94	**0.99**	0.93	0.95	**0.99**	0.97
Black-Box																		
	Fashion-MNIST									CIFAR10								
Attackers	X-Det	PD	TWS	MDS	VG	IG	GBP	LRP	ORG	X-Det	PD	TWS	MDS	VG	IG	GBP	LRP	ORG
FGSM-U	**1.00**	0.99	0.76	0.54	**1.00**	0.98	0.99	**1.00**	**1.00**	0.98	0.99	0.66	0.93	0.88	0.92	0.99	0.99	**1.00**
PGD-U	**1.00**	0.99	0.77	0.53	**1.00**	0.98	0.99	**1.00**	**1.00**	0.97	0.98	0.57	0.59	0.76	0.80	0.91	0.98	**1.00**
PGD-T	**1.00**	0.99	0.78	0.55	**1.00**	0.97	0.99	**1.00**	**1.00**	0.99	0.99	0.72	0.59	0.78	0.83	0.92	0.96	**1.00**
DFool-U	0.94	0.93	0.81	0.52	0.85	0.94	**0.98**	0.91	0.95	0.79	0.74	0.75	0.54	0.70	**0.80**	**0.80**	**0.80**	0.60
CW-U	0.91	0.87	0.81	0.53	0.83	0.91	**0.99**	0.90	0.86	**0.82**	0.75	0.75	0.53	0.71	**0.82**	0.80	0.81	0.70
CW-T	0.97	0.96	0.80	0.52	0.91	**0.99**	0.98	0.95	0.98	**0.82**	0.77	0.76	0.53	0.80	**0.82**	**0.82**	**0.82**	0.77
DDN-U	0.88	0.86	0.80	0.52	0.82	**0.95**	0.94	0.91	0.93	0.80	0.63	0.76	0.54	0.71	0.80	**0.81**	0.80	0.76
DDN-T	0.98	0.96	0.79	0.54	0.92	0.97	**0.99**	0.96	**0.99**	0.82	0.72	0.76	0.54	0.71	0.80	0.82	0.82	**0.89**

AUC score of adversarial example detection for vaccinated training

3. Experiment Results: Detection

Our ensemble detector

Grey-Box								
	Fashion-MNIST				CIFAR-10			
Attacker	X-Det	PD	l_∞-D	l_2-D	X-Det	PD	l_∞-D	l_2-D
PGD-U	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.99	**1.00**	0.39
PGD-T	**1.00**	**1.00**	0.99	0.91	**1.00**	0.99	**1.00**	0.50
CW-U	0.95	0.93	0.73	**0.97**	**0.98**	0.78	0.49	0.97
CW-T	0.98	0.98	0.84	**0.99**	**0.99**	0.84	0.49	0.98
DDN-U	0.99	0.98	0.80	**1.00**	**0.99**	0.70	0.49	0.98
DDN-T	**1.00**	**1.00**	0.93	**1.00**	**0.99**	0.81	0.49	0.98
OnePixel	**0.82**	0.61	0.59	0.75	**0.83**	0.81	0.51	0.77
Black-Box								
	Fashion-MNIST				CIFAR-10			
Attacker	X-Det	PD	l_∞-D	l_2-D	X-Det	PD	l_∞-D	l_2-D
PGD-U	**0.99**	**0.99**	0.98	0.91	0.99	0.99	**1.00**	0.70
PGD-T	**0.99**	**0.99**	0.98	0.92	0.99	0.99	**1.00**	0.78
CW-U	**0.87**	0.85	0.51	0.73	**0.80**	0.75	0.48	0.77
CW-T	**0.97**	0.93	0.78	0.88	**0.80**	0.77	0.49	0.76
DDN-U	0.85	**0.88**	0.53	0.83	**0.80**	0.63	0.49	0.75
DDN-T	0.95	0.98	0.84	0.90	0.82	0.72	0.48	0.77
OnePixel	**0.73**	0.71	0.57	0.69	**0.72**	0.70	0.51	0.69

AUC score of adversarial example detection for <u>invaccinated </u>training

Note that OnePixel is L_0 attack, while our detectors are trained for L_2 and L_∞

3. Experiment Results: Whole Pipeline

Grey-Box																		
	Fashion-MNIST						CIFAR-10						ImageNet					
	Our	PD	DDN$_a$	PGD$_a$	TVM	F	Our	PD	DDN$_a$	PGD$_a$	TVM	F	Our	PD	DDN$_a$	PGD$_a$	TVM	F
Clean	0.90	0.90	0.86	0.84	0.67	**0.92**	0.82	0.79	0.75	0.64	0.35	**0.86**	0.89	0.66	0.78	0.72	0.75	**0.95**
FGSM-U	**0.84**	0.75	0.82	0.82	0.49	0.56	**0.55**	0.36	0.48	0.43	0.29	0.24	**0.60**	0.47	0.49	0.47	0.36	0.44
PGD-U	**0.79**	0.64	0.80	0.81	0.57	0.27	**0.41**	0.30	0.37	0.35	0.32	0.08	**0.75**	0.70	0.38	0.47	0.66	0.02
PGD-T	**0.89**	0.86	0.84	0.87	0.53	0.66	**0.62**	0.60	0.33	0.48	0.32	0.05	**0.73**	0.66	0.29	0.51	0.70	0.00
Dfool-U	0.87	**0.88**	0.26	0.76	0.65	0.00	**0.71**	0.68	0.19	0.29	0.34	0.00	**0.75**	0.58	0.37	0.35	0.71	0.01
CW-U	0.86	**0.88**	0.70	0.73	0.66	0.00	**0.74**	0.73	0.70	0.63	0.34	0.00	**0.74**	0.64	0.50	0.53	0.71	0.00
CW-T	**0.86**	0.85	0.72	0.53	0.65	0.00	0.74	**0.75**	0.45	0.46	0.33	0.00	**0.79**	0.61	0.40	0.39	0.75	0.00
DDN-U	**0.90**	0.89	0.74	0.76	0.66	0.00	0.69	**0.74**	0.66	0.52	0.34	0.00	**0.76**	0.60	0.56	0.44	0.75	0.03
DDN-T	**0.90**	0.89	0.59	0.64	0.65	0.00	0.71	**0.75**	0.53	0.43	0.34	0.00	**0.79**	0.60	0.50	0.39	0.74	0.00

Black-Box																		
	Fashion-MNIST						CIFAR-10						ImageNet					
	Our	PD	DDN$_a$	PGD$_a$	TVM	F	Our	PD	DDN$_a$	PGD$_a$	TVM	F	Our	PD	DDN$_a$	PGD$_a$	TVM	F
Clean	0.90	0.90	0.86	0.84	0.67	**0.92**	0.82	0.79	0.75	0.64	0.35	**0.86**	0.89	0.66	0.78	0.72	0.75	**0.95**
FGSM-U	**0.72**	0.70	0.68	0.71	0.46	0.50	**0.43**	0.27	0.41	0.41	0.31	0.50	**0.60**	0.49	0.51	0.48	0.54	0.50
PGD-U	0.78	0.80	0.77	**0.82**	0.48	0.50	0.66	**0.70**	0.68	0.58	0.31	0.50	**0.63**	0.61	0.58	0.50	0.51	0.50
PGD-T	0.79	0.78	0.74	**0.81**	0.43	0.50	0.63	**0.73**	0.70	0.59	0.30	0.50	**0.65**	0.52	0.55	0.49	0.50	0.50
Dfool-U	**0.87**	0.86	0.84	**0.87**	0.48	0.50	**0.78**	0.76	0.71	0.61	0.29	0.50	**0.67**	0.60	0.58	0.51	0.43	0.50
CW-U	**0.88**	0.87	0.84	0.87	0.48	0.50	**0.78**	0.75	0.71	0.61	0.30	0.50	**0.65**	0.58	0.51	0.51	0.46	0.50
CW-T	**0.87**	**0.87**	0.84	0.85	0.53	0.50	**0.77**	0.75	0.71	0.60	0.29	0.50	**0.67**	0.45	0.56	0.51	0.44	0.50
DDN-U	**0.88**	0.87	0.84	0.87	0.50	0.50	**0.77**	0.76	0.72	0.61	0.30	0.50	**0.67**	0.43	0.57	0.50	0.45	0.50
DDN-T	**0.88**	0.87	0.84	0.87	0.49	0.50	**0.77**	0.74	0.71	0.60	0.28	0.50	**0.68**	0.36	0.53	0.46	0.41	0.50

Image classification accuracy of X-Ensemble and the baselines

3. Experiment Results: Robustness

X-Ensemble			
	Fashion-MNIST	CIFAR-10	ImageNet
PGD-T	0.87	0.67	0.72
CW-T	0.90	0.69	0.83
DDN-T	0.90	0.71	0.78

Classification accuracy of X-Ensemble under white- box attacks

It shows that our model are robust to the counterattack of adversaries

4. Conclusion

1) We proposed X-Ensemble, an ensembled detection-rectification pipeline on high-performance adversary defense;

2) X-Ensemble combines sub-detectors with random forests to achieve satisfying performance against hybrid and unforeseen attacks;

3) The non-differentiable nature of random forests guarantees the robustness of X-Ensemble under white-box attacks.

Interpretability is a Kind of Safety: An Interpreter-based Ensemble for Adversary Defense






Jingyuan Wang, Yufan Wu, Mingxuan Li, Xin Lin, Junjie Wu*, Chao Li

Beihang University, Beijing, China

* Corresponding author

AUGUST 23-27th KDD2020 Virtual Conference