MTDeep: Boosting the Security of Deep Neural Nets Against Adversarial Attacks with Moving Target Defense

Sailik Sengupta • Subbarao Kambhampati

Tathagata Chakraborti

Arizona State University

IBM **Research** AI

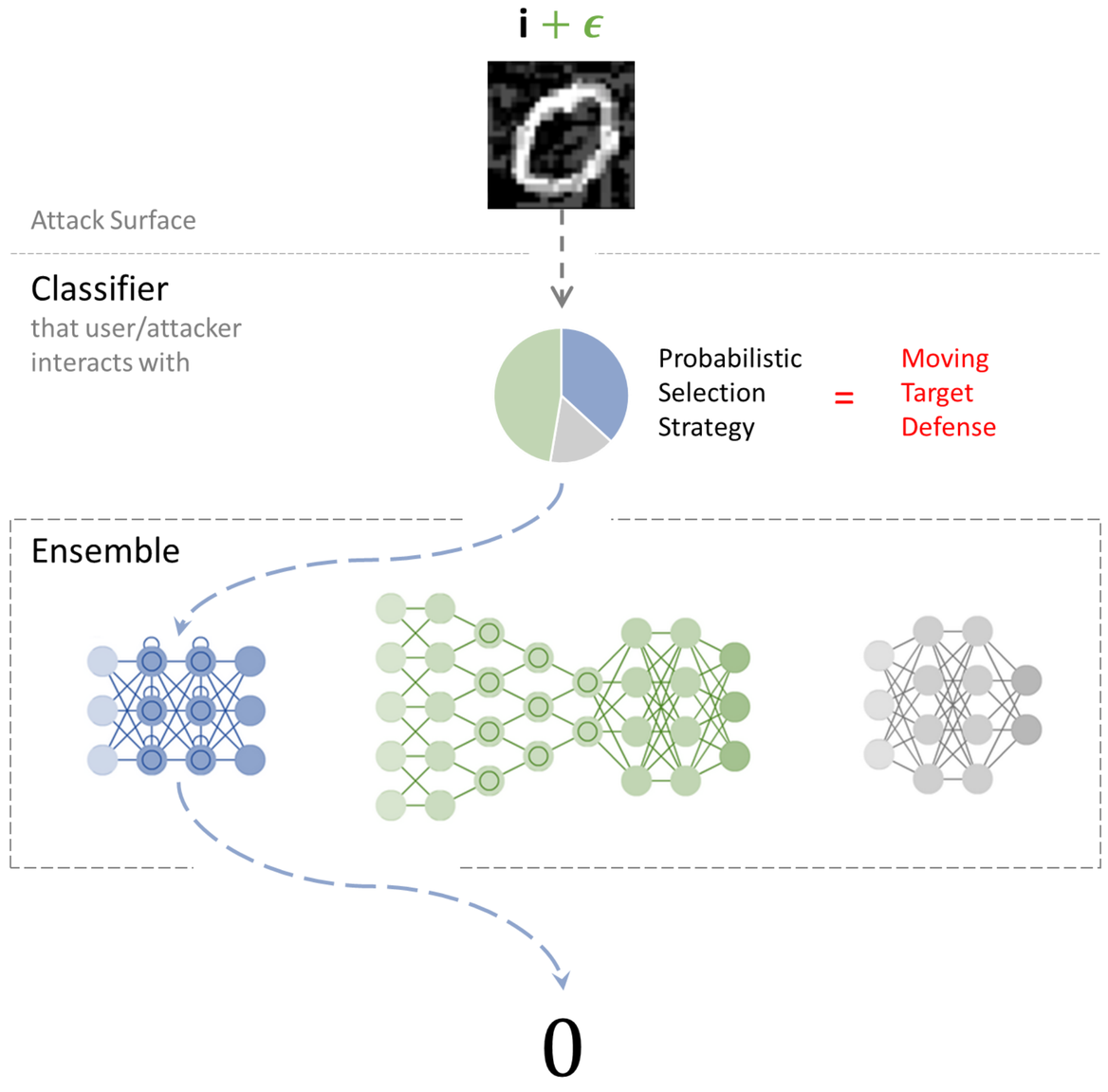
GameSec 2019

In **Moving Target Defense**, an agent defends a system by randomly switching between a set of system configurations in order to take away the adversary's advantage of reconnaissance.

Given an ensemble of classifiers, choose a network strategically at classification time.

This improves robustness against adversarial inputs while ensuring high accuracy on legitimate data.

We investigate the notion of *differential immunity* that allows ensembles to conceive such defense mechanisms.



Selection Strategy

We model the interaction between the classifier and the users (both legitimate and adversarial) as a two-player Bayesian Game.

Legitimate User ($\mathcal{L}$)

MTDeep	Classification Image
MLP	99.1
CNN	98.3
HRNN	98.7

Adversarial User ($\mathcal{A}$)

FGM_m	FGM_c	FGM_h	DF_m	DF_c	DF_h	PGD_m	PGD_c	PGD_h
3.1	20.39	38.93	1.54	89.8	93.83	0.00	49.00	61.00
55.06	10.28	71.39	98.87	0.87	98.55	78.00	0.00	90.0
25.12	27.24	11.43	95.38	83.17	3.66	23.00	51.00	0.00

Normal form illustrating the payoff matrix. Utilities of the defender in this constant sum game are proportional to the accuracy of the networks.

Mixed strategy at Stackelberg equilibrium improves robustness while not substantially loosing out on accuracy.

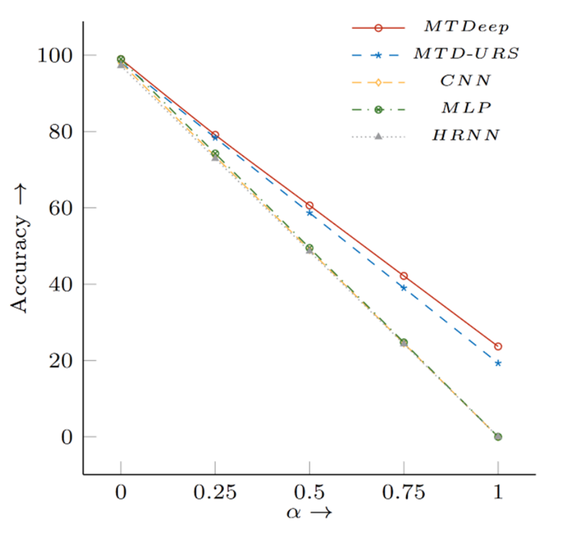
$$\max_{x,q} \sum_{n \in N} (\alpha \cdot \sum_{u \in U} R^D_{n,u}\ x_n q^{\mathcal{A}}_u\ +\ (1-\alpha) \cdot R^D_{n,u}\ x_n q^{\mathcal{L}}_u)$$

The switching strategy is better than uniform random which weighs less accurate and vulnerable configurations equally.

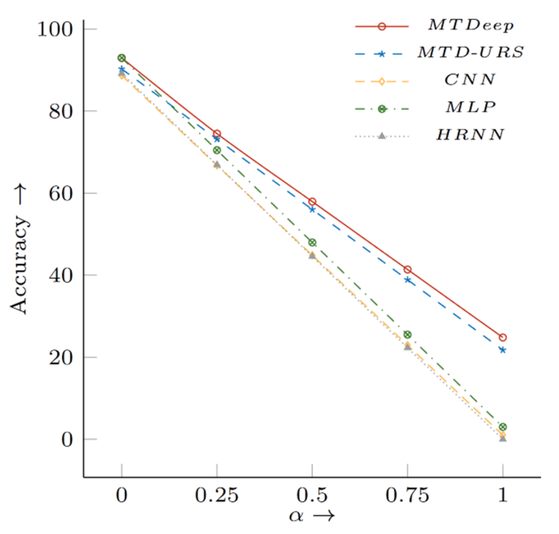
We notice impressive gains across various datasets.

Works even better when coupled with existing defense techniques like Ensemble Adversarial Training (EAT).

Due to the randomization in the selection strategy, black box attacks are also less effective against MTDeep as a whole than against the individual networks.



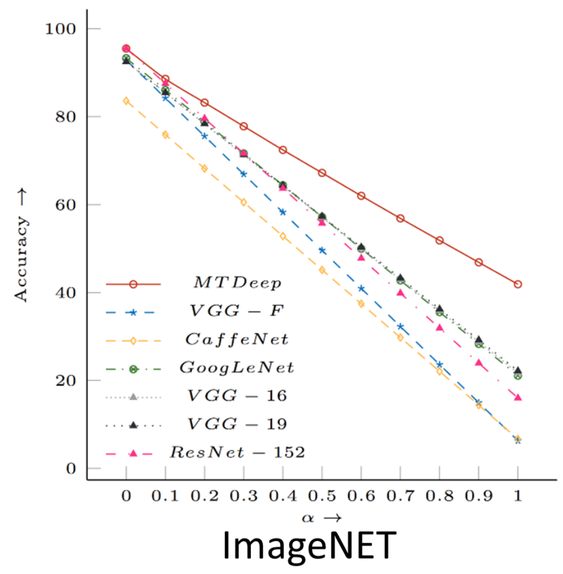
MNIST



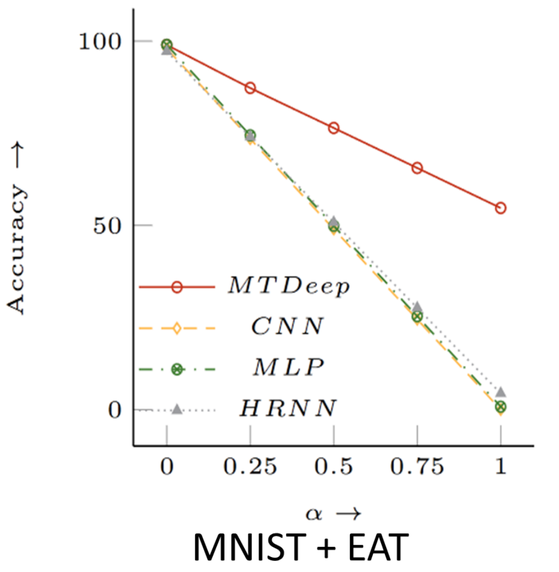
Fashion MNIST

Low Transferability of attacks
= High differential immunity
= More gains!

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%



ImageNET



MNIST + EAT