Attack as Defense: Characterizing Adversarial Examples using Robustness

Zhe Zhao, Guangke Chen, Jingyi Wang,
Yiwei Yang, Fu Song, Jun Sun

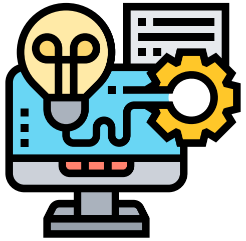
上海科技大学
ShanghaiTech University

ZHEJIANG UNIVERSITY

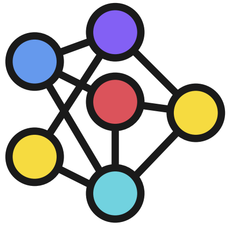
SMU
SINGAPORE MANAGEMENT UNIVERSITY

Zhe Zhao (zhaozhe1@shanghaitech.edu.cn)
Fu Song (songfu@shanghaitech.edu.cn)

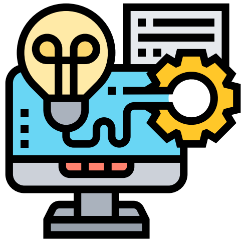
Deep learning and adversarial examples



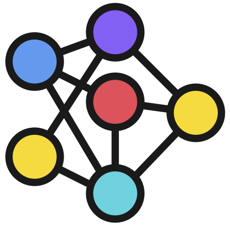
Deep Learning
(DL)



Deep learning and adversarial examples

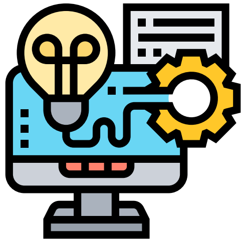


Deep Learning
(DL)

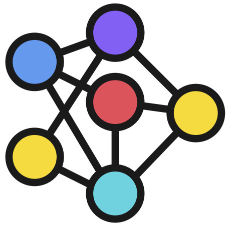


However, DL is vulnerable to adversarial examples…

Deep learning and adversarial examples



Deep Learning
(DL)



However, DL is vulnerable to adversarial examples…

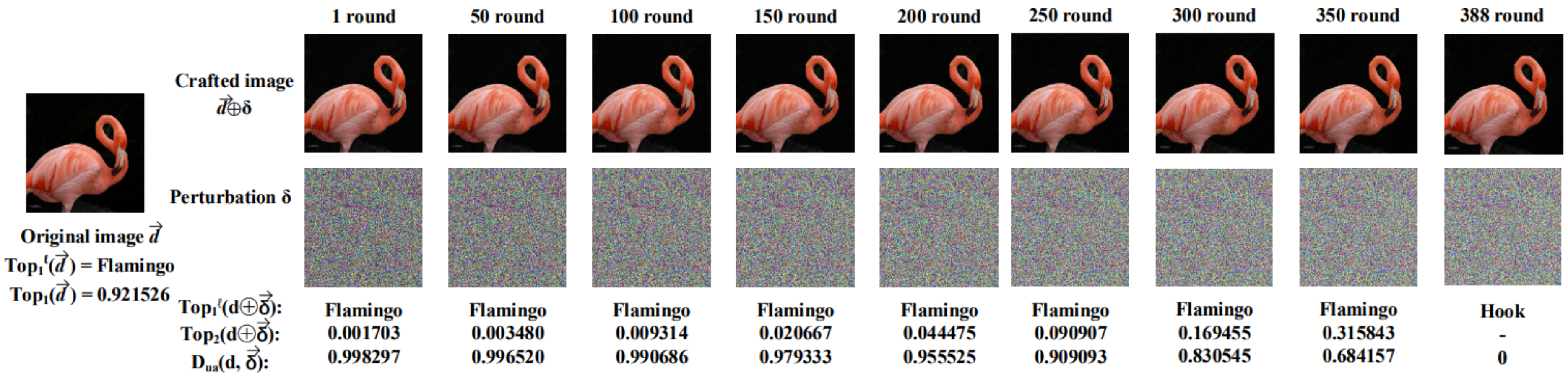


Figure from "Taking Care of The Discretization Problem: A Comprehensive Study of the Discretization Problem and A Black-Box Adversarial Attack in Discrete Integer Domain", Lei Bu; Zhe Zhao; Yuchao Duan; Fu Song.

Attack and defense

An extensive number of adversarial attacks have been proposed since C. Szegedy et al.

White-box attack
Black-box attack

Targeted attack
Untargeted attack

Distance constraint:
L_0, L_2, L_∞

Reference:
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In Proceedings of International Conference on Learning Representations.

Attack and defense

An extensive number of adversarial attacks have been proposed since C. Szegedy et al.

White-box attack
Black-box attack

Targeted attack
Untargeted attack

Distance constraint:
L_0, L_2, L_∞

Attempted defenses against adversarial examples:

Adversarial train

Input transformation

Adversarial detector

Reference:
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In Proceedings of International Conference on Learning Representations.

Attack and defense

Attacks

Adaptive attacks [Athalye et al., 2018]

Optimization-based attacks [Calini and Wagner, 2017]

Iterative attacks [kurakin et al., 2016]

One-step attacks [Goodfellow et al., 2014]

Defenses

Randomization, denoising [Xie et al., 2018; Liao et al., 2018]

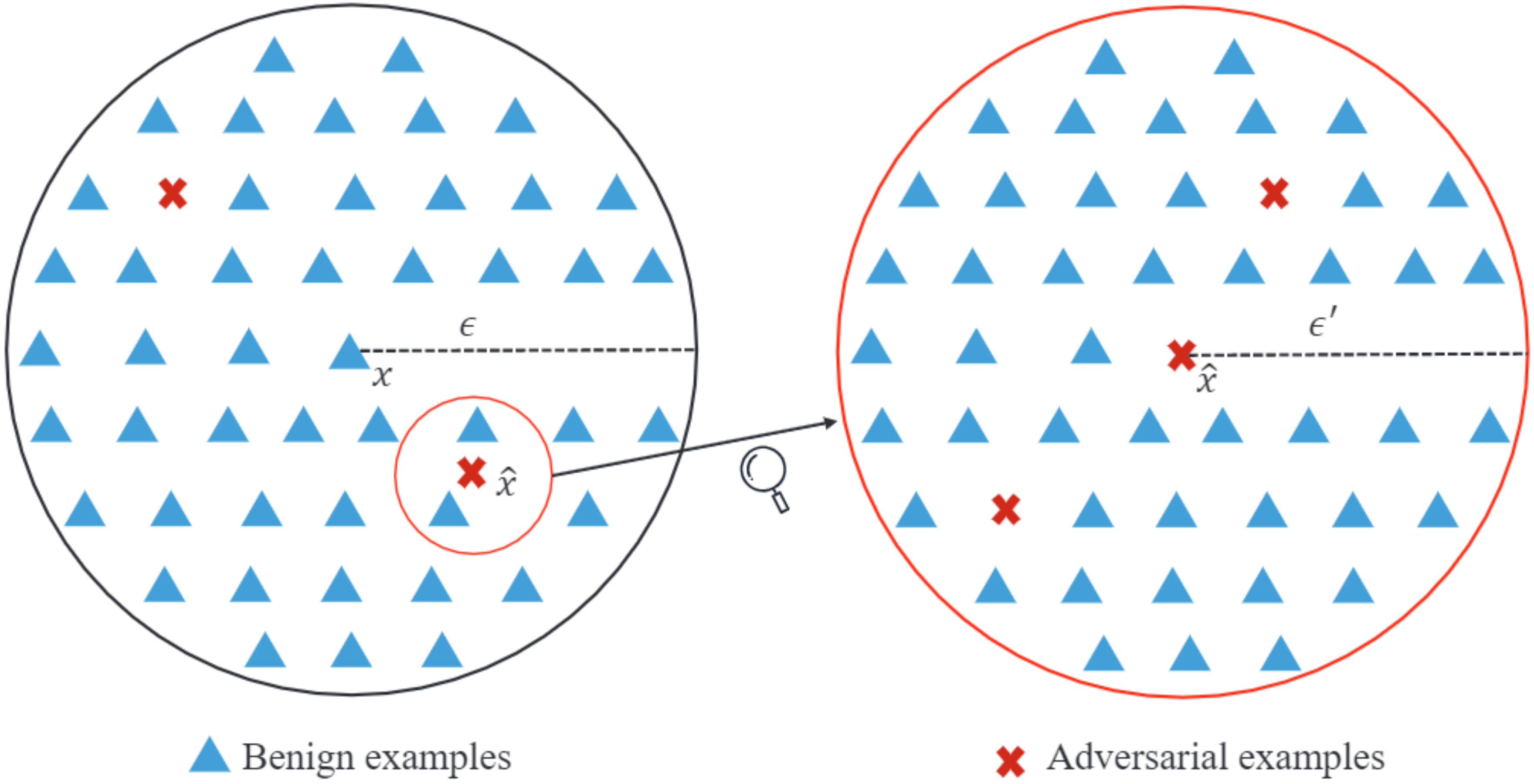
Defensive distillation [Papernot et al., 2016]

Adversarial training with FGSM [Kurakin et al., 2015]

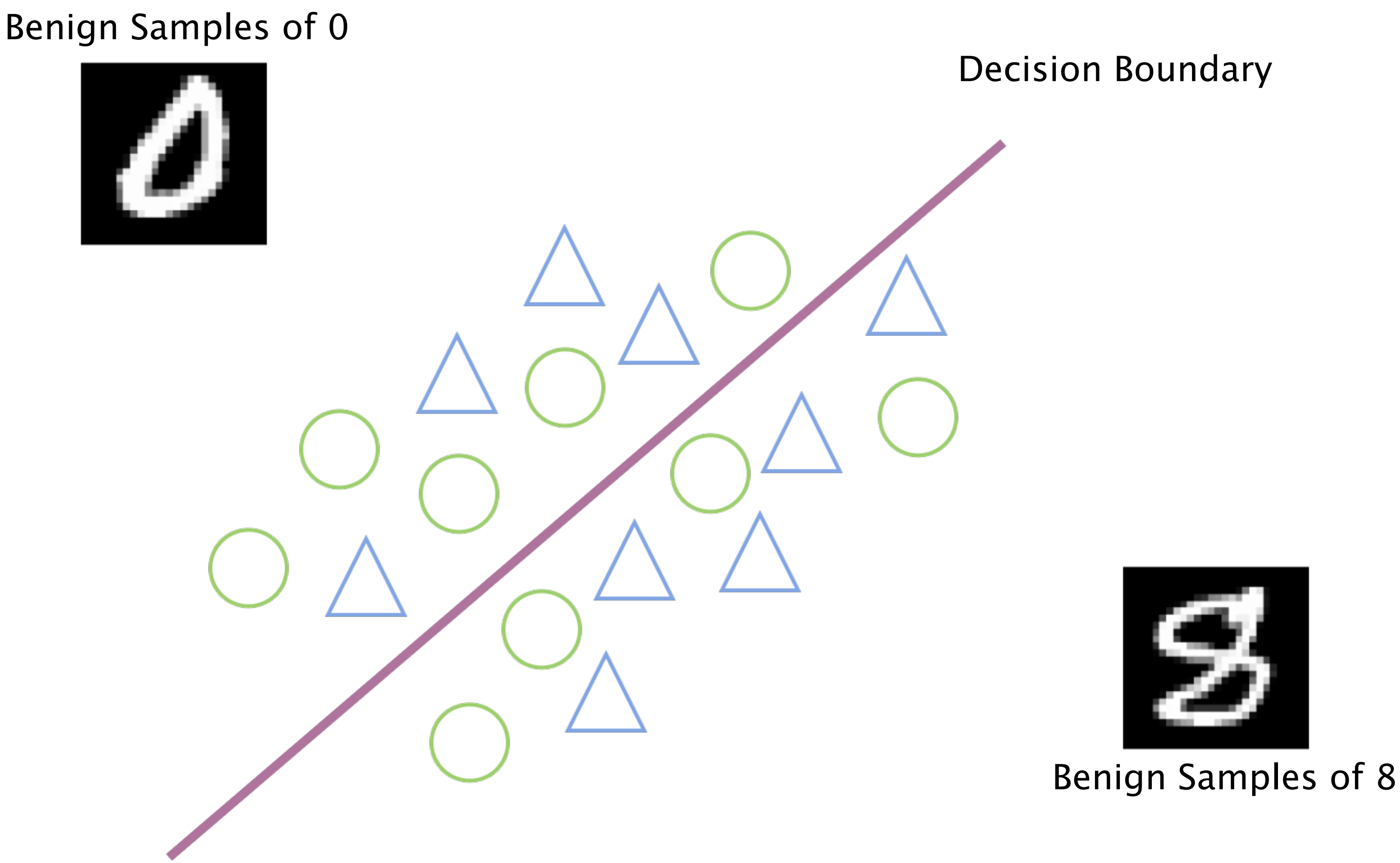
Reference:
Slides made by Prof. Hang Su from AML-CV Workshop at CVPR 2021

Attack as defense: Idea

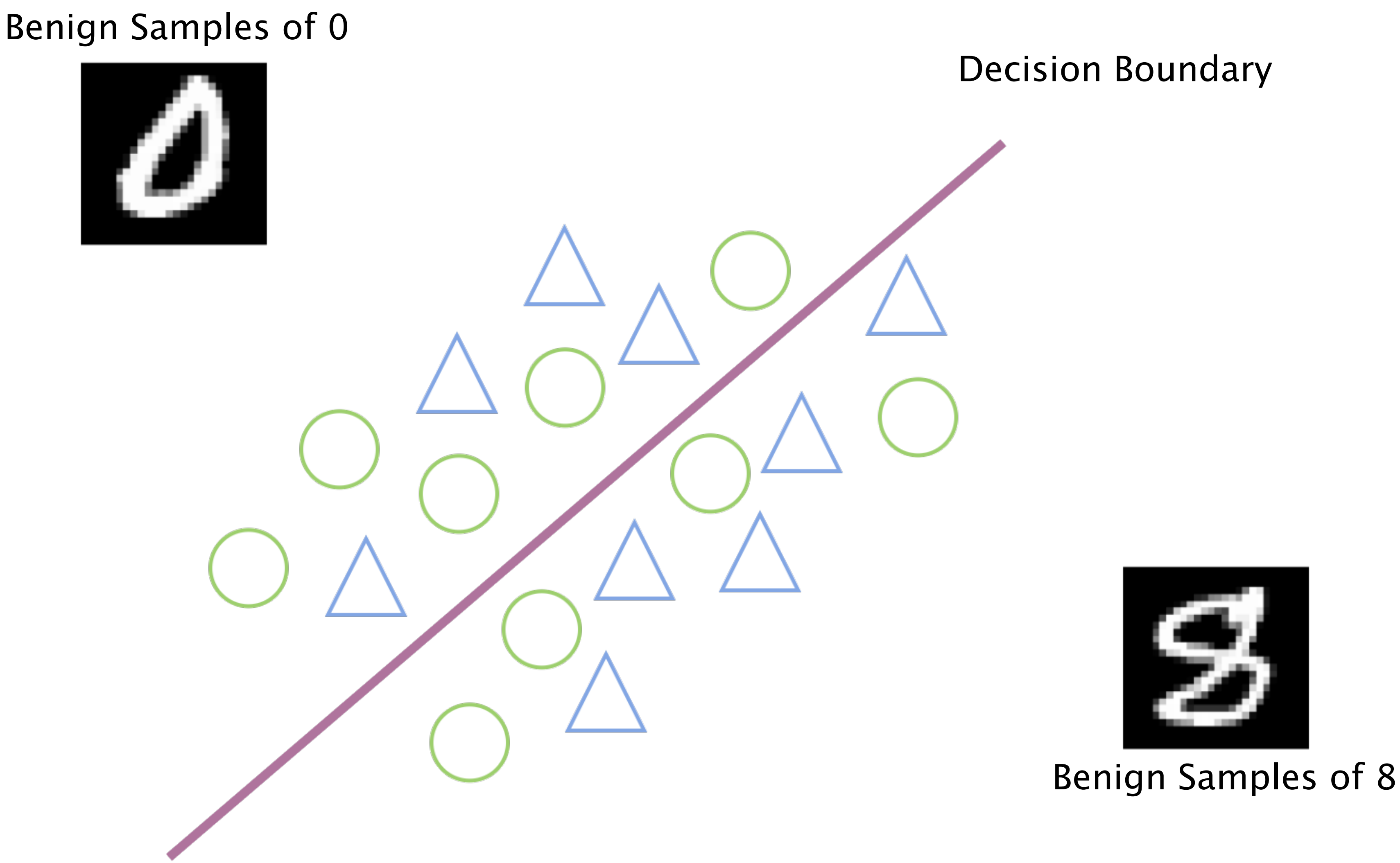
Attack as defense: Idea



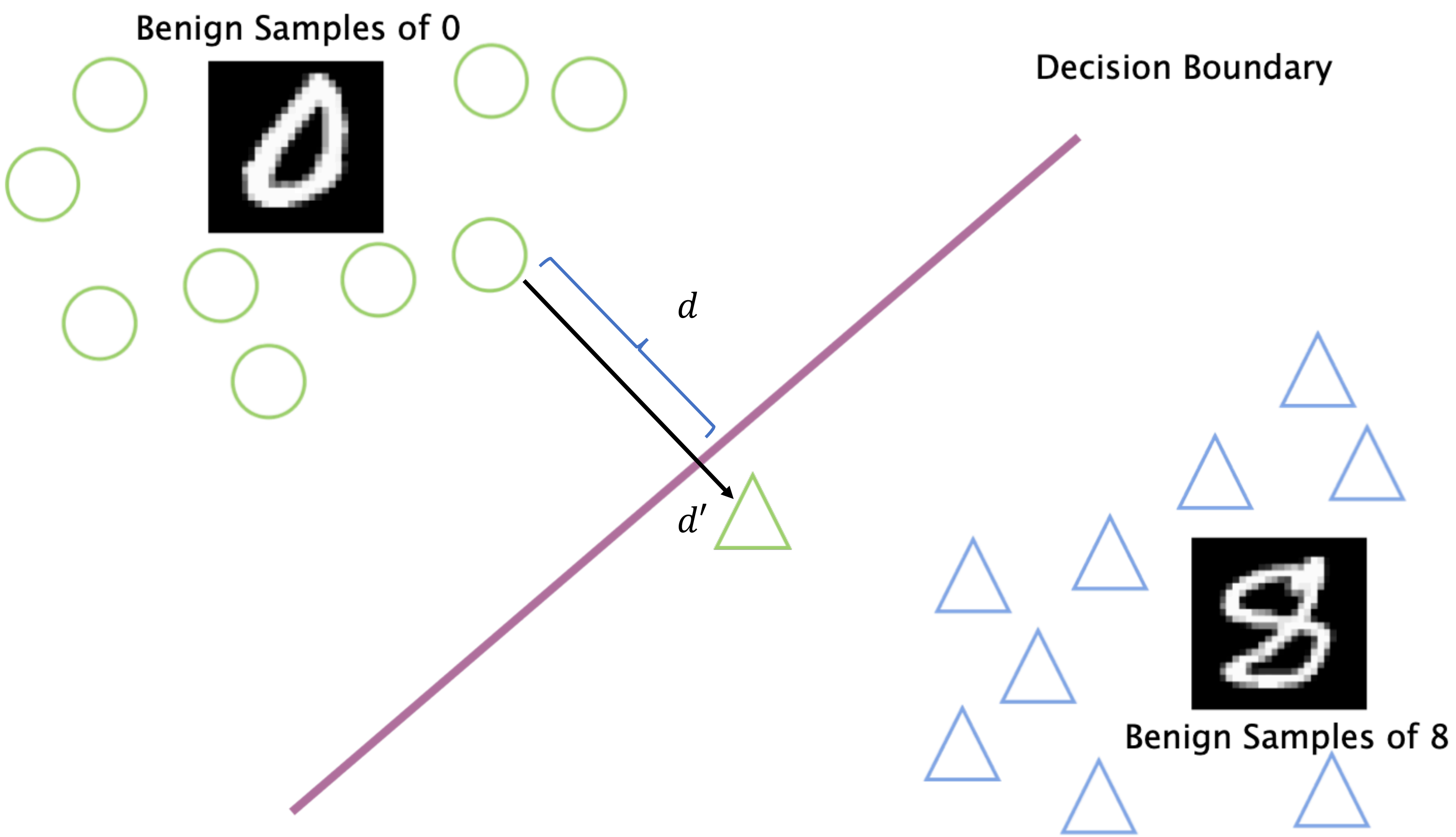
Attack as defense: Intuition



Attack as defense: Intuition



Attack as defense: Intuition



Characterization: Robustness

How to quantify the above observation?

(Local) Robustness

$$\|x - x'\|_p \leq \delta,\ \mathcal{D}(x) = \mathcal{D}(x')$$

Characterization: Robustness

How to quantify the above observation?

(Local) Robustness

$$\|x - x'\|_p \leq \delta,\ \mathcal{D}(x) = \mathcal{D}(x')$$

CLEVER Score

Reference:
Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. In Proceedings of International Conference on Learning Representations.

Characterization: Robustness

How to quantify the above observation?

(Local) Robustness

$$\|x - x'\|_p \leq \delta,\ \mathcal{D}(x) = \mathcal{D}(x')$$

CLEVER Score

Dataset	Label for Evaluate	Benign examples	Adversarial examples								Avg.
			FGSM	λ	BIM	λ	JSMA	λ	C&W	λ	λ
MNIST	Untarget	3.5572 ± 0.3342	0.1093 ± 0.0506	32.55	0.0256 ± 0.0031	138.95	0.0550 ± 0.0060	64.68	0.0004 ± 0.0001	8893	74.77
	Target-2	3.6711 ± 0.3296	0.1148 ± 0.0427	31.98	0.0258 ± 0.0031	142.29	0.0558 ± 0.0063	65.79	0.0004 ± 0.0001	9178	74.62
	Target-5	3.8303 ± 0.3113	0.2047 ± 0.0431	18.71	0.1582 ± 0.0084	24.21	0.1898 ± 0.0096	20.18	0.1384 ± 0.0043	27.68	22.17
	LLC	3.8372 ± 0.3097	0.2390 ± 0.0421	16.06	0.1647 ± 0.0071	23.30	0.2120 ± 0.0076	18.10	0.1406 ± 0.0045	27.29	20.29
CIFAR10	Untarget	0.3851 ± 0.1850	0.2743 ± 0.1627	1.40	0.0329 ± 0.0033	11.71	0.0128 ± 0.0021	30.09	0.0005 ± 0.0002	770	4.81
	Target-2	0.4141 ± 0.1806	0.2971 ± 0.1675	1.39	0.0380 ± 0.0044	10.90	0.0129 ± 0.0021	32.10	0.0005 ± 0.0002	828	4.75
	Target-5	0.4657 ± 0.1913	0.3389 ± 0.1675	1.37	0.0971 ± 0.0117	4.80	0.0610 ± 0.0061	7.63	0.0925 ± 0.0168	5.03	3.16
	LLC	0.4829 ± 0.1913	0.3572 ± 0.1713	1.35	0.1091 ± 0.0132	4.43	0.0918 ± 0.0095	5.26	0.1035 ± 0.0180	4.67	2.92

Reference:
Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. In Proceedings of International Conference on Learning Representations.

Characterization: Robustness

How to quantify the above observation?

(Local) Robustness

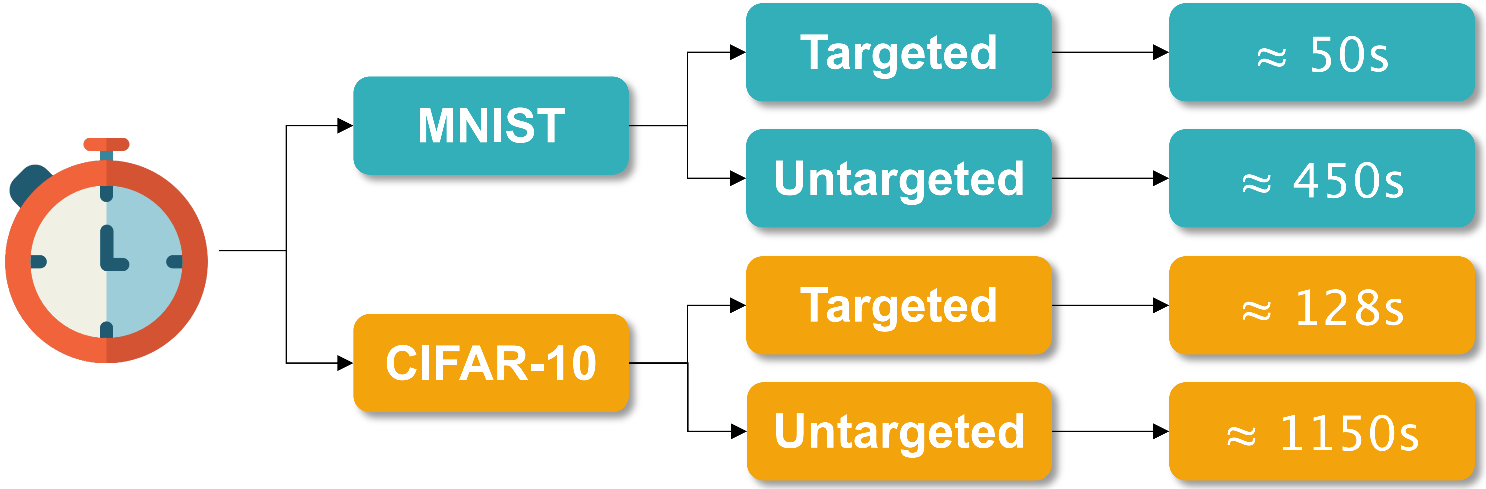
$$\|x - x'\|_p \leq \delta,\ \mathcal{D}(x) = \mathcal{D}(x')$$

CLEVER Score

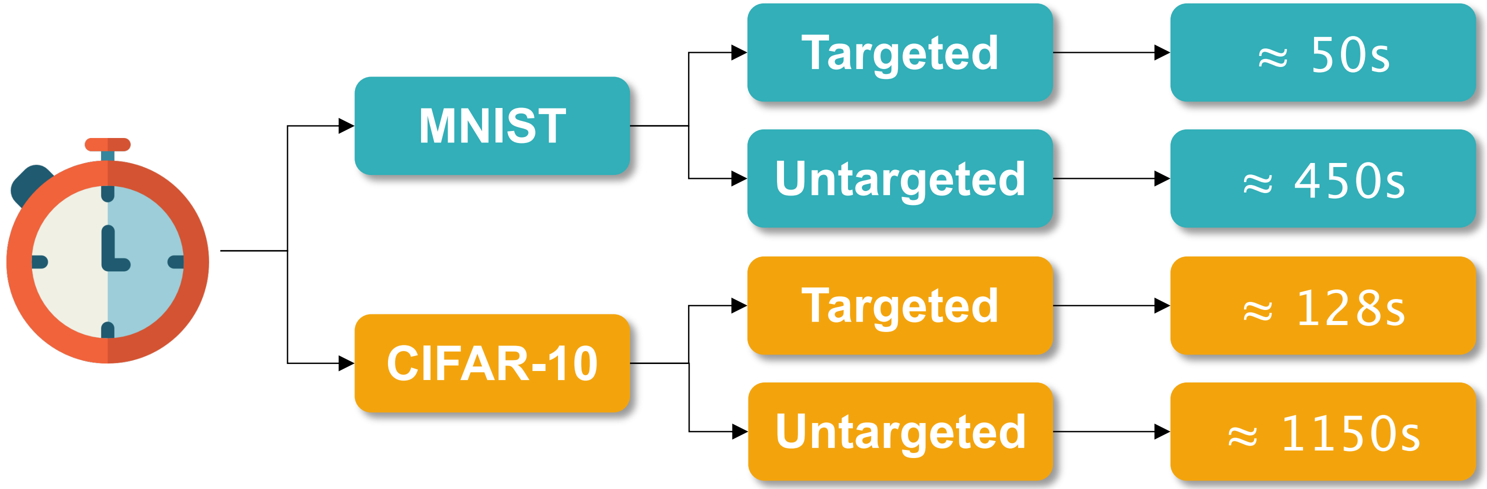
Dataset	Label for Evaluate	Benign examples	Adversarial examples								Avg.
			FGSM	λ	BIM	λ	JSMA	λ	C&W	λ	λ
MNIST	Untarget	3.5572 ± 0.3342	0.1093 ± 0.0506	32.55	0.0256 ± 0.0031	138.95	0.0550 ± 0.0060	64.68	0.0004 ± 0.0001	8893	74.77
	Target-2	3.6711 ± 0.3296	0.1148 ± 0.0427	31.98	0.0258 ± 0.0031	142.29	0.0558 ± 0.0063	65.79	0.0004 ± 0.0001	9178	74.62
	Target-5	3.8303 ± 0.3113	0.2047 ± 0.0431	18.71	0.1582 ± 0.0084	24.21	0.1898 ± 0.0096	20.18	0.1384 ± 0.0043	27.68	22.17
	LLC	3.8372 ± 0.3097	0.2390 ± 0.0421	16.06	0.1647 ± 0.0071	23.30	0.2120 ± 0.0076	18.10	0.1406 ± 0.0045	27.29	20.29
CIFAR10	Untarget	0.3851 ± 0.1850	0.2743 ± 0.1627	1.40	0.0329 ± 0.0033	11.71	0.0128 ± 0.0021	30.09	0.0005 ± 0.0002	770	4.81
	Target-2	0.4141 ± 0.1806	0.2971 ± 0.1675	1.39	0.0380 ± 0.0044	10.90	0.0129 ± 0.0021	32.10	0.0005 ± 0.0002	828	4.75
	Target-5	0.4657 ± 0.1913	0.3389 ± 0.1675	1.37	0.0971 ± 0.0117	4.80	0.0610 ± 0.0061	7.63	0.0925 ± 0.0168	5.03	3.16
	LLC	0.4829 ± 0.1913	0.3572 ± 0.1713	1.35	0.1091 ± 0.0132	4.43	0.0918 ± 0.0095	5.26	0.1035 ± 0.0180	4.67	2.92

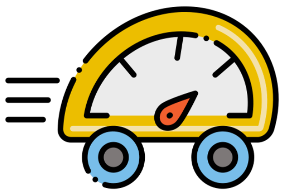
Reference:
Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. In Proceedings of International Conference on Learning Representations.

Characterization: Attack Costs



Characterization: Attack Costs



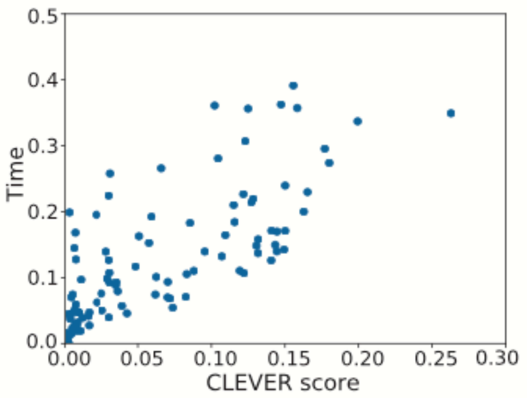


Characterization: Attack Costs

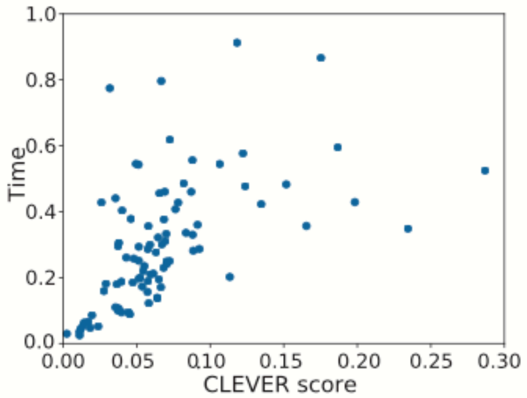
More robust, more difficult to attack.

Verification

Approximation



(a) Score vs. time on MNIST



(b) Score vs. time on CIFAR10

Characterization: Attack Costs

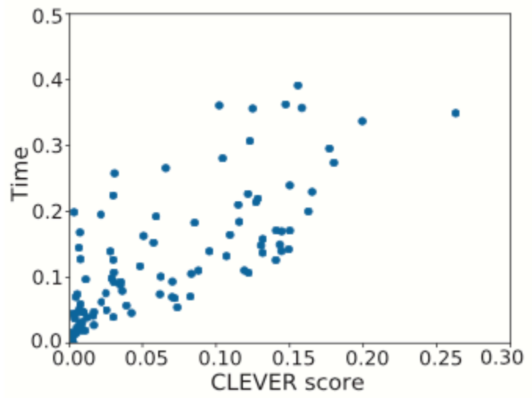
More robust, more difficult to attack.

Verification

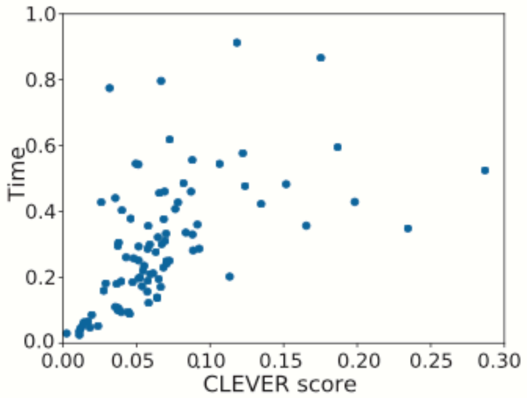
Approximation

Time

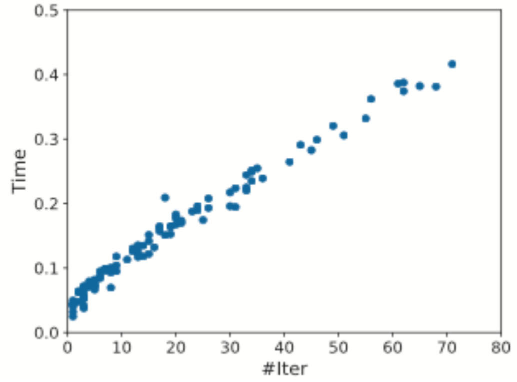
Iteration



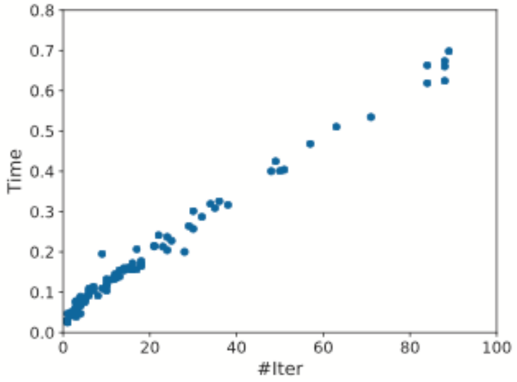
(a) Score vs. time on MNIST



(b) Score vs. time on CIFAR10

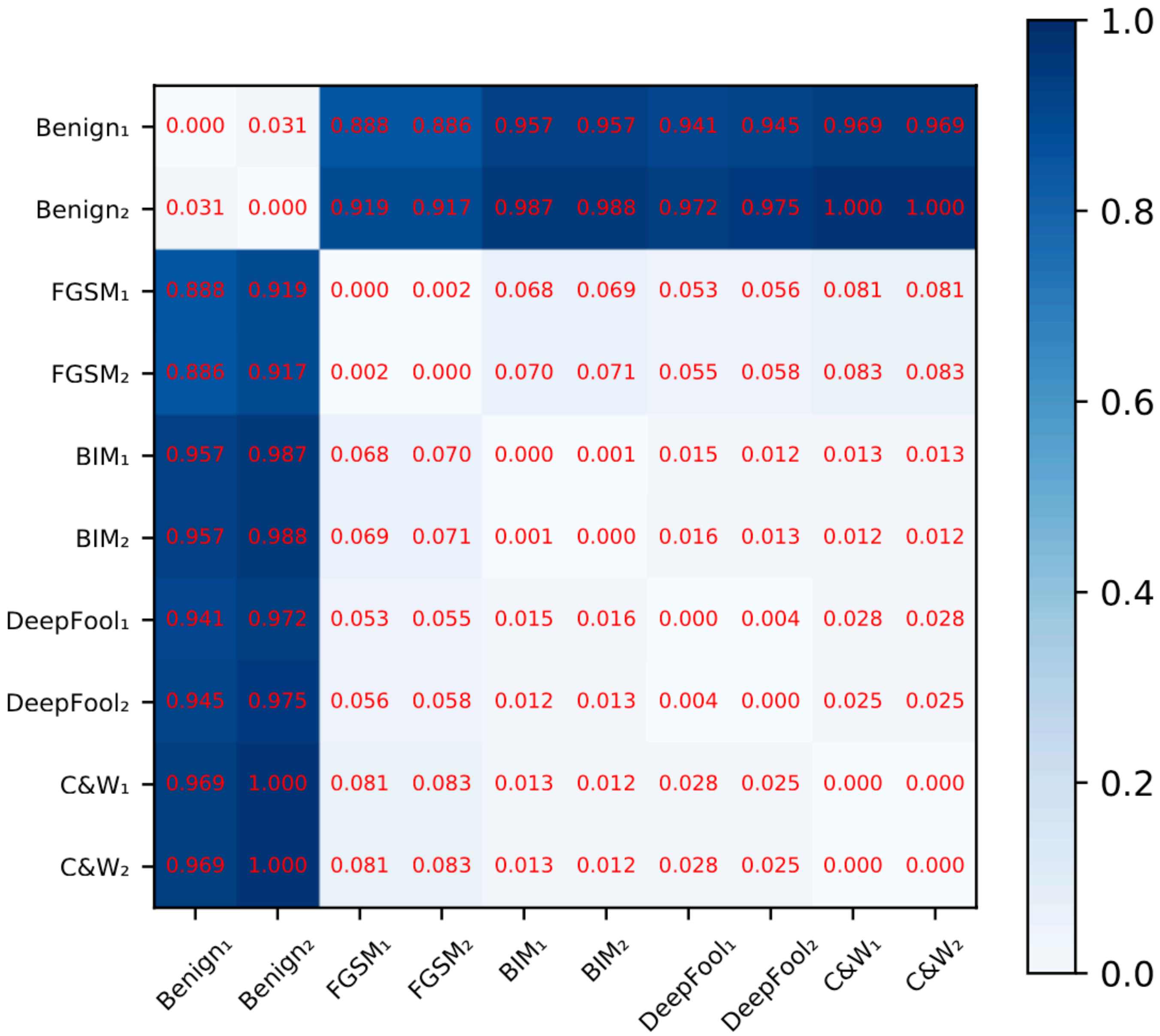


(a) Time vs. ♯iter on MNIST



(b) Time vs. ♯iter on CIFAR10

Characterization: Attack Costs



Detection Approach

K-NN based

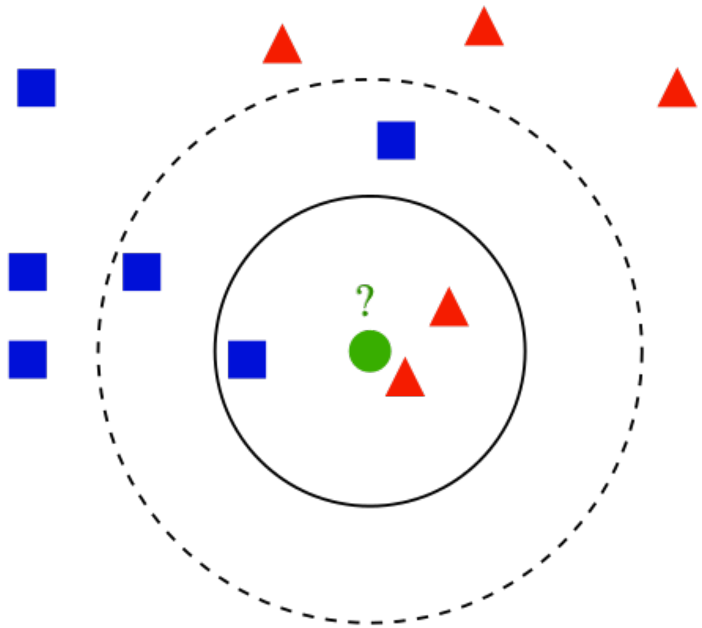
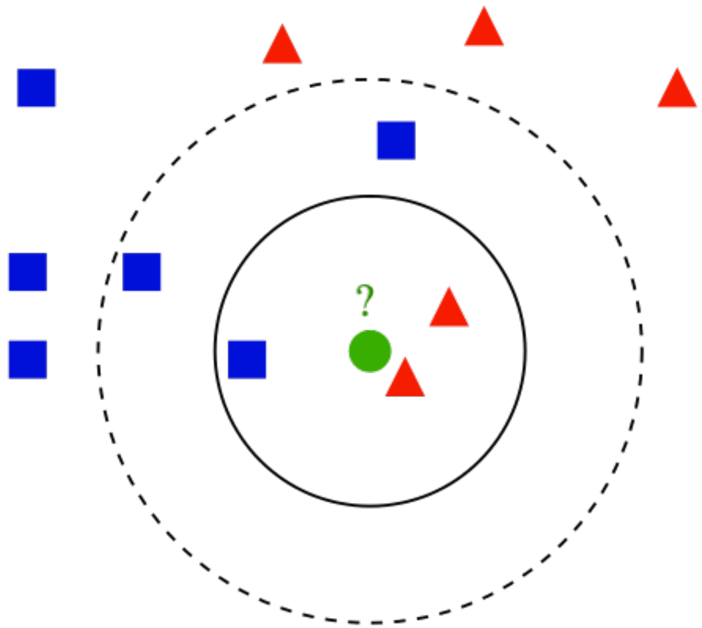


Figure from Wikipedia

Detection Approach

K-NN based



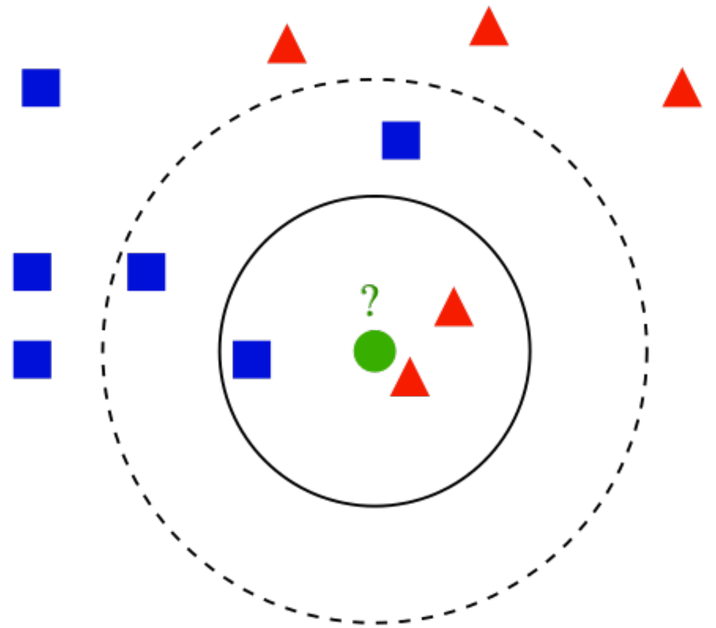
k – nearest neighbors based detector

Training set contains: benign samples attack costs
adv examples attack costs

Figure from Wikipedia

Detection Approach

K-NN based

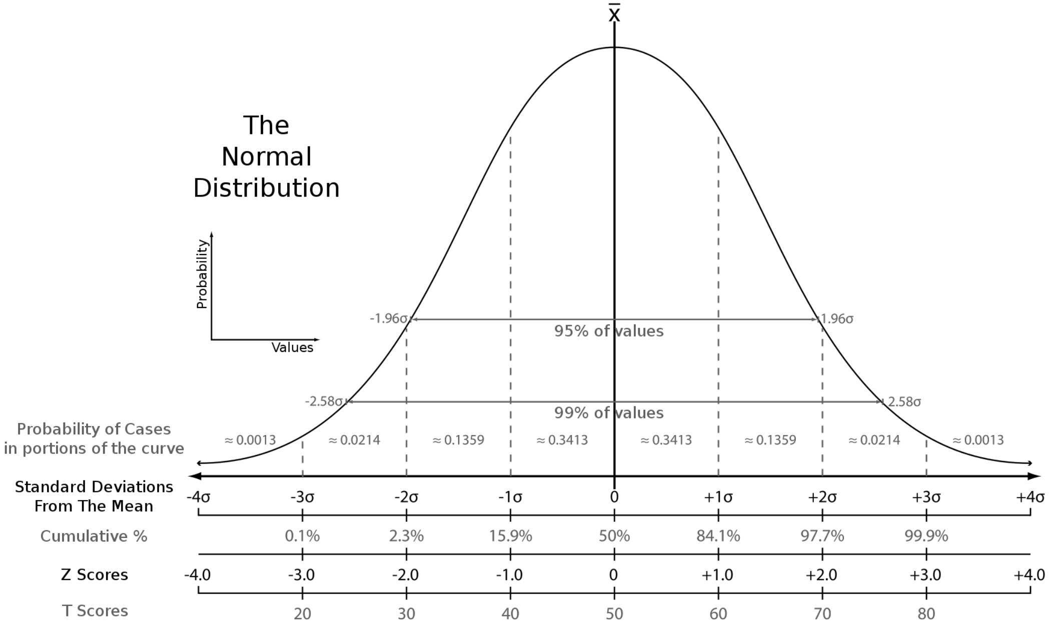


k – nearest neighbors based detector

Training set contains: benign samples attack costs
adv examples attack costs

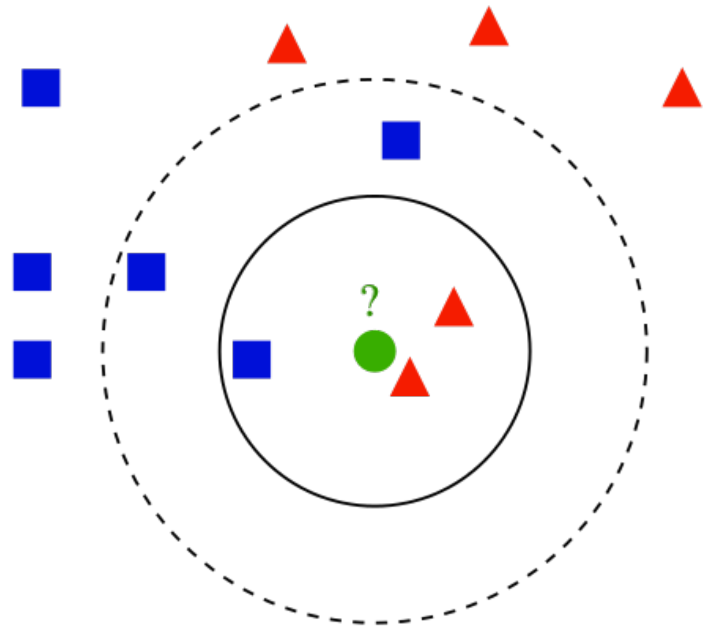
Figure from Wikipedia

Z-score based



Detection Approach

K-NN based

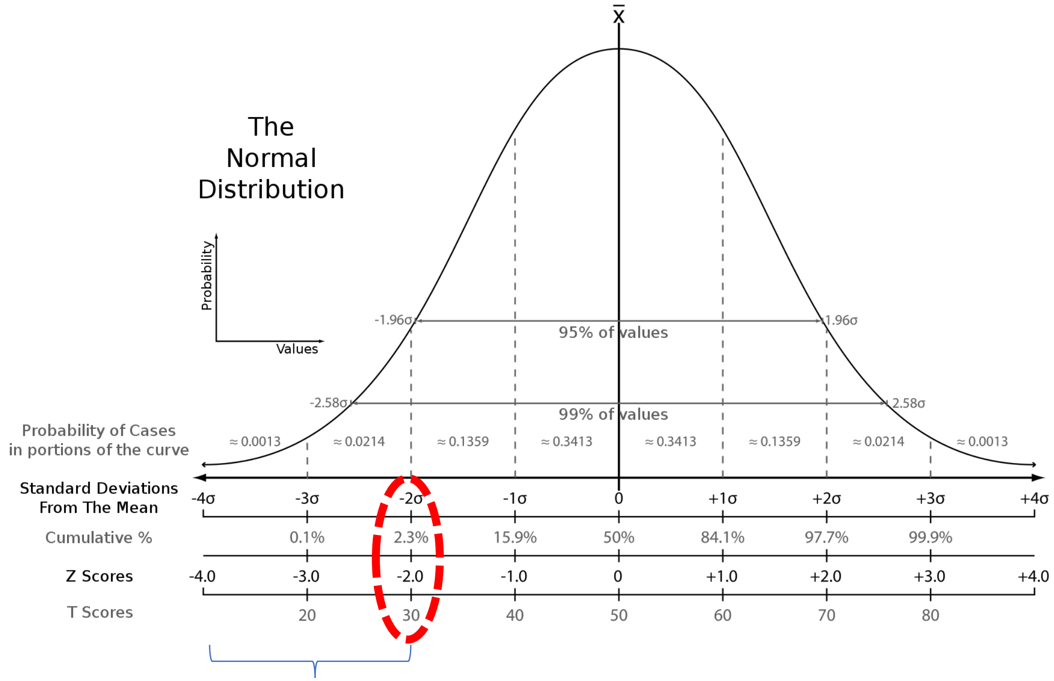


k – nearest neighbors based detector

Training set contains: benign samples attack costs
adv examples attack costs

Figure from Wikipedia

Z-score based



Statistics based detector

Only needs benign samples for training

If $z_x < h$, then x is adversarial example

Figure from Wikipedia

Ensemble Detection Approach

Different attack methods have different characteristics.

Can these ‘attack as defense’ methods be combined?

Figure from Wikipedia

Ensemble Detection Approach

Different attack methods have different characteristics.

Can these 'attack as defense' methods be combined?

K-NN based

Train the detector with n-dimension attack iterations, where n is the number of attacks.

Z-score based

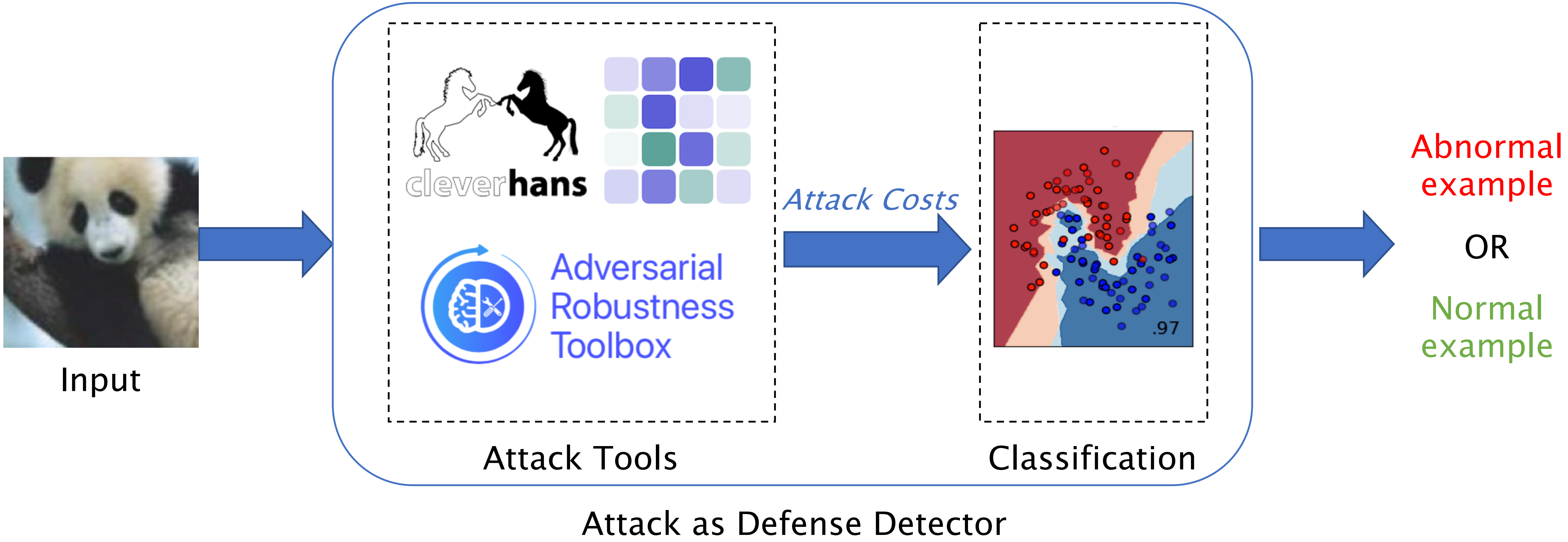
For each attack, we can construct a Z-Score detector, so we have n independent detectors.

Consider k as a hyper-parameter, the ensemble detector classifies an input to adversarial if at least k detectors classify the input to adversarial, otherwise benign.

Overview

Attack

Overview



Input
cleverhans
Adversarial
Robustness
Toolbox
Attack Costs
.97
Attack Tools
Classification
Attack as Defense Detector
Abnormal
example
OR
Normal
example

Experiments

Experiments

RQ1. How to select effective attacks for defense?

- Generate adversarial examples with codes and models from [1]
- Select 8 famous adversarial attack methods as defense
- Implemented by Foolbox (https://github.com/bethgelab/foolbox)
- Compare the attack costs between benign and adversarial examples

RQ2. How effective are the selected attacks for defense?

RQ3. How effective and efficient is A^2D (i.e., detection)?

Reference:
[1] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner.2017. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410 (2017).

RQ1: How to select effective attacks for defense?

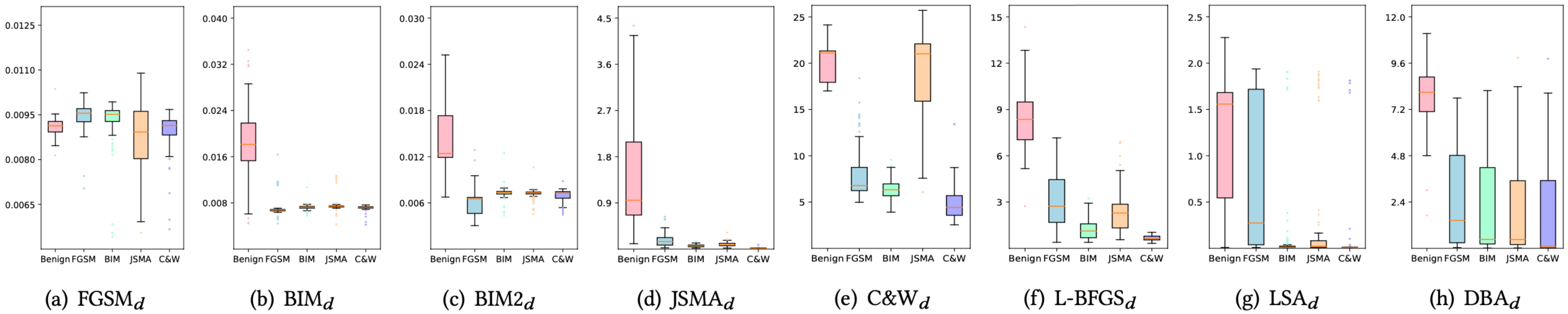
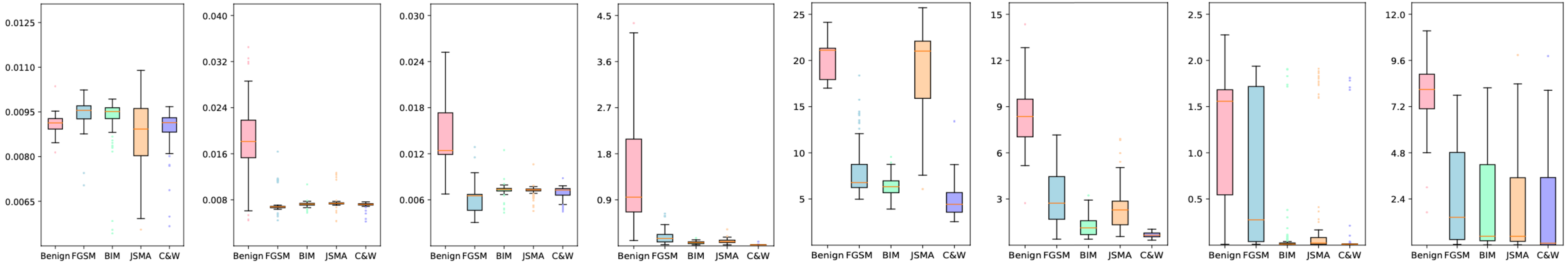


Figure. Attack time of benign and adversarial examples, where y-axis means seconds

RQ1: How to select effective attacks for defense?



(a) FGSM_d (b) BIM_d (c) BIM2_d (d) JSMA_d (e) $\text{C\&W}_d$ (f) L-BFGS_d (g) LSA_d (h) DBA_d

RQ1: How to select effective attacks for defense?

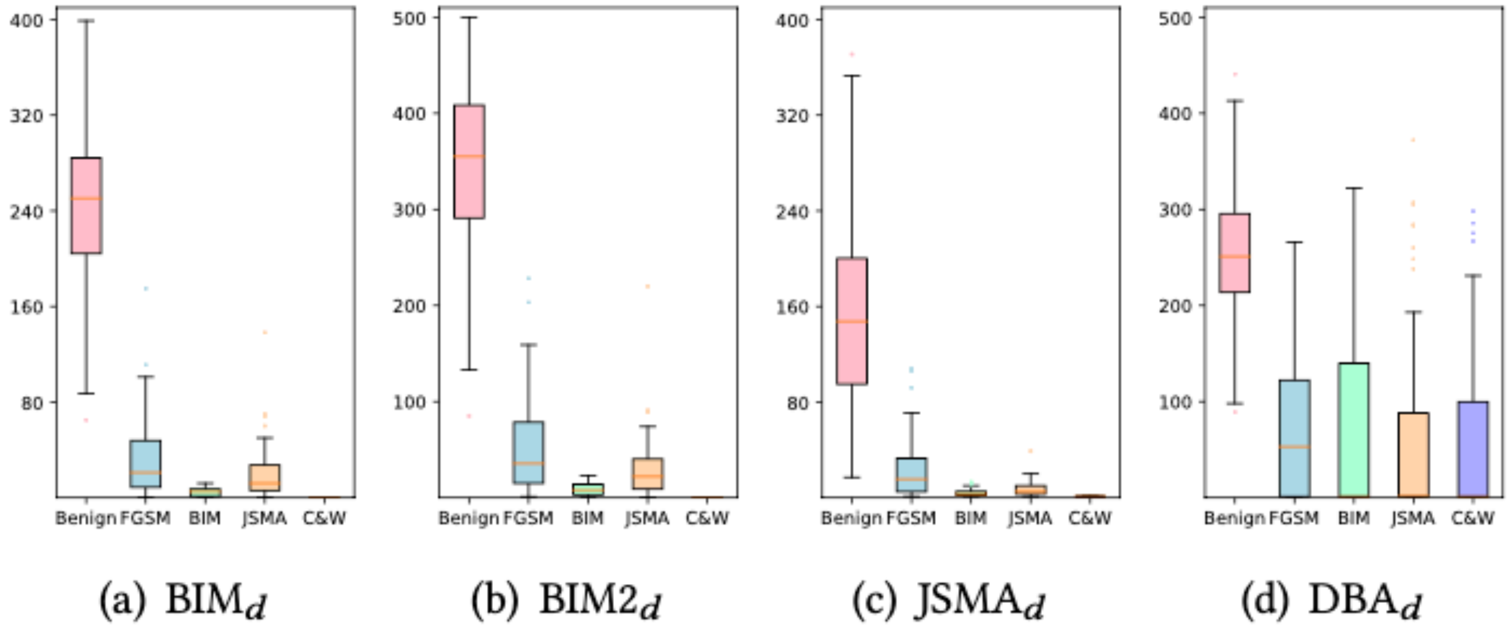


Figure. Attack iterations of benign and adversarial examples

Answer to RQ1: Both attack time and the number of iterations can be used to select effective attacks for defense, while non-iterative attacks are not effective.

White-box Attack

Black-box Attack

L_0 Distance Metrics

L_2 Distance Metrics

L_∞ Distance Metrics

Experiments

RQ1. How to select effective attacks for defense?

RQ2. How effective are the selected attacks for defense?

- Select 4 baselines,
  KD+BU, LID (ICLR'18), mMutant (ICSE'19), Dissector (ICSE'20)
- Evaluation metric: AUROC
- For a fair comparison, we conduct comparison directly using the same target models and attacks provided by baselines

RQ3. How effective and efficient is A^2D (i.e., detection)?

Reference:
[1] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewick- rema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. 2018. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. In Proceedings of International Conference on Learning Representations.
[2] Jingyi Wang, Guoliang Dong, Jun Sun, Xinyu Wang, and Peixin Zhang. 2019. Adversarial sample detection for deep neural network through model mutation testing. In Proceedings of the 41st International Conference on Software Engineering. IEEE, 1245–1256.
[3] Huiyan Wang, Jingwei Xu, Chang Xu, Xiaoxing Ma, and Jian Lu. 2020. Dissector: Input Validation for Deep Learning Applications by Crossing-layer Dissection. In The 42th International Conference on Software Engineering. ACM, 727–738.

RQ2: How effective are the selected attacks for defense?

Env_1	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_1	BL_2
MNIST	FGSM	0.9653	**0.9922**	0.9883	0.9504	0.8267	0.9161
	BIM	0.9986	**0.9996**	0.9995	0.9625	0.9786	0.9695
	JSMA	**0.9923**	0.9922	0.9914	0.9497	0.9855	0.9656
	C&W	**1.0**	**1.0**	**1.0**	0.9672	0.9794	0.9502
CIFAR10	FGSM	0.6537	0.712	0.6474	0.6977	0.7015	**0.7891**
	BIM	0.8558	**0.8636**	0.861	0.8276	0.8255	0.8496
	JSMA	0.9459	**0.955**	0.9526	0.9452	0.8421	0.9475
	C&W	0.9905	0.9984	**0.9988**	0.9833	0.9217	0.9799

Env_2	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_3
MNIST	FGSM	0.9665	**0.9883**	0.9846	0.9595	0.9617
	JSMA	0.9971	**0.9984**	0.9974	0.984	0.9941
	DeepFool	0.9918	**0.9971**	0.9951	0.9587	0.9817
	C&W	0.9456	**0.9870**	0.9769	0.8672	0.9576
	BB	0.9746	**0.9895**	0.9852	0.9535	0.9677
CIFAR10	FGSM	0.8808	0.8994	**0.8998**	0.8746	0.8617
	JSMA	0.9774	**0.9890**	0.9873	0.9566	0.9682
	DeepFool	0.9832	0.9898	**0.9902**	0.9769	0.9614
	C&W	0.8842	**0.9176**	0.9175	0.9004	0.9063

Env_3	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_4
MNIST	FGSM	0.9985	0.9999	**1.0**	0.9674	0.9993
	JSMA	0.9972	0.9998	**0.9999**	0.9113	0.9993
	DeepFool	0.9702	0.9877	0.9874	0.9255	**0.9892**
	C&W	0.9985	**1.0**	**1.0**	0.9623	0.9996
CIFAR10	FGSM	0.9945	0.9979	**0.9983**	0.9629	0.9981
	JSMA	0.9934	0.9962	0.9961	0.976	**0.9966**
	DeepFool	**0.9713**	0.9703	0.9692	0.9604	0.9618
	C&W	0.9951	0.9981	**0.9985**	0.9928	0.9968
ImageNet	FGSM	0.973	0.9763	**0.9782**	0.9625	0.9617
	JSMA	**0.9962**	0.9805	0.99	0.9937	0.9695
	DeepFool	**0.9958**	0.9793	0.9892	0.9891	0.9924
	C&W	0.9873	0.9731	0.9801	**0.9924**	0.9636

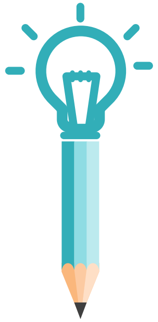
Answer to RQ2: Against most attacks on 3 environments, the selected white-box attacks $JSMA_d$, BIM_d and $BIM2_d$ are more effective than the baselines.

RQ2: How effective are the selected attacks for defense?

Env_1	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_1	BL_2
MNIST	FGSM	0.9653	**0.9922**	0.9883	0.9504	0.8267	0.9161
	BIM	0.9986	**0.9996**	0.9995	0.9625	0.9786	0.9695
	JSMA	**0.9923**	0.9922	0.9914	0.9497	0.9855	0.9656
	C&W	**1.0**	**1.0**	**1.0**	0.9672	0.9794	0.9502
CIFAR10	FGSM	0.6537	0.712	0.6474	0.6977	0.7015	**0.7891**
	BIM	0.8558	**0.8636**	0.861	0.8276	0.8255	0.8496
	JSMA	0.9459	**0.955**	0.9526	0.9452	0.8421	0.9475
	C&W	0.9905	0.9984	**0.9988**	0.9833	0.9217	0.9799

Env_2	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_3
MNIST	FGSM	0.9665	**0.9883**	0.9846	0.9595	0.9617
	JSMA	0.9971	**0.9984**	0.9974	0.984	0.9941
	DeepFool	0.9918	**0.9971**	0.9951	0.9587	0.9817
	C&W	0.9456	**0.9870**	0.9769	0.8672	0.9576
	BB	0.9746	**0.9895**	0.9852	0.9535	0.9677
CIFAR10	FGSM	0.8808	0.8994	**0.8998**	0.8746	0.8617
	JSMA	0.9774	**0.9890**	0.9873	0.9566	0.9682
	DeepFool	0.9832	0.9898	**0.9902**	0.9769	0.9614
	C&W	0.8842	**0.9176**	0.9175	0.9004	0.9063

Env_3	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_4
MNIST	FGSM	0.9985	0.9999	**1.0**	0.9674	0.9993
	JSMA	0.9972	0.9998	**0.9999**	0.9113	0.9993
	DeepFool	0.9702	0.9877	0.9874	0.9255	**0.9892**
	C&W	0.9985	**1.0**	**1.0**	0.9623	0.9996
CIFAR10	FGSM	0.9945	0.9979	**0.9983**	0.9629	0.9981
	JSMA	0.9934	0.9962	0.9961	0.976	**0.9966**
	DeepFool	**0.9713**	0.9703	0.9692	0.9604	0.9618
	C&W	0.9951	0.9981	**0.9985**	0.9928	0.9968
ImageNet	FGSM	0.973	0.9763	**0.9782**	0.9625	0.9617
	JSMA	**0.9962**	0.9805	0.99	0.9937	0.9695
	DeepFool	**0.9958**	0.9793	0.9892	0.9891	0.9924
	C&W	0.9873	0.9731	0.9801	**0.9924**	0.9636



Q: Why the AUROC results on ImageNet of $JSMA_d$ and DBA_d are close to or surpass BIM_d?

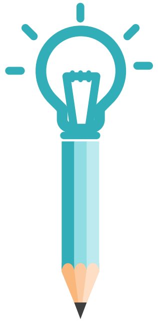
A: Image dimension.

RQ2: How effective are the selected attacks for defense?

Env_1	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_1	BL_2
MNIST	FGSM	0.9653	**0.9922**	0.9883	0.9504	0.8267	0.9161
	BIM	0.9986	**0.9996**	0.9995	0.9625	0.9786	0.9695
	JSMA	**0.9923**	0.9922	0.9914	0.9497	0.9855	0.9656
	C&W	**1.0**	**1.0**	**1.0**	0.9672	0.9794	0.9502
CIFAR10	FGSM	0.6537	0.712	0.6474	0.6977	0.7015	**0.7891**
	BIM	0.8558	**0.8636**	0.861	0.8276	0.8255	0.8496
	JSMA	0.9459	**0.955**	0.9526	0.9452	0.8421	0.9475
	C&W	0.9905	0.9984	**0.9988**	0.9833	0.9217	0.9799

Env_2	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_3
MNIST	FGSM	0.9665	**0.9883**	0.9846	0.9595	0.9617
	JSMA	0.9971	**0.9984**	0.9974	0.984	0.9941
	DeepFool	0.9918	**0.9971**	0.9951	0.9587	0.9817
	C&W	0.9456	**0.9870**	0.9769	0.8672	0.9576
	BB	0.9746	**0.9895**	0.9852	0.9535	0.9677
CIFAR10	FGSM	0.8808	0.8994	**0.8998**	0.8746	0.8617
	JSMA	0.9774	**0.9890**	0.9873	0.9566	0.9682
	DeepFool	0.9832	0.9898	**0.9902**	0.9769	0.9614
	C&W	0.8842	**0.9176**	0.9175	0.9004	0.9063

Env_3	Attack	$JSMA_d$	BIM_d	$BIM2_d$	DBA_d	BL_4
MNIST	FGSM	0.9985	0.9999	**1.0**	0.9674	0.9993
	JSMA	0.9972	0.9998	**0.9999**	0.9113	0.9993
	DeepFool	0.9702	0.9877	0.9874	0.9255	**0.9892**
	C&W	0.9985	**1.0**	**1.0**	0.9623	0.9996
CIFAR10	FGSM	0.9945	0.9979	**0.9983**	0.9629	0.9981
	JSMA	0.9934	0.9962	0.9961	0.976	**0.9966**
	DeepFool	**0.9713**	0.9703	0.9692	0.9604	0.9618
	C&W	0.9951	0.9981	**0.9985**	0.9928	0.9968
ImageNet	FGSM	0.973	0.9763	**0.9782**	0.9625	0.9617
	JSMA	**0.9962**	0.9805	0.99	0.9937	0.9695
	DeepFool	**0.9958**	0.9793	0.9892	0.9891	0.9924
	C&W	0.9873	0.9731	0.9801	**0.9924**	0.9636



Q: Why the AUROC results on ImageNet of $JSMA_d$ and DBA_d are close to or surpass BIM_d?

A: Image dimension.

Q: Why BL_2 performs better than the others on CIFAR10 adversarial examples crafted by FGSM?

A: Model accuracy.

Experiments

RQ1. How to select effective attacks for defense?

RQ2. How effective are the selected attacks for defense?

RQ3. How effective and efficient is A^2D (i.e., detection)?

- K-NN based detectors and Z-Score based detectors
- Evaluation metric: detection accuracy

RQ3. How effective and efficient is A2D (i.e., detection)?

Using K-NN based detector on MNIST dataset as a demo:

The average detection accuracy and time cost:

- $JSMA_d$: 90.84%, 1.8ms
- BIM_d : 98.09%, 2.1ms
- $BIM2_d$: 96.17%, 2.1ms
- DBA_d : 87.42%, 11ms
- END (Ensemble detector) : 99.35%, NA

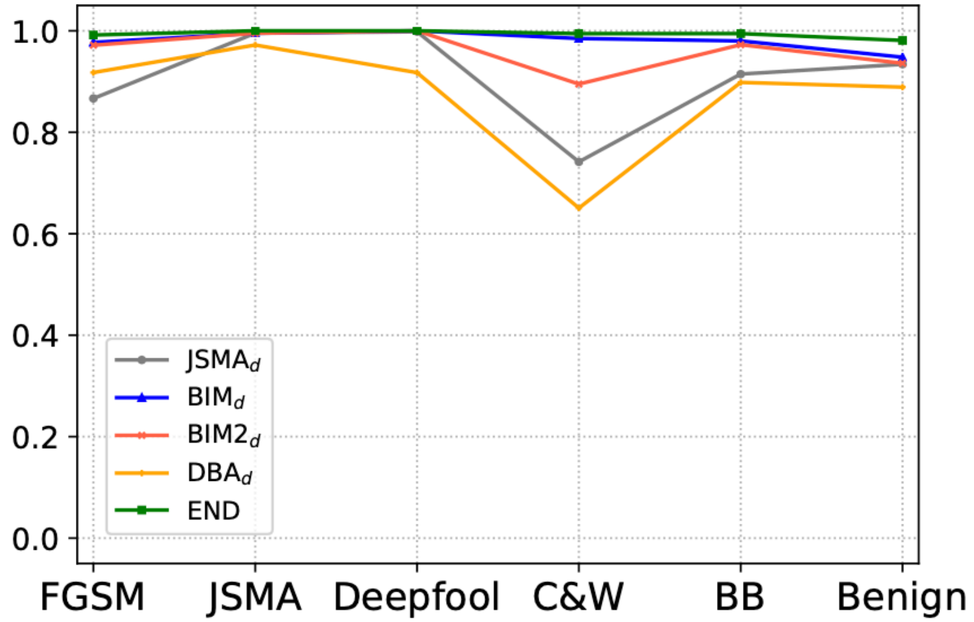


Figure. Detection accuracy, where x-axis means the class of inputs, different lines represent the detection results of different detectors

RQ3. How effective and efficient is A2D (i.e., detection)?

Some findings:

- DBA_d performs worse, but could protect the privacy of the model

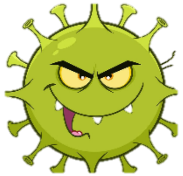
- END performs better

- Z-Score based detectors are able to achieve comparable or even better accuracy than K-NN based detectors, although Z-score based detectors only use benign examples

- For white-box attacks, attacking an adversarial examples requires only about 10 gradient queries on average

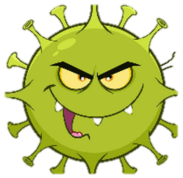
- Our detectors and corresponding parameters have good interpretability, the defenders can adjust FPR and other results according to their needs

Adaptive attack



If the attacker know the existence of ‘attack as defense’, what would they do?

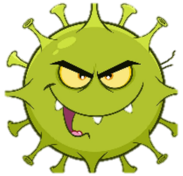
Adaptive attack



If the attacker know the existence of ‘attack as defense’, what would they do?

Encode the attack cost into the loss function?

Adaptive attack



If the attacker know the existence of 'attack as defense', what would they do?

Encode the attack cost into the loss function?

Do we have any other ways to increase the attack cost?

- Increase the confidence/strength of adversarial examples
- Initially considered by Carlini and Wagner for increasing transferability
- Confidence is controlled by the parameter κ

Reference:
Nicholas Carlini and David A.Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In Proceedings of IEEE Symposium on Security and Privacy (S&P). 39–57.

Adaptive attack

Increasing κ from 0 to 8 on MNIST:

$\kappa = 0$	CLEVER Score ≈ 0
	No. of Attack Iterations $= 1.01$

Adaptive attack

Increasing κ from 0 to 8 on MNIST:

$\kappa = 0$

CLEVER Score ≈ 0

No. of Attack Iterations $= 1.01$

$\kappa = 8$

CLEVER Score $= 0.17$

No. of Attack Iterations $= 42.59$

Adaptive attack

Increasing κ from 0 to 8 on MNIST:

$\kappa = 0$

CLEVER Score ≈ 0

No. of Attack Iterations = 1.01

$\kappa = 8$

CLEVER Score = 0.17

No. of Attack Iterations = 42.59

Does this mean that attack as defense is invalid?

$\kappa = 0$

CLEVER Score ≈ 0

No. of Attack Iterations = 1.01

L_2 distance = 1.71

$\kappa = 8$

CLEVER Score = 0.17

No. of Attack Iterations = 42.59

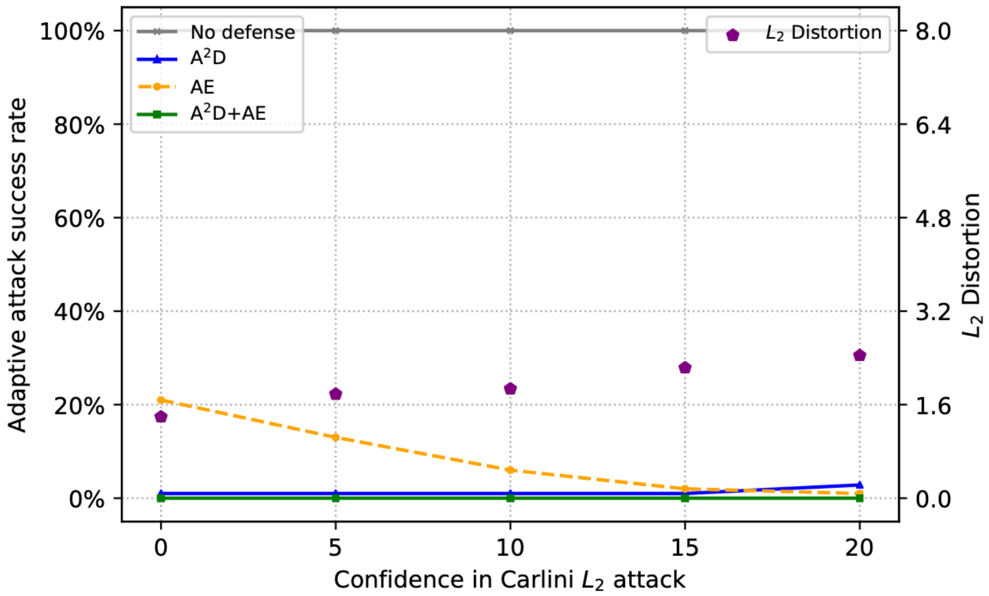
L_2 distance = 2.53

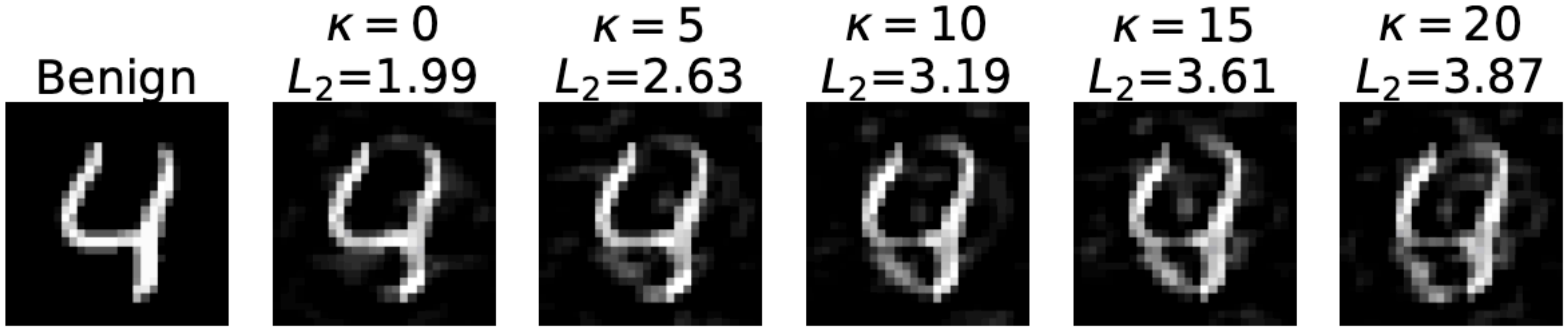
Adaptive attack

Combine A^2D with other detectors that are aimed at large distortion.

Adaptive attack

Combine A²D with other detectors that are aimed at large distortion.



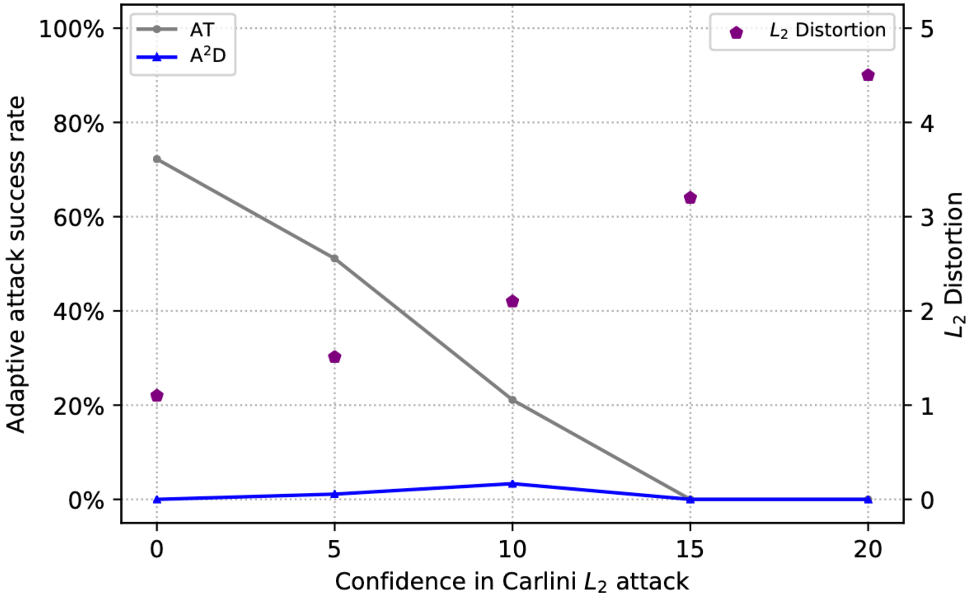


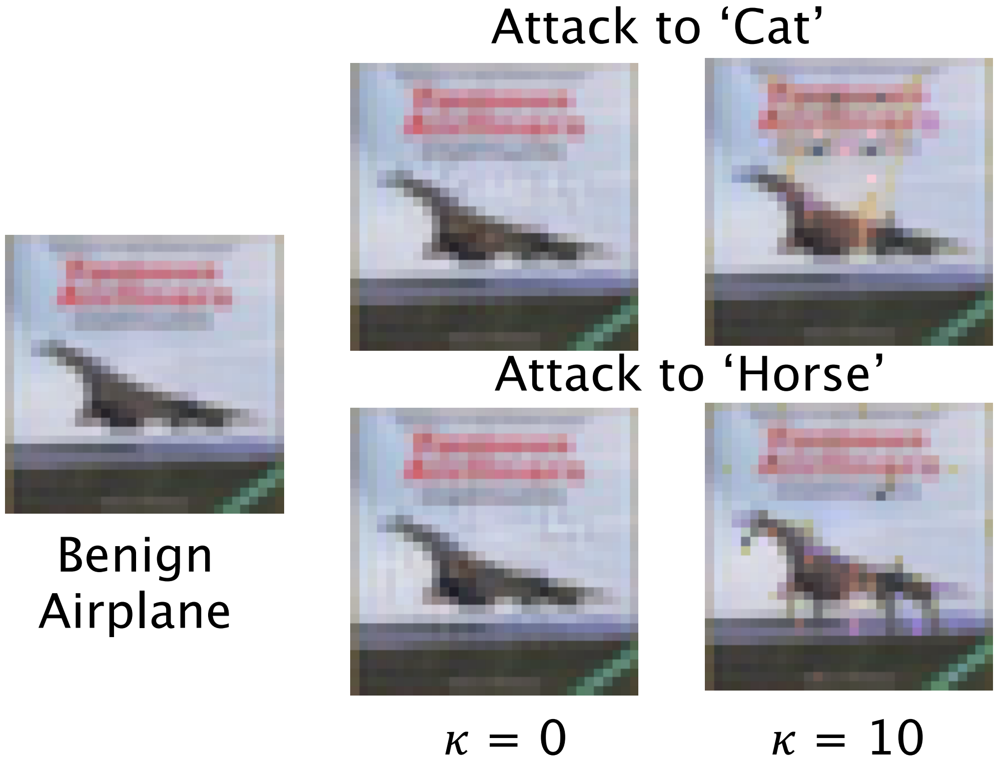
Adaptive attack

Combine with adversarial training which enhances the DL model,
so the attackers cannot generate adversarial examples with high κ easily.

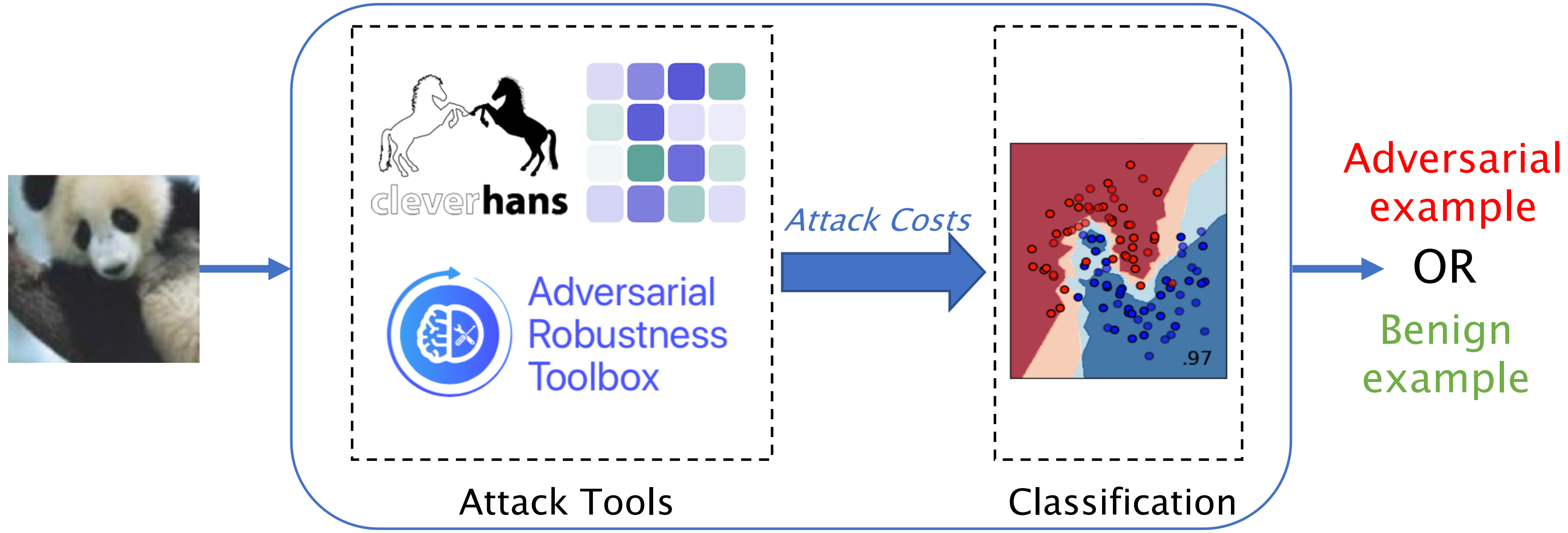
Adaptive attack

Combine with adversarial training which enhances the DL model,
so the attackers cannot generate adversarial examples with high κ easily.





Conclusion



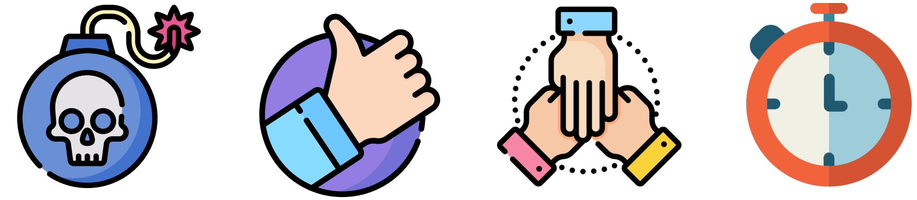
系统与软件安全实验室
System and Software Security Laboratory

System and Software Security Lab (S3L), ShanghaiTech University, Shanghai, China, http://s3l.shanghaitech.edu.cn/

S3L WeChat QR Code

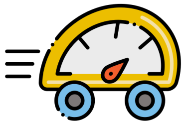
Zhe Zhao (zhaozhe1@shanghaitech.edu.cn)
Fu Song (songfu@shanghaitech.edu.cn)

Icon made by Freepik from www.flaticon.com

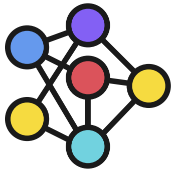


Icon made by Eucalyp from www.flaticon.com

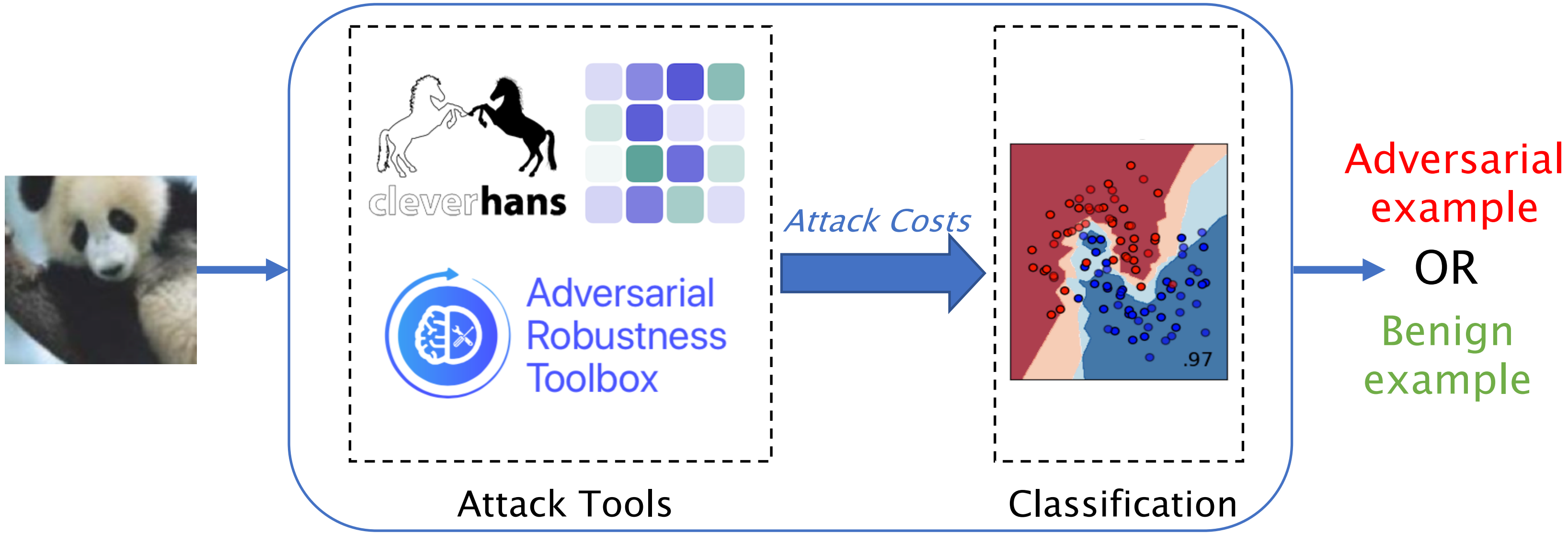
Icon made by Flat Icons from www.flaticon.com



Icon made by Becris from www.flaticon.com



Conclusion



S3Lab
系 统 与 软 件 安 全 实 验 室
System and Software Security Laboratory

System and Software Security Lab (S3L), ShanghaiTech University, Shanghai, China, http://s3l.shanghaitech.edu.cn/

S3L WeChat QR Code

Zhe Zhao (zhaozhe1@shanghaitech.edu.cn)
Fu Song (songfu@shanghaitech.edu.cn)