463.14 Automotive Security in Adversarial Learning Environments

Computer Security II
CS463/ECE424
University of Illinois
Tesla's 'full self-driving' rolls back its privacy protection of trip videos

Previous versions of Tesla's North American owners manual for the Model 3 and Model Y said that "to protect your privacy, cabin camera images and video clips transmitted to Tesla servers ... are not associated with your Vehicle Identification Number." (Tesla's "cabin camera" films inside alerts, to remind you to keep your eyes on the road when Autopilot is engaged. To protect your privacy, camera images do not leave the vehicle itself and are not transmitted to anyone, including Tesla, unless you enable data sharing. Even in the event of a safety concern like a collision, only a log of the alerts displayed to the driver may be transmitted to Tesla, excluding camera data. If you choose to enable data sharing, in the occurrence of a serious safety risk or a safety event like a collision, it allows the vehicle to share short video clips with Tesla to help us further develop future safety features and software enhancements such as collision avoidance updates and more. To adjust your data sharing preferences, use your vehicle's touchscreen to tap Controls > Safety & Security > Data Sharing > Allow Cabin Camera Analytics. To protect your privacy, this data is not linked to your VIN, and used to continuously improve the intelligence of features that rely on Cabin Camera. You may change this setting at any time.

https://www.tesla.com/legal/privacy#location-data
In The News: Crashing of Self-Driving Car (Uber 2018)

• “Inadequate safety risk assessment procedures”

• The system is not trained to react to pedestrians crossing the street outside of designated crosswalks

• Vehicle operator distracted by personal cellphone

In The News: Crashing of Self-Driving Car (Tesla 2016)

• Car's cameras failed to pick out a **white trailer** against a **bright sky** in Florida
Case Study 1: Adversarialial Examples to Attack Vision Sensors

Perils of Stationary Assumption

Traditional machine learning approaches assume

Training Data \approx \text{Testing Data}
Autonomous Driving in Practice
Adversarial Examples

Adversarial Perturbation In ML

\[
\min_{\theta} J(\theta, x, y)
\]

Model parameters \quad Input feature vector \quad label

\[
\max_{\epsilon} J(\theta, x + \epsilon, y)
\]

Adversarial perturbation

How to solve the adversary strategy

- Local search
- Combinatorial optimization
- Convex relaxation
Optimization Based Attack

\[
\begin{align*}
\text{minimize} & \quad \mathcal{D}(x, x + \delta) \\
such that & \quad C'(x + \delta) = t \\
x + \delta & \in [0, 1]^n
\end{align*}
\]

Large probability of \(x+\delta\) belonging to a target class \(t\)

Small perturbation

\[
\begin{align*}
\text{minimize} & \quad \mathcal{D}(x, x + \delta) + c \cdot f(x + \delta) \\
such that & \quad x + \delta \in [0, 1]^n
\end{align*}
\]

Vulnerabilities of Perceptron Systems of Automobiles

Robust Physical-World Attacks on Deep Learning Visual Classification. CVPR, 2018
However, What We Can See Everyday...
The Physical World Is... Messy

Varying Physical Conditions (Angle, Distance, Lighting, ...)

Physical Limits on Imperceptibility

Fabrication/Perception Error (Color Reproduction, etc.)

Background Modifications

Evtimov, Eykholt, Fernandes, Kohno, Li, Prakash, Rahmati, and Song, 2017
Creating Robust Physical Adversarial Examples

\[ \text{argmin}_{\delta} \lambda \| \delta \|_p + J(f_\theta(x + \delta), y^*) \]

Perturbation/Noise Matrix \( \rightarrow \) L\(_p\) norm (L-0, L-2, ...) \( \rightarrow \) Loss Function \( \rightarrow \) Adversarial Target Label

\[ \text{argmin}_{\delta} \lambda \| \delta \|_p + \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x_i + \delta), y^*) \]

\{ STOP, STOP, STOP, STOP, STOP, STOP, STOP \}
Optimizing Spatial Constraints
(Handling Limits on Imperceptibility)

\[ \arg\min_{\delta} \lambda \| M_x \delta \|_p + \frac{1}{k} \sum_{i=1}^{k} J( f_\theta(x_i + M_x \cdot \delta), y^*) \]

Subtle Poster
Camouflage Sticker

Mimic vandalism
“Hide in the human psyche”
Handling Fabrication/Perception Errors

\[
\arg\min_{\delta} \lambda ||M_x \cdot \delta||_p + \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x_i + M_x \cdot \delta), y^*) + NPS(M_x \cdot \delta)
\]

\[
NPS(\delta) = \sum_{\hat{p} \in \delta} \prod_{p' \in P} |\hat{p} - p'|
\]

P is a set of printable RGB triplets

Color Space

Sampled Set of RGB Triplets

NPS based on Sharif et al., “Accessorize to a crime,” CCS 2016
How Can We Realistically Evaluate Attacks?

**Lab Test (Stationary)**

- Angles = 0°, 15°, 30°, ...
- Road Sign (Top View)
- Camera

**Field Test (Drive-By)**

- ~ 250 feet, 0 to 20 mph
- Record video
- Sample frames every k frames
- Run sampled frames through DNN
Lab Test (Stationary)

Target Class:
Speed Limit 45
Art Perturbation
Subtle Perturbation
Thinking more about Physical objects
Similar attack against LiDAR sensors
Numerous Defenses Proposed
Case Study 2: Attacking GPS Sensors

GPS Navigation Systems used by 1+Billion Users

• GPS navigation is widely used by drivers around the world

• Self-driving cars rely on GPS for navigation and on-road decisions
GPS Navigation Systems used by 1+ Billion Users

GPS malfunction can lead to real-world consequences
Known Threat: GPS Spoofing

- **Civilian GPS** is vulnerable to spoofing attacks
  - A lack of authentication of signal source
- Take over victim GPS via brute-force jamming or smooth methods
Portable GPS Spoofer is Affordable ($223)

- A Pen (for size reference)
- Raspberry Pi ($35)
- Mobile Charger ($10)
- Antenna ($3)
- HackRF One SDR ($175)
GPS Spoofing in **Free Space (Air, Water)**

In 2012, a drone was diverted in White Sands, New Mexico

In 2013, a yacht was diverted on the way from Monaco to Greece
Spoofing Road Navigation: More Challenging

“Turn left” - physically impossible instruction!
Easily alert human drivers
Making the Attack More Stealthy

Physical World

Digital Map

Navigation instructions lead to attacker’s pre-defined location
Core Idea: Calculate Spoofing Location and Timing

Goal: find ghost route to mimic the shape of victim route

Assumption: know rough destination area or checkpoint

Victim route

Ghost route

Ghost location
Route Searching Algorithm

Map

Directed Graph

Exhaustive BFS

Iterative searching

Turn pattern matching

Concatenate victim routes

Trace Driven Evaluation

- 600 real-world trips from the taxi datasets of New York City and Boston
- Deviating attack: 3,507 qualified victim routes per trip
- Endangering attack: 599 out of 600 (99.8%) contains wrong-way
Real-world Experiments

- Experiments with legal permission from local authority and IRB approval
  - After midnight, spoofing signals do not affect outside (-127.41 dBm)

Trigger instructions in time and divert to 2.1 & 2.5 km away
Can Human Users Detect the Attack?

• Let users drive in a simulator
  – Play truck drivers to “deliver packages” from location A to B
  – Advertise the study as a usability study, spoof locations in real time
  – Post-study interview to know why users can/cannot detect the attack
User Study Results

• Attack success rate: 95% (38 out of 40)
  – Two users detect it by cross-checking surrounding environment and the map to find inconsistency (Highway vs. local way)

• Users are more likely to use GPS in unfamiliar areas
  – Not enough pre-knowledge/time to check the inconsistency
  – Heavily rely on voice and turn-by-turn instructions

• Most users experienced GPS malfunction in real life
  – Unstable GPS signal does not alert users
Take-aways: Learning from Users

• It is feasible to manipulate road navigation systems
  – Advanced GPS spoofing strategies
  – Works even when humans are in the loop

• Defense ideas inspired by the user study
  – Cross-check data from digital and physical worlds
    o Computer vision-based localization
  – GPS-free localization & navigation
Remarks

• Different sensors in automobile could be vulnerable against adversarial attacks
• Different attacks are optimized differently but they have common adversarial goals
• General/universal defense is hard, but we can leverage certain properties of learning tasks and develop more robust models
Discussion Questions

• What does it take to make you feel safe to ride in a self-driving vehicle?

• Do you prefer a world of autonomous vehicles or the coexistence of human drivers and autonomous driving (or human drivers only)?

• Can you point out other security/privacy challenges faced by autonomous driving systems?