UNIVERSITY OF MICHIGAN

Morphence

Moving Target Defense Against Adversarial Examples

Abderrahmen Amich and Birhanu Eshete

Adversarial Examples

Input

Picture of a Cat

Adv. crafting

Careful perturbations of the input

ML model

A highly accurate Cats vs Dogs classifier

Output

Wrong prediction

Important Milestones

2014
Gradient-based Attacks
FGSM[1], BIM[2] and PGDM[3]

2016
Defences against Gradient Attacks
Defensive Distillation[4], Gradient Masking, etc

2017
Carlini & Wagner
C&W[5]

2017
Adversarial training
Adv training[6]

2019
Certified Defenses
Certified Defenses[7], PixelDP[8]

2021
Morphence (this work)
Moving Target Defense

[1] https://arxiv.org/abs/1412.6572
[2] https://arxiv.org/pdf/1611.01236.pdf
[3] https://arxiv.org/pdf/1706.06083.pdf
[4] https://arxiv.org/abs/1511.04508
[5] https://arxiv.org/abs/1608.04644
[6] https://arxiv.org/abs/1611.01236
[7] https://arxiv.org/abs/1705.07204
[8] https://arxiv.org/abs/1801.09344
[9] https://arxiv.org/abs/1802.03471

Why Moving Target Defense?

Fixed Model

- Highly vulnerable to model approximation
- Given enough time, the adversary will eventually find a way to evade it
- Repeated Attack: Once successful, it is always successful

Moving Model

- Fitting the target model could be harder
- The defender is always one step ahead
- An attack can succeed only once

A Moving Model

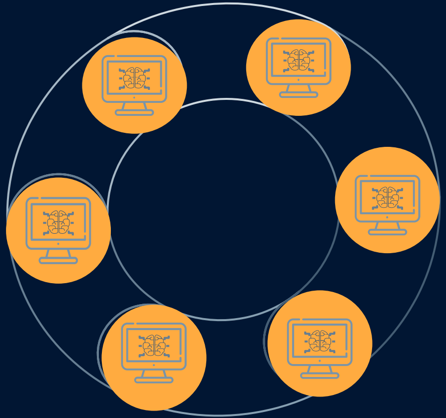
Input

Picture of a Cat

Adv. crafting

Careful perturbations of the input

n = 6

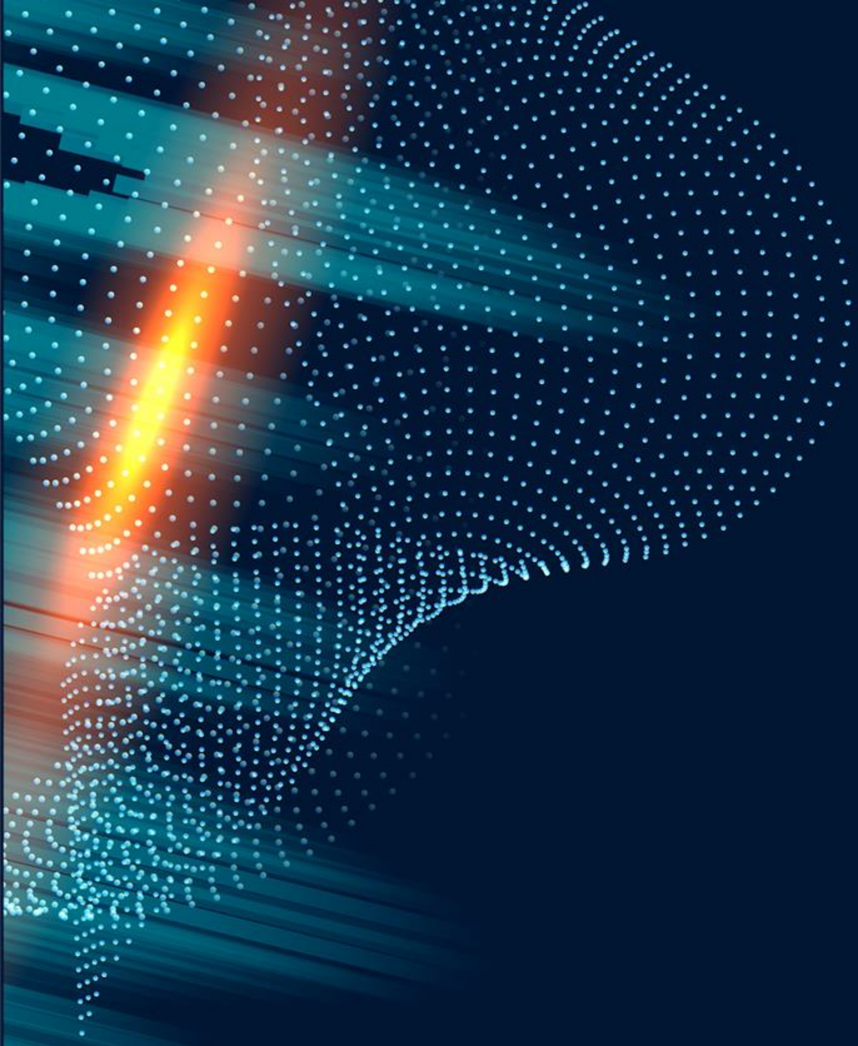


Moving Model

A highly accurate Cats vs Dogs classifier

Output

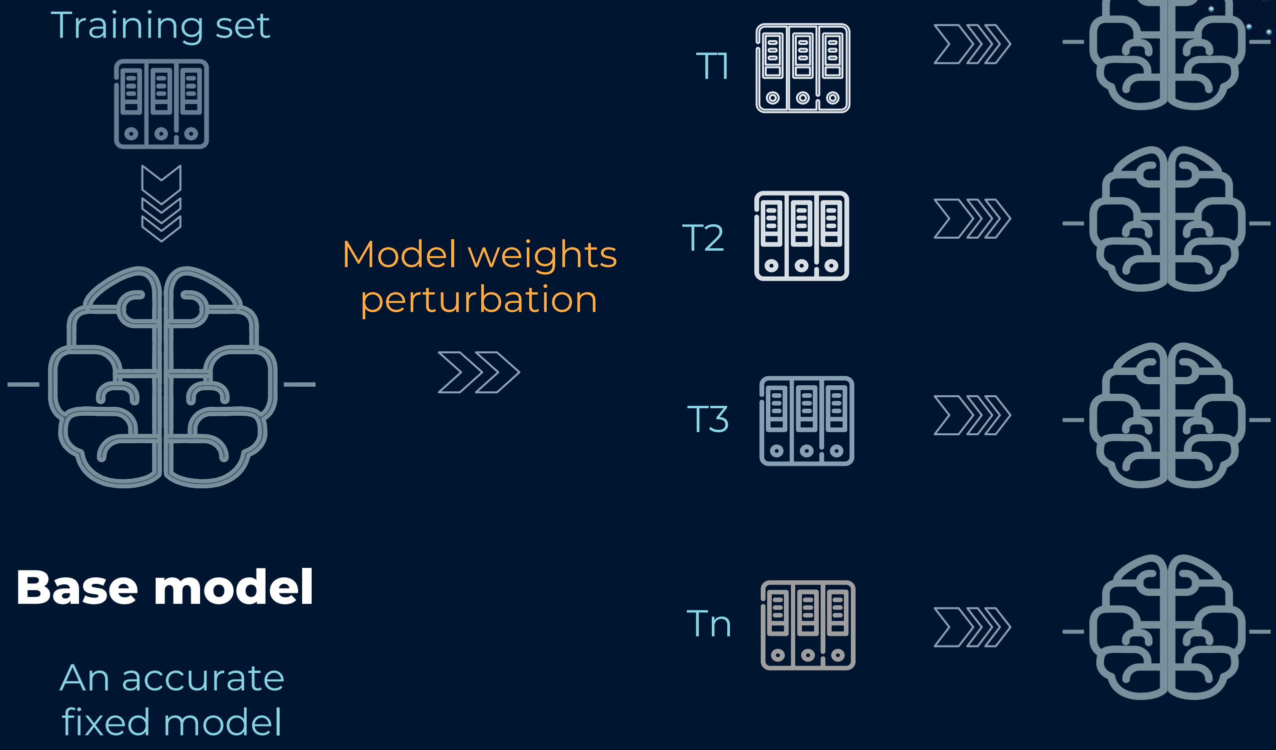
Correct prediction



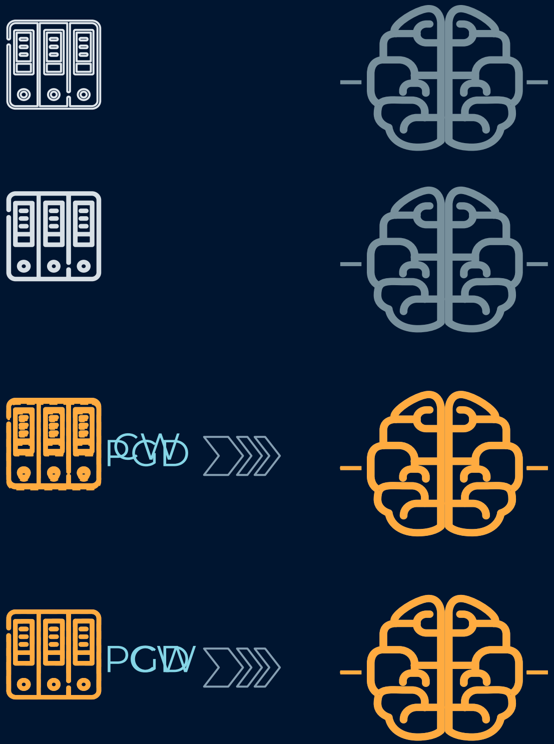
Morphence

Towards Moving Target Defenses against Adversarial Examples ...

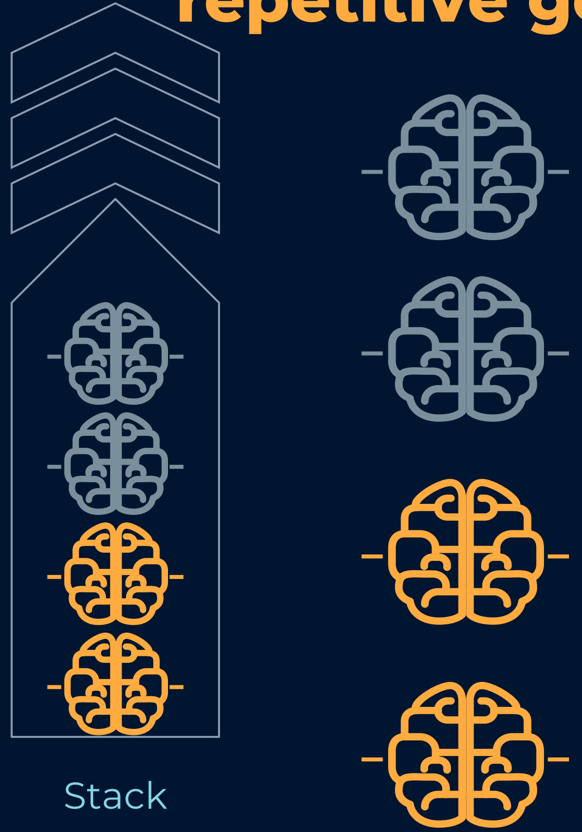
Model Pool Generation

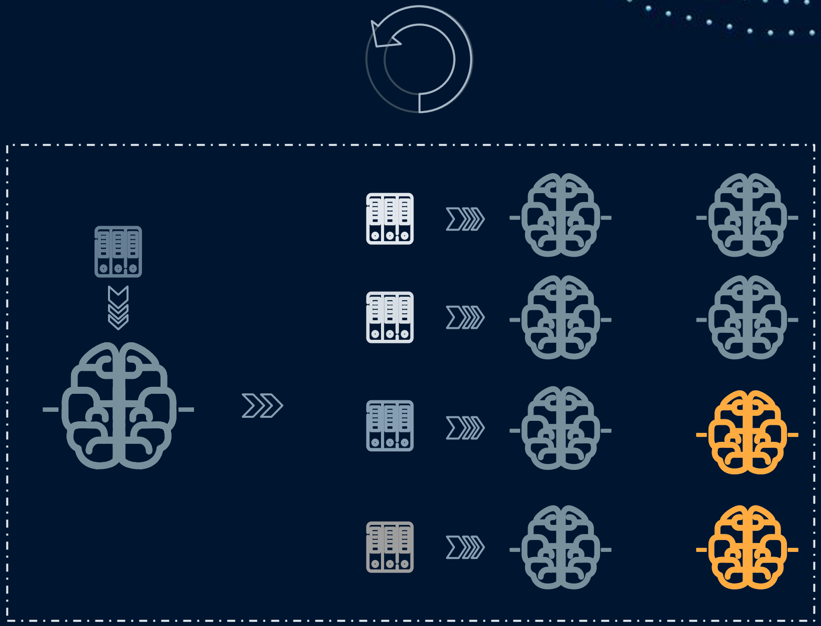


Another Layer of Robustness

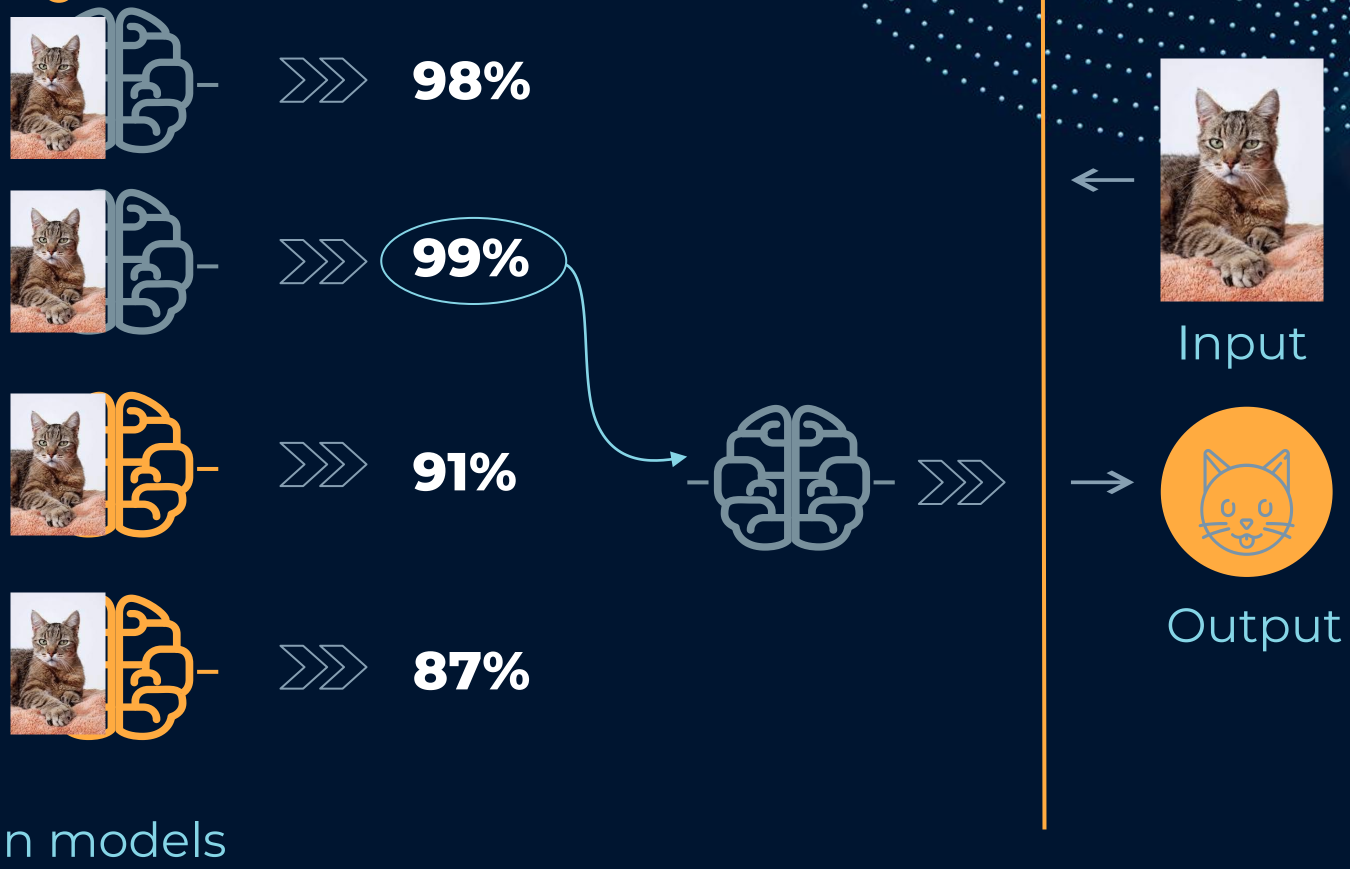


Local storage and repetitive generation

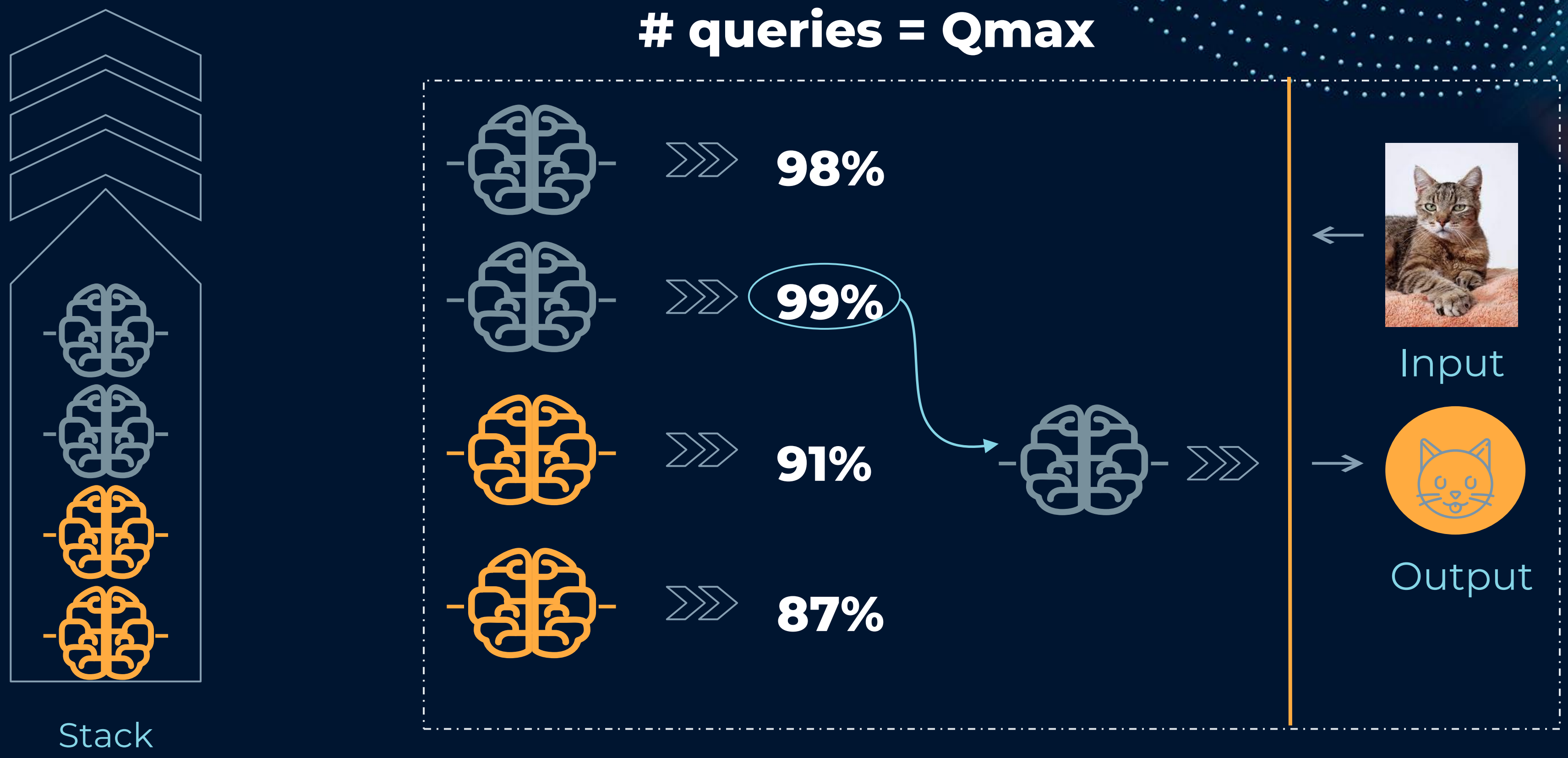




Deployment



Model Pool Renewal



Results on MNIST

	Undefended	Adversarially-trained	*Morphence*
No Attack	99.72%	97.17%	**99.04%**
FGSM	9.98%	42.38%	**71.43%**
C&W	0.0%	0.0%	**97.75%**
SPSA [10]	29.04%	59.43%	**97.77%**

Do not sacrifice accuracy on benign data

Significant increase compared to adv training

Overcomes C&W

Robust against iterative-query attacks

[10] https://arxiv.org/abs/1802.05666

Results on CIFAR10

	Undefended	Adversarially-trained	*Morphence*
No Attack	83.63%	75.37%	**84.64%**
FGSM	9.98%	36.62%	**38.78%**
C&W	1.25%	1.34%	**44.50%**
SPSA	38.96%	59.43%	**62.83%**

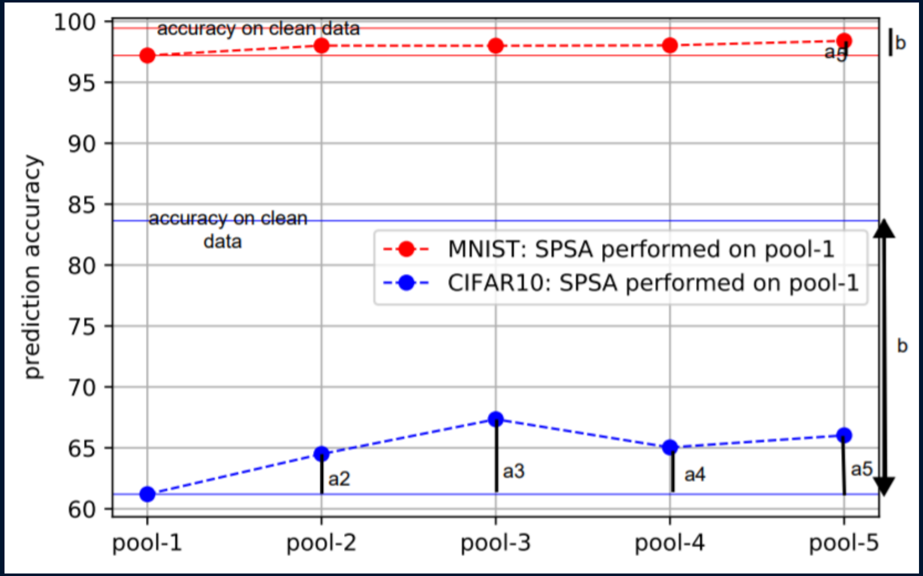
Can improve accuracy on benign data

Improvement compared to adv training

Significant improvement on C&W

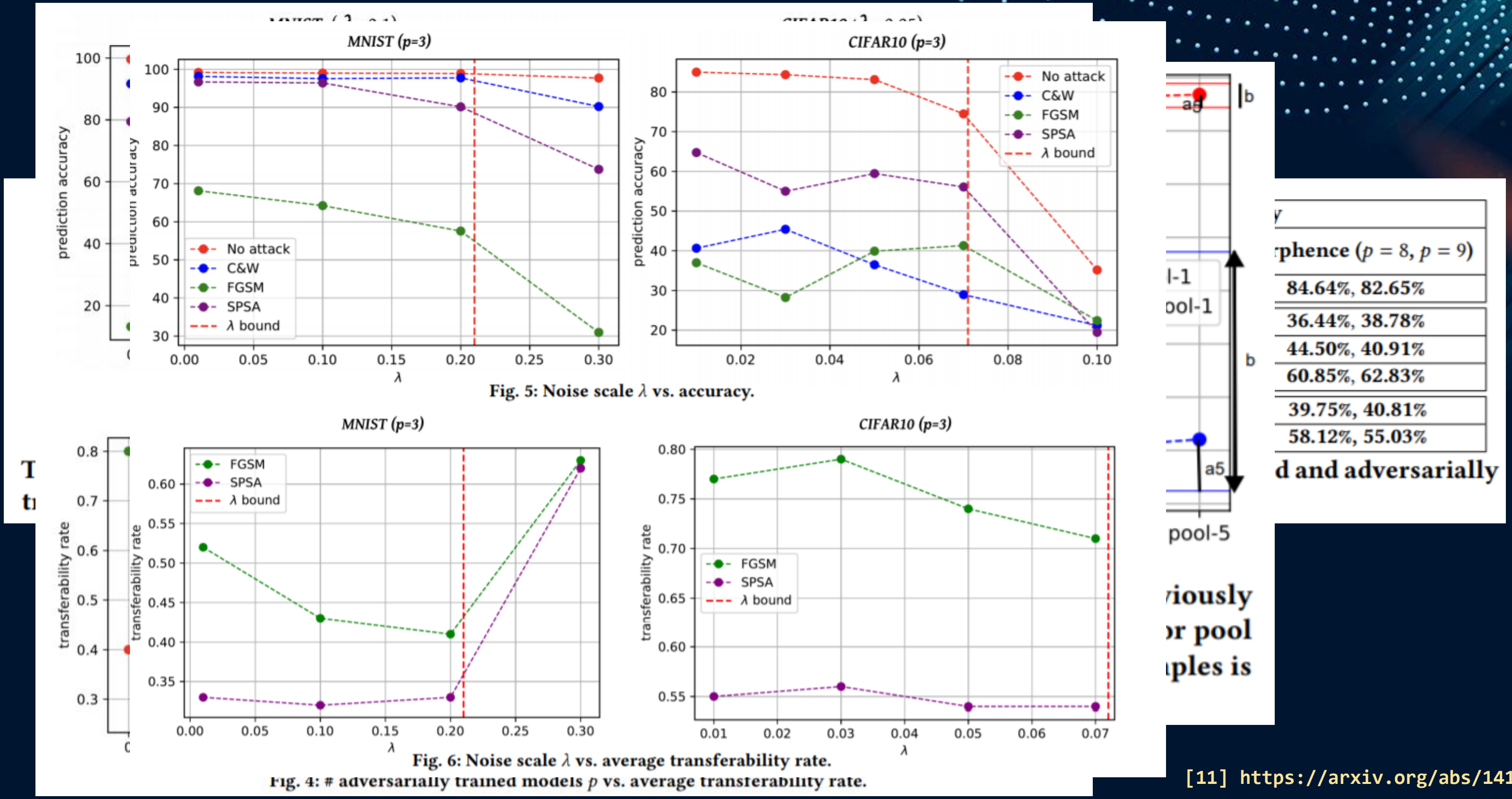
Higher robustness against iterative-query attacks

Robustness Against Repeated Attacks



Does successful attacks on pool-1 remain evasive on different pool of models?

Detailed Results in the paper



[11] https://arxiv.org/abs/1412.6572

Conclusions

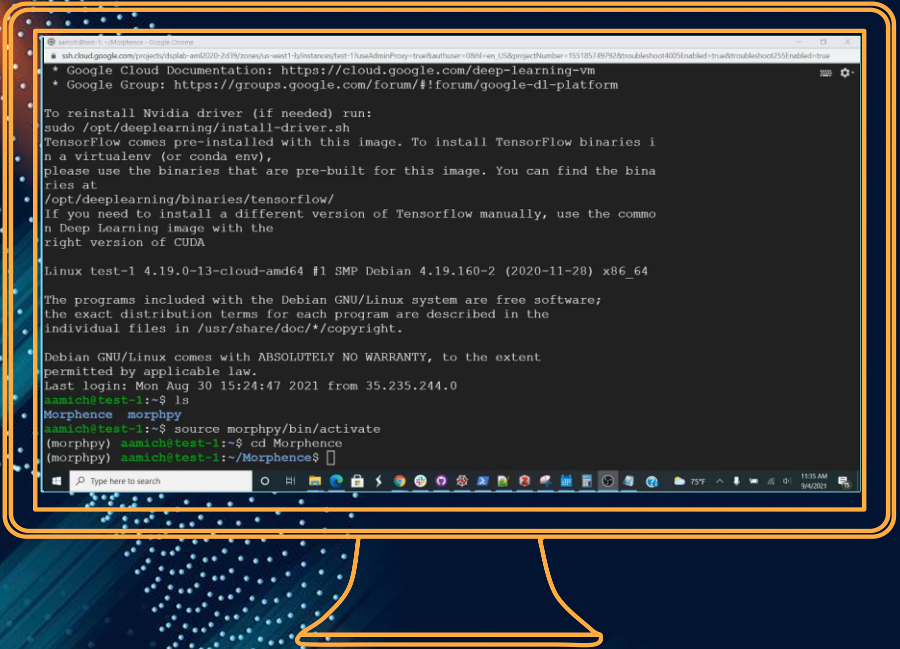
Morphence

Moving Target Defense Against Adversarial Examples

- A Moving target model is more robust than the best fixed model defense.
- A Moving target model can prevent falling to the same attack multiple times.
- Iteratively querying a moving target model is not effective to optimize adversarial perturbations.
- We hope that Morphence will be used as a new benchmark for robustness against evasion attacks

Available Artifact

https://github.com/um-dsp/Morphence



THANKS!

Do you have any questions?

aamich@umich.edu

@AbderrahmenAmi2

https://abderrahmen-amich.netlify.app/