Intelligent System Security: Adversarial Attacks and Defense

Boqi (Percy) Chen

What is in the Image?

You have the latest self-driving car with a road sign detector that can easily detect stop signs and keep you in safe

Your Car:
"stop sign"
99.85% Confidence

But one day, someone hacked your car's camera system. And the image from the camera has changed a bit

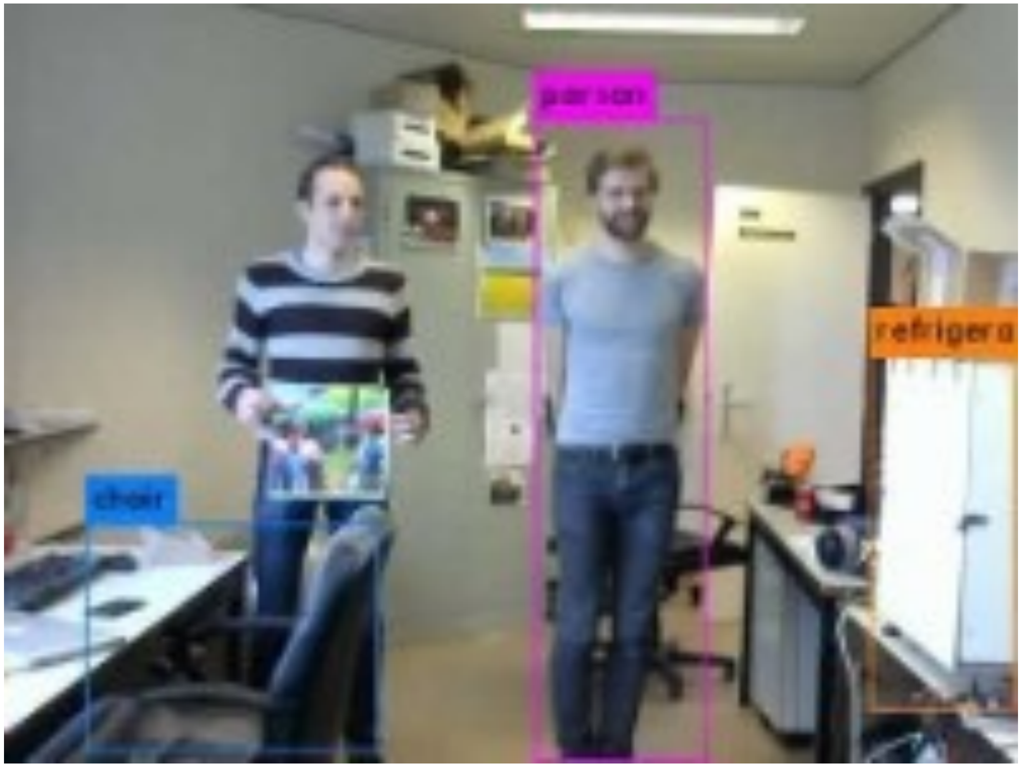
You: "Is the hacker stupid? This is still a stop sign!"

Your Car:
"120 km/hr"
99.90% Confidence

https://kennysong.github.io/adversarial.js/

Adversarial Examples in Physical World

Your car is also equipped with a perfect pedestrian (person) detector.
Until one day, there is a pedestrian carrying a weird painting…

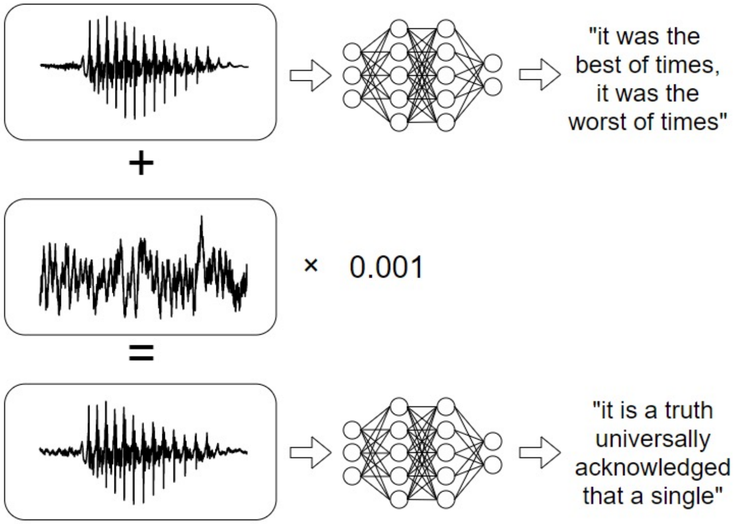


https://youtu.be/MIbFvK2S9g8?t=46

Other domains?

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: Positive (77%)
Adversarial example [Visually similar]	Aonnoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: Negative (52%)
Adversarial example [Semantically similar]	Connoisseurs of Chinese footage will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: Negative (54%)

Adversarial examples for NLP



Adversarial examples for Sound

Adversarial Attacks

Above are adversarial examples

- Adversarial examples refers to samples that are perceptually indifferentiable from the correct samples but causes the machine learning model making wrong prediction

- In traditional machine learning, adversarial examples main refer to the out-of-distribution samples (i.e. violate the statistical assumption)
  - E.g. spammers insert many "good words" into the spam emails

- In 2013, Szegedy et al. find that neural networks are particular venerable to a classes of data perturbation even on the training data

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

Threats of Adversarial Attacks

- Small perturbation using adversarial methods are more likely to be misclassified than larger random noise
- Adversarial examples that were designed to be misclassified by a model *M1* is often also misclassified by a model *M2*
  - Black-box attack is possible
- Reduces the **reliability** of intelligent systems with ML components

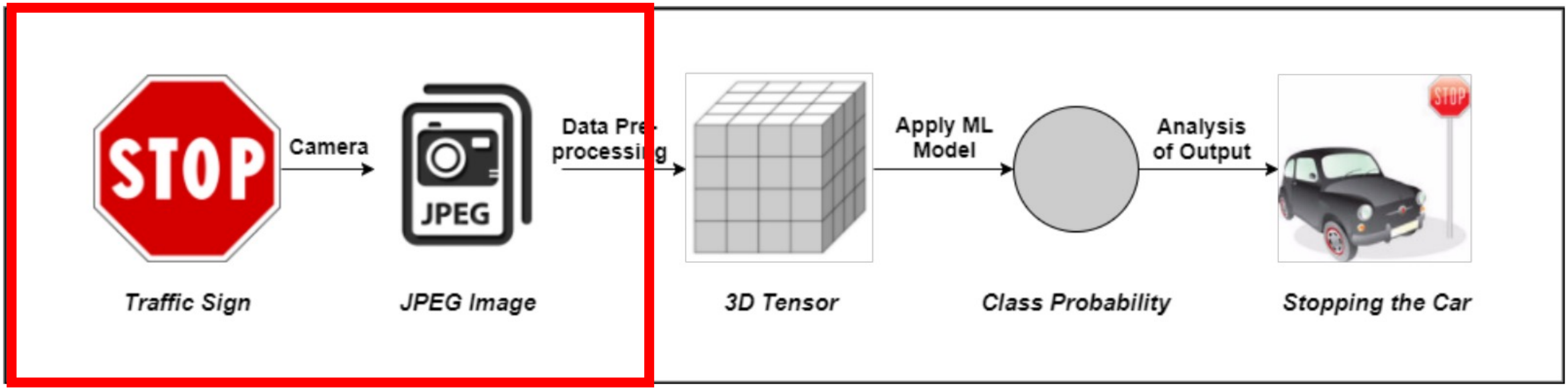
BRIAN BARRETT SECURITY 02.22.2020 09:00 AM

Security News This Week: A Tiny Piece of Tape Tricked Teslas Into Speeding Up 50 MPH

An MGM Resorts breach, natural gas ransomware, and more of the week's top security news.

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

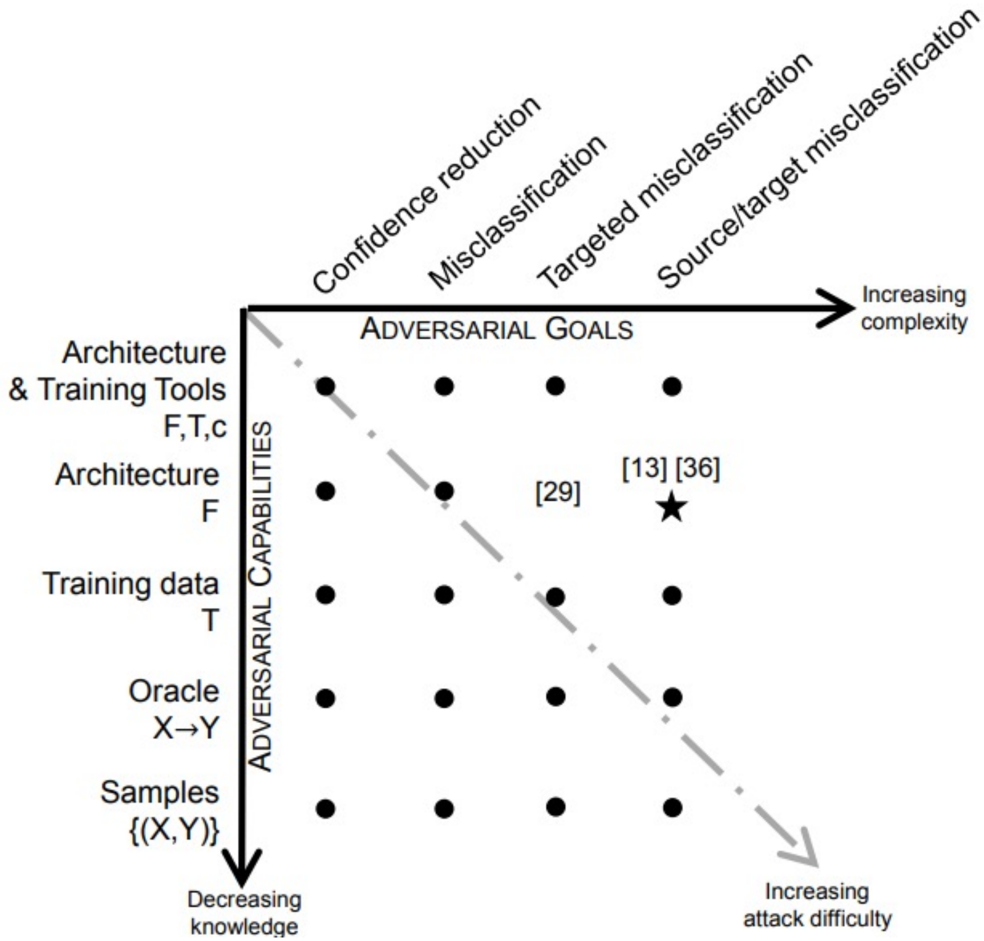
The Threat Model: Target Process



Main Target of Adversarial Attack

Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." *2016 EuroS&P*. IEEE, 2016.

The Threat Model: Adversarial Ability



Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." *2016 EuroS&P*. IEEE, 2016.

How Adversarial Attack Works (1)

- Assume we want to produce an adversarial example for an image
  - Image: x
  - Model*:* M_θ, Loss Function: $\mathrm{l}oss$
  - Correct prediction: $M(x) = l$
- We want to apply some perturbation r, such that $M(x+r) \neq l$

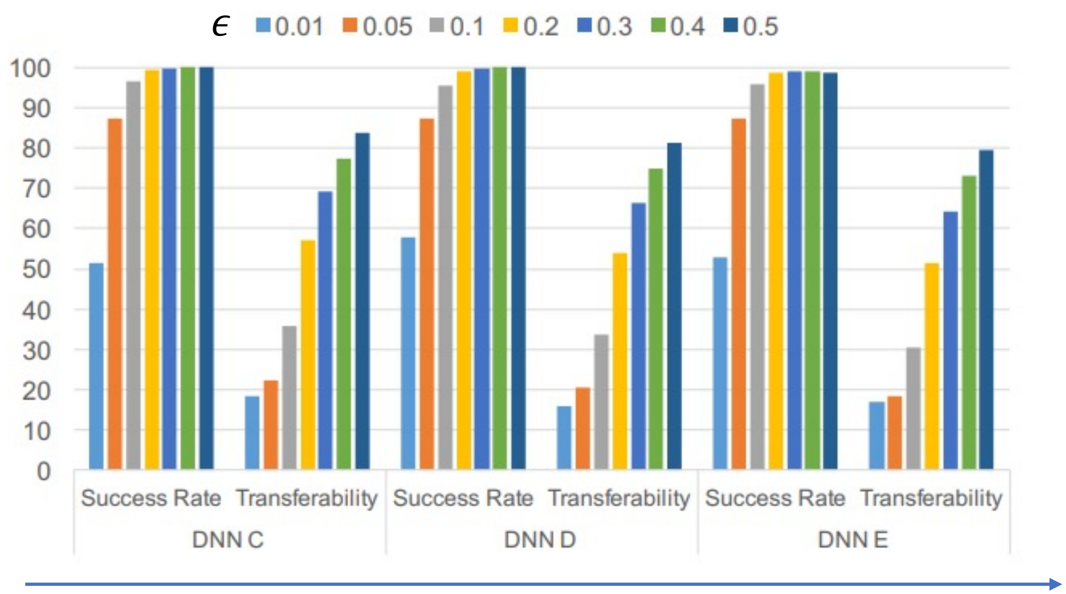
	Training of Machine Learning Model:	**Generate Adversarial Example**
Target	Estimate θ $\mathrm{s.t.\,loss}\left(\mathrm{l}, \mathrm{M}_\theta(\mathrm{x})\right)$ is minimized	Estimate small r s.t. $\mathrm{loss}\left(l, \mathrm{M}_\theta(\mathrm{x}+r)\right)$ is **maximized.** $(x + r \in [0,1])$
Simplest Method	Gradient descent: $\theta = \theta - \alpha \nabla_\theta \mathrm{loss}(\mathrm{l}, \mathrm{M}_\theta(\mathrm{x}))$	FGSM: $x = x + \epsilon\, sign\left(\nabla_x \mathrm{loss}\left(l, \mathrm{M}_\theta(\mathrm{x})\right)\right)$

Goodfellow, Ian J. et al. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

How Adversarial Attack Works (2)

What if we don't have access to the model architecture, but can only use the model as an API?

1. Train a local substitute model with collected data and augmentation
2. Create adversarial examples using gradient based method
3. Feed these examples back to the target model

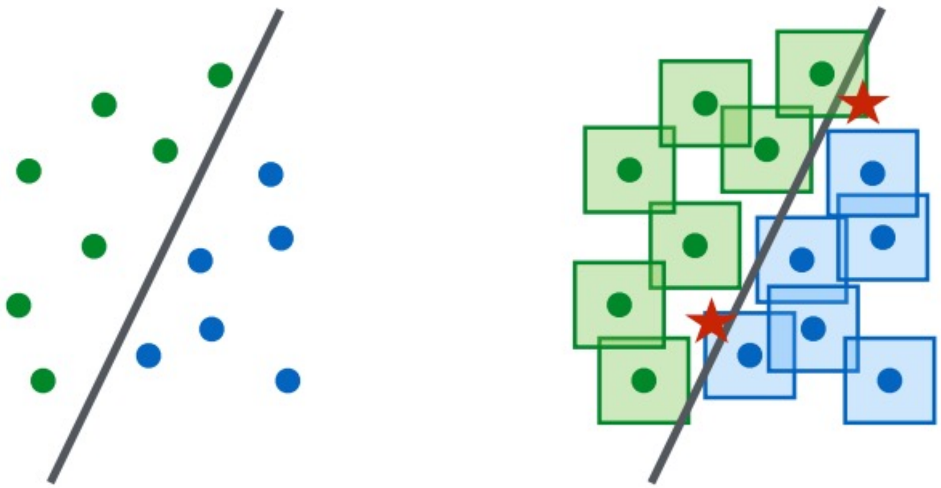


Local substitution

Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." *Proceedings of the 2017 ACM ASIACCS*. 2017.

Why Adversarial Examples Exists

The Deep Learning models we have currently are still too linear!

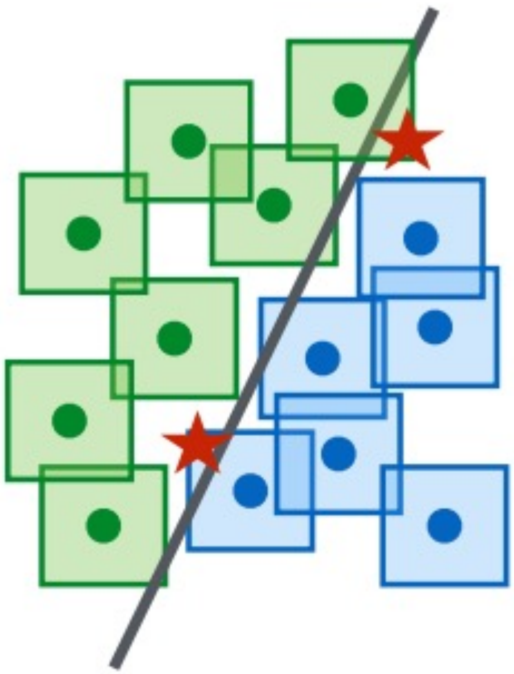


Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017)

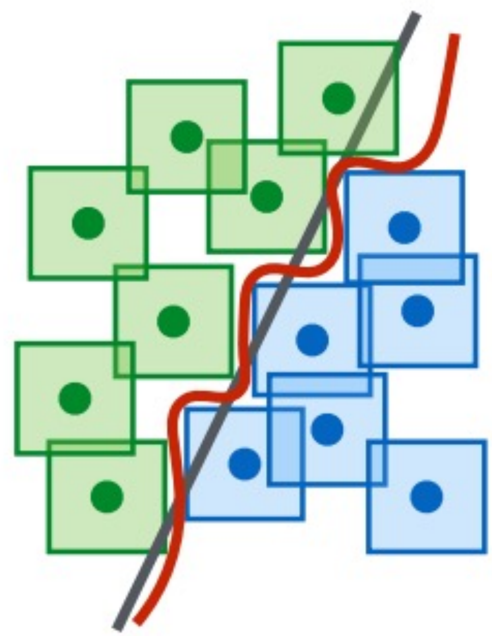
Potential Solutions?

Take the adversarial effect into consideration during training!

Adversarial Training



Standard Training



Adversarial Training

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

How to Defend Adversarial Examples (1)

Instead of using the original data sample, we want to use the worst data sample.

If we want a model M to be robust to perturbation within range ϵ

1. Given a training sample (x, l)
2. Find x' such that $\text{loss}\big(M(l, x')\big)$ is **maximized** and $\max(|x - x'|) < \epsilon$
   - This step is normally very computationally intensive, but we can approximate it with **adversarial methods**
3. Feed (x', l) into the training of the model

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017)

Results of Adversarial Training

CIFAR10

	(a) Standard training: Simple	(a) Standard training: Wide	(b) FGSM training: Simple	(b) FGSM training: Wide	(c) PGD training: Simple	(c) PGD training: Wide
Natural	92.7%	95.2%	87.4%	90.3%	79.4%	87.3%
FGSM	27.5%	32.7%	90.9%	95.1%	51.7%	56.1%
PGD	0.8%	3.5%	0.0%	0.0%	43.7%	45.8%

(a) Standard training (b) FGSM training (c) PGD training

What is the observation?

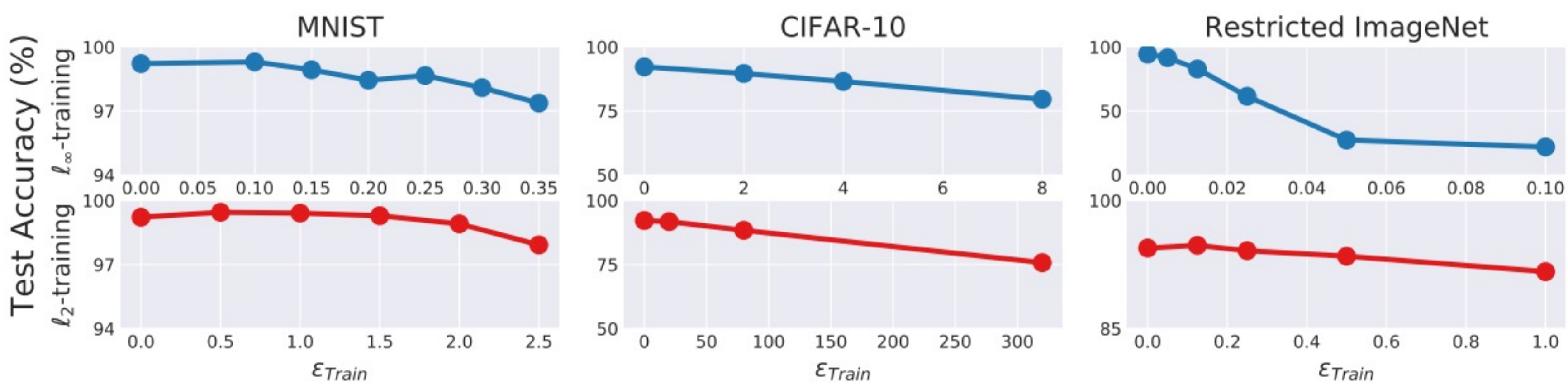
With adversarial training, the accuracy increases for adversarial examples, but decreases for natural examples

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017)

"No Free Lunch"

if we want a robust model, decrease in standard accuracy is inevitable!

Basic intuition: If we want the model to be robust against perturbations, we need to **discard some features that only weakly correlate with the label, but contribute to the standard accuracy**



Tsipras, Dimitris, et al. "There is no free lunch in adversarial robustness (but there are unexpected benefits)." *arXiv preprint*

Future Direction

Like any security problem, defenses to existing adversarial attacks are proposed, but new attack comes and defeat these defense

Can we end this battle once and forever? **Certified Training!**

- *"Certified Defenses against Adversarial Examples", Aditi Raghunathan et al.:* **Defense against all perturbation-based attacks, but only works on small networks**

- *"Certified Adversarial Robustness via Randomized Smoothing",* Jeremy Cohen et al: **Scale certified training to ImageNet**

- "Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing": **Robust model for predicting Top-k labels on ImageNet**

Thank you