AI/ML Security
CyberWarriors Summer Camp 2022

Amanda Fernandez, PhD
Assistant Professor - UTSA
AI/ML Security

Outline

• Artificial Intelligence & Machine Learning

• Attacking ML Models

• Generative Networks

• Deepfakes
Artificial Intelligence

Robots

Agents

Learn

Communication

Reason

Plan

Remember

Interact

Big Data

Automated Scheduling

Data Analytics

Language Processing

AI
Artificial Intelligence

Types: Expert Systems & Reinforcement Learning
Machine Learning

*Types: Supervised & Unsupervised*
Machine Learning

Building models
Machine Learning
..easy, right?

(Pinna and Gregory, 2002)
Adversarial Attacks

Examples & patches

- Attacking a neural network involves providing data that affects its performance.

- Goal: data should look innocuous!

Tesla’s autopilot tricked into driving on the wrong side of the road

NewScientist

Keen Security Labs, of Chinese tech company Tencent, confused a Tesla Model S by placing 3 stickers on the road.
Adversarial Attacks

Examples

- **fish** (92%)
- **boat** (79%)
Adversarial Attacks

Examples

Image source: TensorFlow.org
How robust are AI models?

*Attacking the state-of-the-art*

- Access to model, training data
- Known architectures
  - \textit{Backbone CNN models increase susceptibility to white box attacks!}
- Transferability is key

![Diagram showing the process of attacking AI models](image)
Adversarial Attacks

Examples & patches

Fig. 1: A silhouette of the eyeglasses we use.

Fig. 2: Examples of raw images of eyeglasses that we collected (left) and their synthesis results (right).

Brown et al, 2017

Sharif et al, 2017
Machine Learning
Generative Models

Generating candy hearts

...or “motivational” posters!
Deep fakes

https://www.youtube.com/playlist?list=PLpaGT3slbH0AGdScmPBAqKZ3SZEEoGnpd
Deep fakes

“Everybody Dance Now”
https://www.youtube.com/watch?v=PCBTZh41Ris
Interested in AI/ML?

Start here:

• Teachable Machine

• YouTube, podcasts

• Course websites, eDx, Coursera, Udemy, …

• Futurism, MIT News

• *Start slow! Learn Python, review linear algebra*
UTSA VAIL
Vision & Artificial Intelligence Lab

Amanda Fernandez, Ph.D.
Assistant Professor
Department of Computer Science

Michael Geyer
PhD Student

David Patrick
PhD Student

Ethan Payne
PhD Student

- Secure adversarial learning
- Semantic segmentation
- Applications of deep learning & computer vision

http://cs.utsa.edu/~fernandez/vail.html
Q&A
How well can AI understand visual data?

How generalized is this understanding...
..across datasets?
..across domains?

Deep Learning
• Few shot learning
• Applications in nuclear materials

Computer Vision
• Semantic segmentation
• Visual saliency estimation
• Applications in VR/AR
• Applications in autonomous driving

How can it be fooled?

Cyber Security
• Adversarial examples
• Defenses for visual models