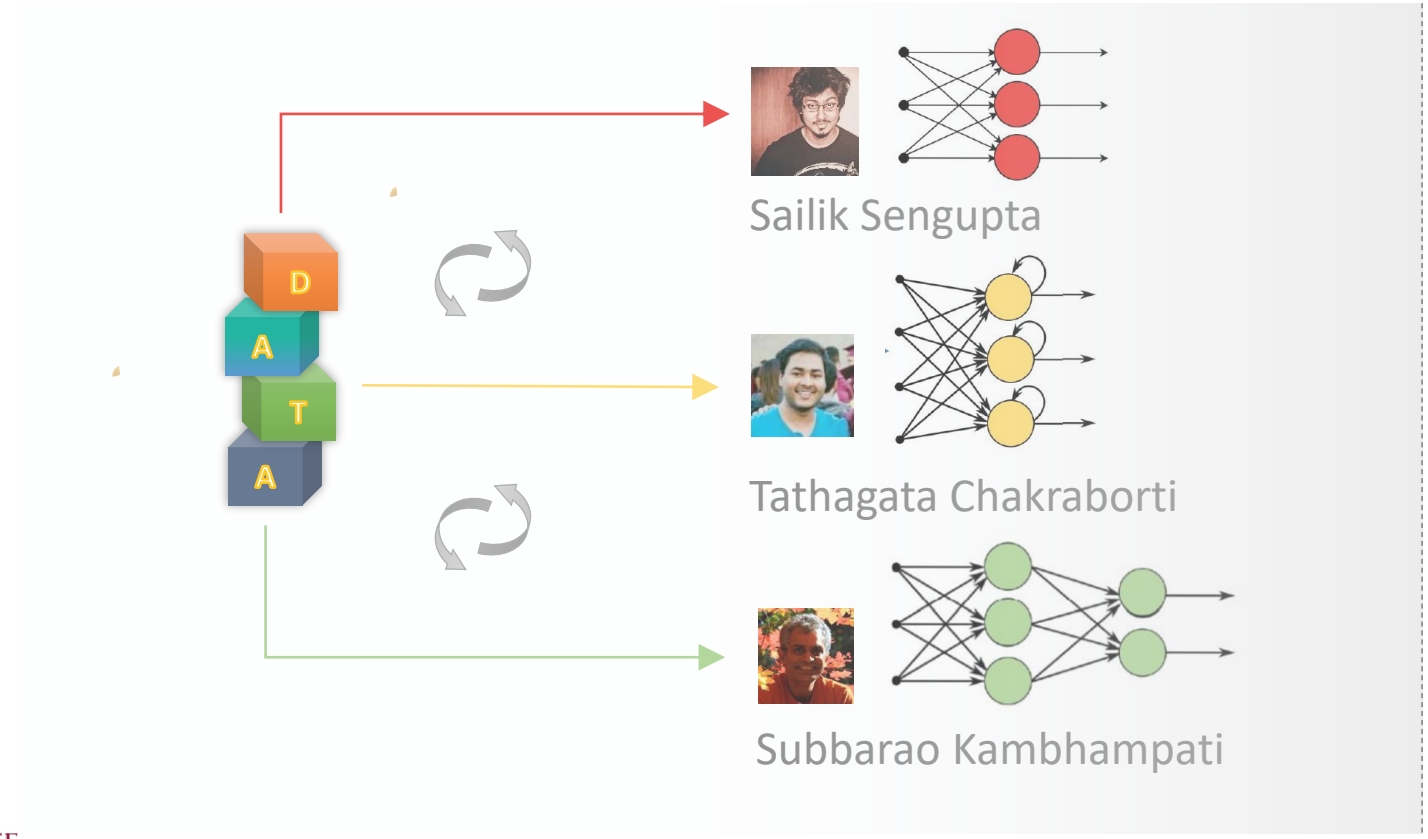
MTDeep: Moving Target Defense to Boost the Security of Deep Neural Networks Against Adversarial Attacks



Decision-time(/Test-time) Attacks

- If noise ϵ is strategically generated to intentionally make a classifier misclassify the modified input $i + \epsilon$, we consider these as attacks.
- Too much noise in input data can be easily detected (by humans). Thus, one has to ensure that ϵ is minimum (minimize $|\epsilon|_p$).

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*.

Picture courtesy:
Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*(2013).

Defense Techniques

- Train the neural net on the attack distribution (with the correct labels) and the classifier becomes immune to the particular type of adversarial inputs. This has been shown to be one of the most effective methods:

$$\min_{\theta} \max_{\epsilon} \alpha\, L(\theta, x+\epsilon, y) + (1-\alpha)\, L(\theta, x, y)$$ Madry et. al. 2018

  - Ensemble adversarial training – Use the constituent networks of an ensemble to get more adv. Input images. Use it to strengthen ***a*** network. Tramer et al., 2017

- Defenses may not be be effective against other attack methods eg. Adv. Universal Perturbation Moosavi et al., 2016

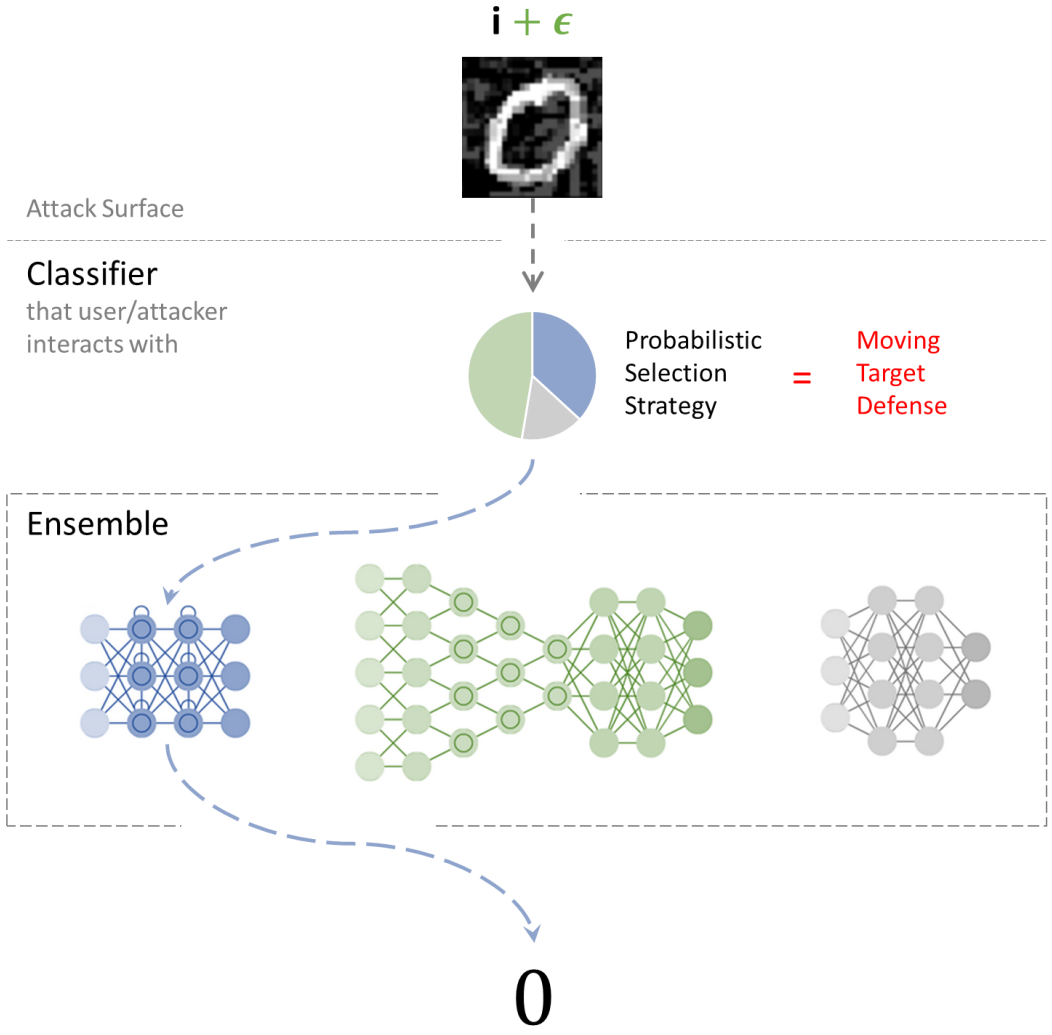
Moving Target Defense for Deep Neural Networks (MTDeep)

- Moving Target Defense
  - Keep shifting; the attacker's attack designed for a particular configuration does not work.

Moving Target Defense for Deep Neural Networks (MTDeep)

- Moving Target Defense
  - Keep shifting; the attacker's attack designed for a particular configuration does not work.

- Assumes an attack designed for one classifier does not work well on the other.
- Starting with a set of classifiers, (how to?) randomly choose between them at test-time.



Can one attack kill all?

- Differential immunity (δ)

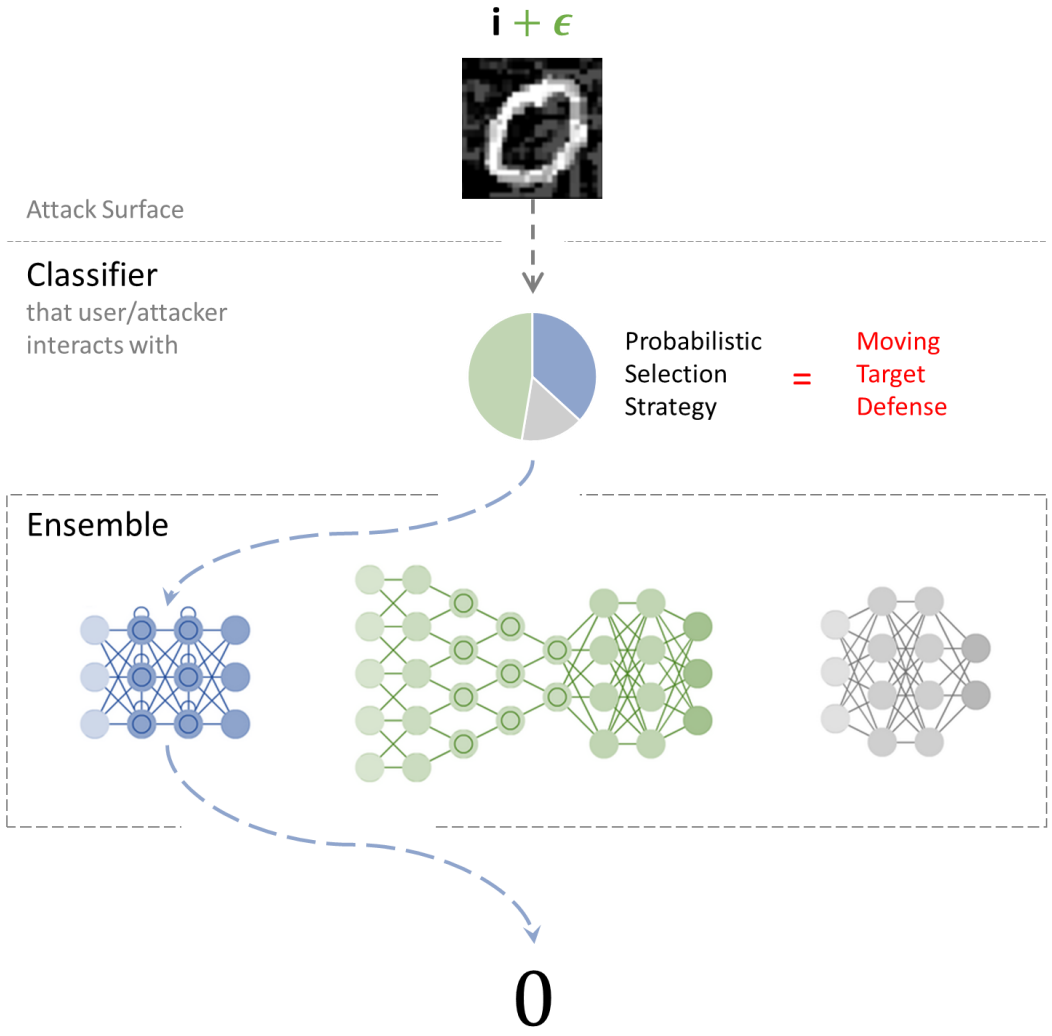
  Given a set of classifiers N and a set of attacks U, metric to measure *worst-case transferability of an ensemble*.

$$\delta(U, N) = \min_{u} \frac{\max_n E(n, u) - \min_n E(n, u) + 1}{\max_n E(n, u) + 1}$$

$$0 \leq \delta \leq 1$$

No differential immunity

High differential immunity



How to pick a classifier?

Let's play a game!

$$\min_{\theta} \max_{\epsilon} L(\theta, x + \epsilon, y)$$

$$\min_{\theta} \max_{\epsilon} \alpha L(\theta, x + \epsilon, y) + (1 - \alpha) L(\theta, x, y)$$

$$\min_{p} \max_{\epsilon} \sum_{i} p_i [\alpha L(\theta_i, x + \epsilon, y) + (1 - \alpha) L(\theta_i, x, y)]$$

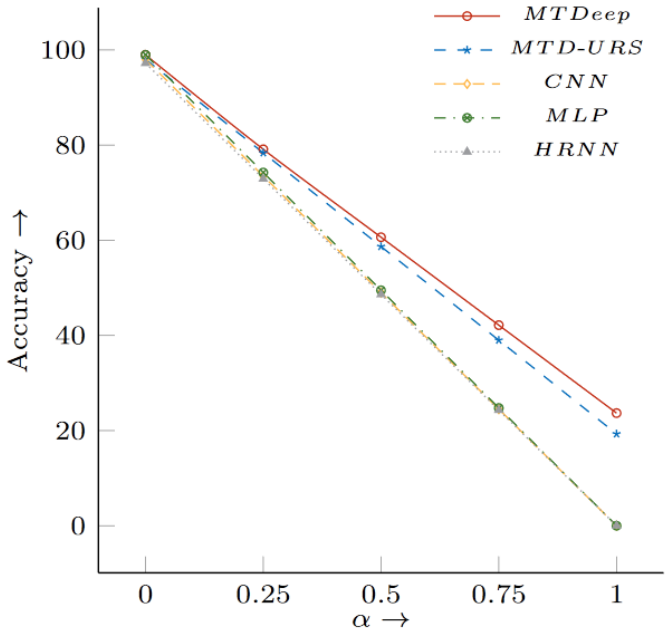
Stackelberg equilibrium now solves the multi-objective function.
Increases defender's utility, which
☺ Reduce misclassification rate on adversarial inputs.
☺ Limit the reduction in classifiaction accuracy on legitimate samples.

	Legitimate User ($\mathcal{L}$)	Adversarial User ($\mathcal{A}$)							
MTDeep	Classification Image	FGM_m	FGM_c	FGM_h	DF_m	DF_c	DF_h	PGD_m	PGD_c	PGD_h
MLP	99.1	3.1	20.39	38.93	1.54	89.8	93.83	0.00	49.00	61.00
CNN	98.3	55.06	10.28	71.39	98.87	0.87	98.55	78.00	0.00	90.0
HRNN	98.7	25.12	27.24	11.43	95.38	83.17	3.66	23.00	51.00	0.00

(a) MNIST

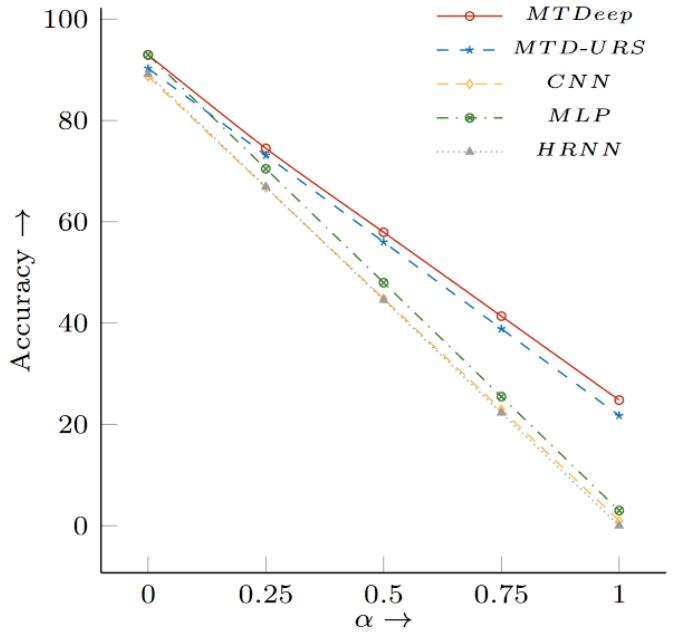
How well does this work?

+23.68



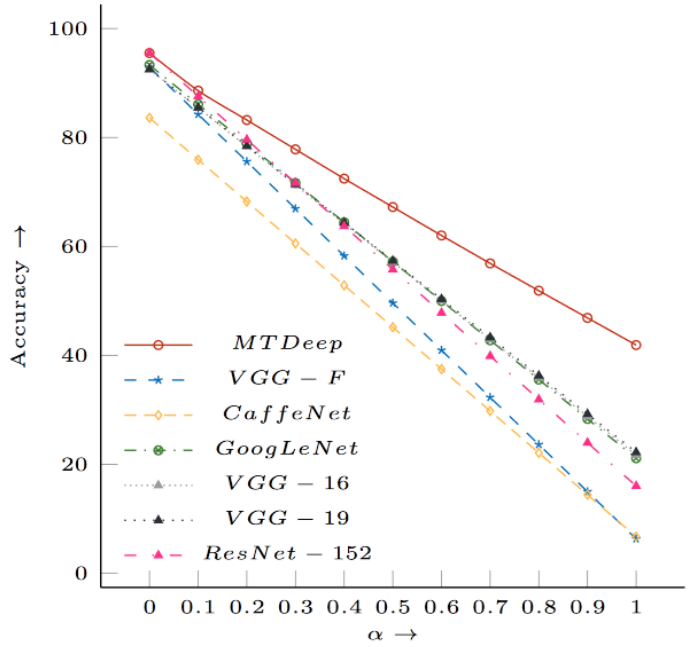
MNIST
(FGM, DF, PGD)

+21.8



F-MNIST
(FGM, DF, PGD)

+20.68



ImageNET
(Universal Perturbations)

MTDeep can be an Add-on Defense too!

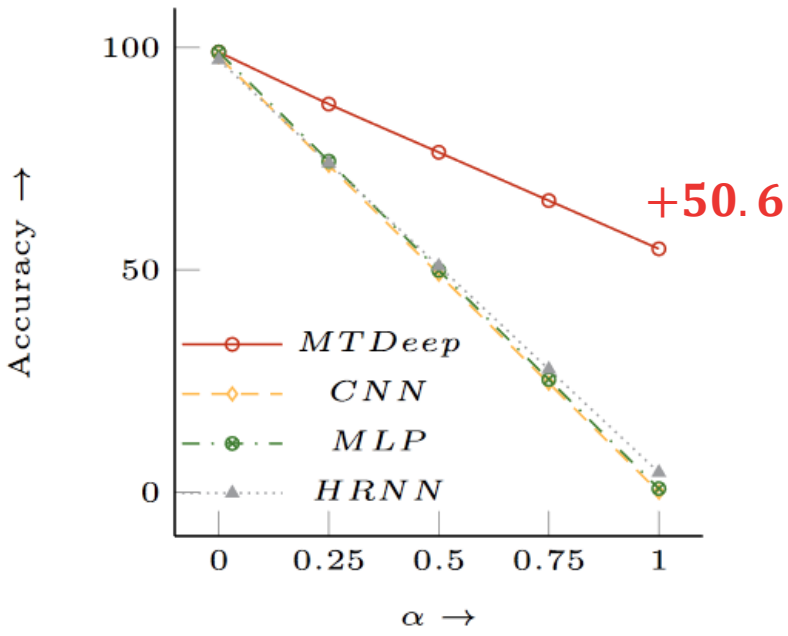
- MTDeep can be used on top of any existing defense mechanisms.
  - We strengthen the constituent networks with Ensemble Adversarial Training.

Legitimate User ($\mathcal{L}$)

MTDeep	Classification Image
MLP_{eat}	97.99
CNN_{eat}	98.97
$HRNN_{eat}$	97.22

Adversarial User ($\mathcal{A}$)

FGM_m	FGM_c	FGM_h	DF_m	DF_c	DF_h	PGD_m	PGD_c	PGD_h
95.06	75.32	70.1	1.5	96.97	95.73	0.00	88.00	69.00
61.44	96.55	68.58	98.36	0.79	96.09	72.00	20.00	81.00
81.24	84.79	93.1	96.85	95.9	4.41	82.00	71.00	10.00



MNIST EAT Classifiers
(FGM, DF, PGD)

- We notice that Ensemble Adversarial Training (EAT) increases the differential immunity of the ensemble against the 3 attacks mentioned! (Why?)

Crafting attacks for an ensemble!

- Black-box attack Papernot et al., 2016
  - Design a distilled network that can capture the behavior of the MTDeep ensemble.
  - Design attacks on the distilled network.
  - See if the attack transfers to the MTDeep ensemble.

- Optimal white-box attacks are stronger than the black-box attack (MTDeep has 8% higher accuracy against black-box attacks compared to the optimal white-box attack for MNIST).

Differential Immunity and Performance

- Higher differential immunity (δ) yields higher gains in security.

Networks	Differential Immunity (δ)	Accuracy of Best Constituent Net	Accuracy of MTDeep	Gain
FashionMNIST	0.11	3%	24.8%	21.8%
MNIST	0.19	0%	23.68%	23.68%
ImageNET	0.34	22.2%	42.88%	20.68%
MNIST + EAT	0.78	4.41%	54.71%	50.3%

Note that ImageNET has 1000 classes compared to 10 classes MNIST or Fashion-MNIST.

Conclusions and Future Work

- MTDeep: Moving Target Defense for Deep neural networks increases the an ensemble's robustness to decision-time adversarial attacks.

How to obtain a good set of θ_is? Adam et al 2018

$$\min_{p} \max_{\epsilon} \sum_{i} p_i[\alpha L(\theta_i, x+\epsilon, y) + (1-\alpha) L(\theta_i, x, y)]$$

$$\min_{p,\theta} \max_{\epsilon} \sum_{i} p_i[\alpha L(\theta_i, x+\epsilon, y) + (1-\alpha) L(\theta_i, x, y)]$$

- Needs to consider differential immunity during training.
- Jointly optimize for policy and obtaining differentially immune networks?

sailiks@asu.edu
https://sailik1991.github.io/