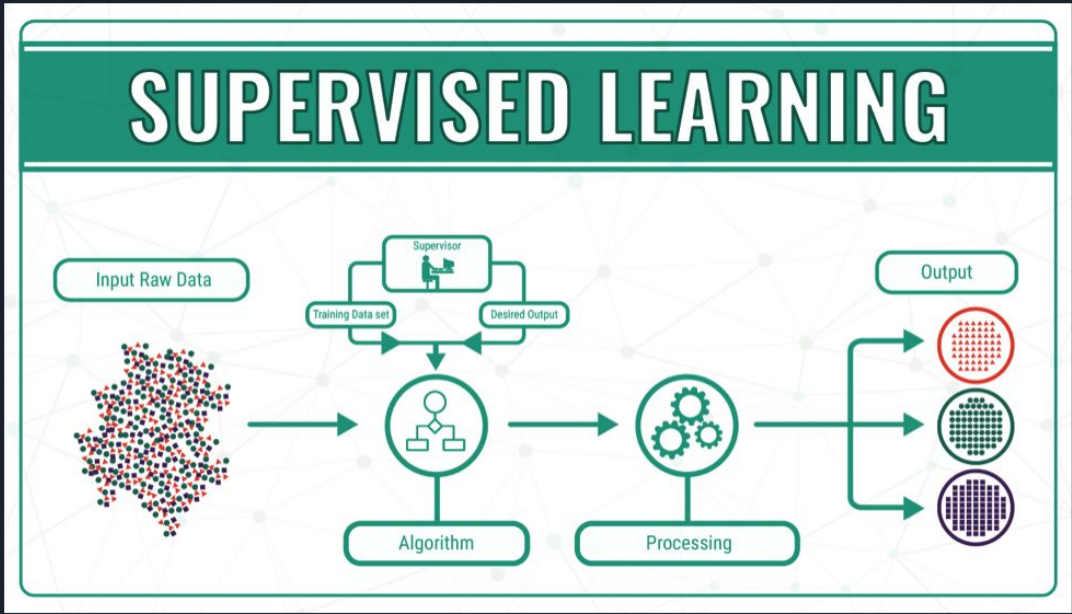
Machine learning and Adversarial Attacks

By Vantou Xiong

What is Machine Learning?

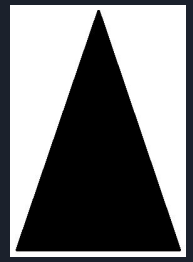
- Classifiers
- Appliances:
  - SnapChat
  - Youtube
  - Siri
  - Email Spam Filter

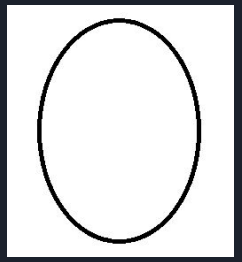


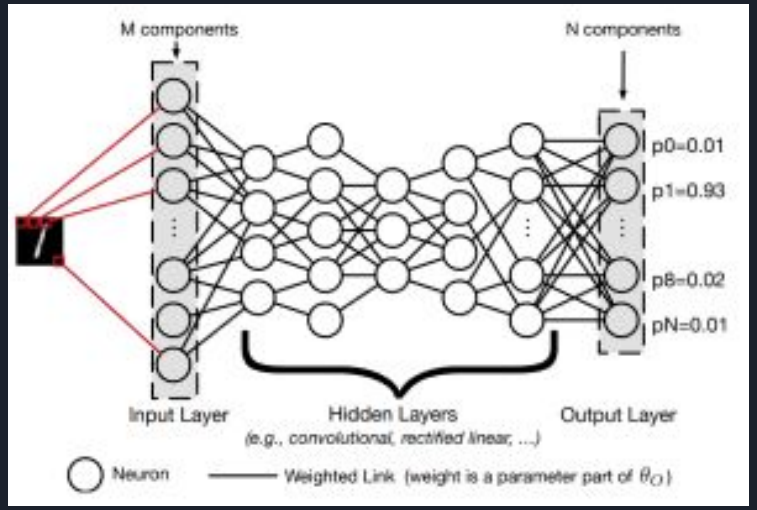
Loon, R. V. [2]

Categorizing Images with Deep Learning

- Input: Pixels in image
- Neural Network
  - Feed Forward
  - Recurrent
- Output: Probability distribution of labels



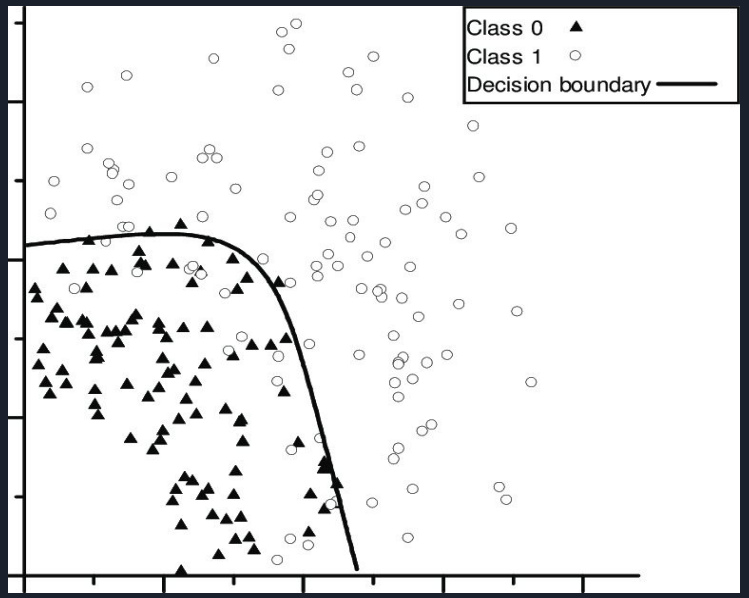




Papernot et al [3]

Categorizing Images with Deep Learning

- Decision Boundaries
- Categorizing inputs



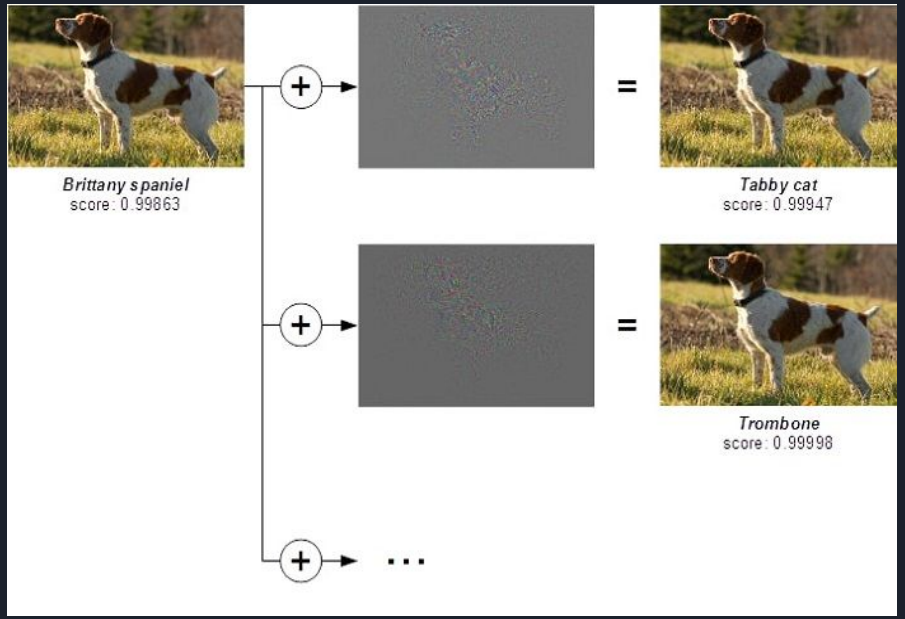
Kemp et al [5]

Adversarial Examples

- Evasion Attack
  - Existing model
  - Minimal perturbed inputs
- Poison Attack
  - Training process
  - Future misclassification

Adversarial Examples - Evasion Attacks

- Fake inputs
- Perturbation function
- Designed to fool ML models



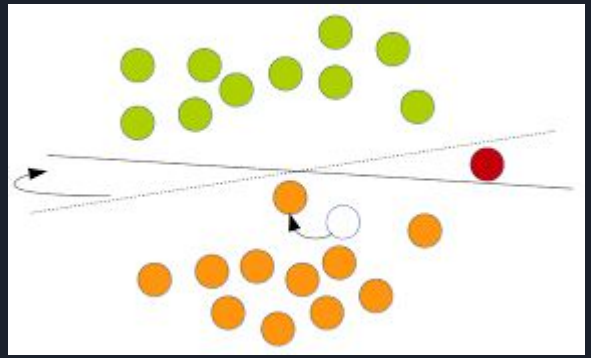
Tanay [4]

Evasion Attacks

- Substitute or surrogate model
- Black box
- White box

Adversarial Examples - Poison Attacks

- Has access to training data
- Inserting bad data
- Happens during training process
- Targets decision boundaries



Polyakov [6]

Poison Attacks

- Backdoor/Integrity attack
- Availability attack
- Influence the machine

Computer Vision

- How computers can see images and classify them
- Able to recognize pictures
- Understanding images

Concrete Examples?

Autonomous Vehicles

- Six levels of Autonomy
  - Level 0: No Automation
  - Level 1: Driver Assistance
  - Level 2: Partial Automation
  - Level 3: Conditional Automation
  - Level 4: High Automation
  - Level 5: Full Automation

Lidar

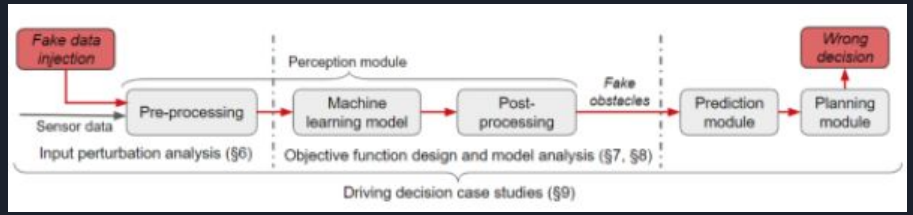
- Sensor using light to measure distances
- 3D point cloud data
- Used for Feature Generation

Feature	Description
Max height	Maximum height of points in the cell.
Max intensity	Intensity of the brightest point in the cell.
Mean height	Mean height of points in the cell.
Mean intensity	Mean intensity of points in the cell.
Count	Number of points in the cell.
Direction	Angle of the cell's center with respect to the origin.
Distance	Distance between the cell's center and the origin.
Non-empty	Binary value indicating whether the cell is empty or occupied.

Cao et al [1]

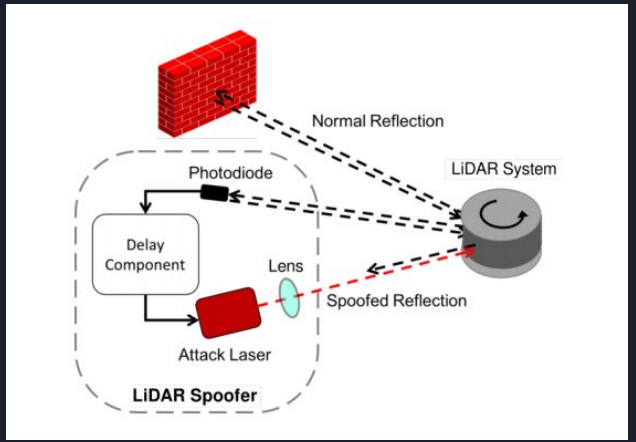
Adversarial Attacks on AV systems

- White Box Attack
- Insert Lidar data via laser
- Target DNN processes the data
- Makes a decision



Adversarial Attacks on AV systems - How it can be done

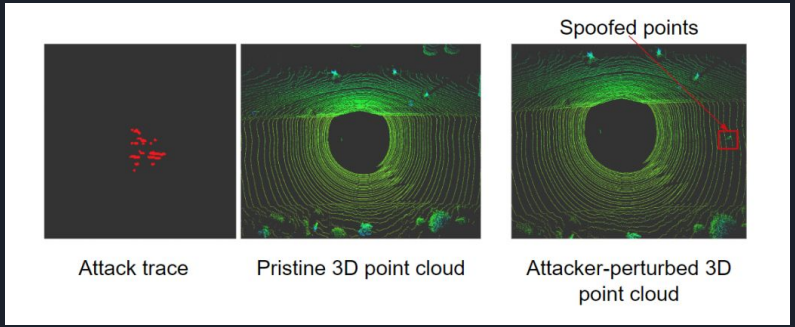
- Need your own laser
- Receives pulse from sensor
- Sends back spoofed reflection



Cao et al [1]

Adversarial Attacks on AV systems

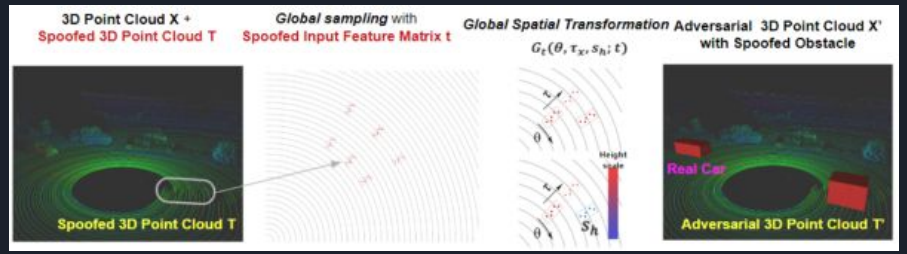
- Perturbation Function
- Merging Function



Cao et al [1]

Adversarial Attacks on AV systems - Result

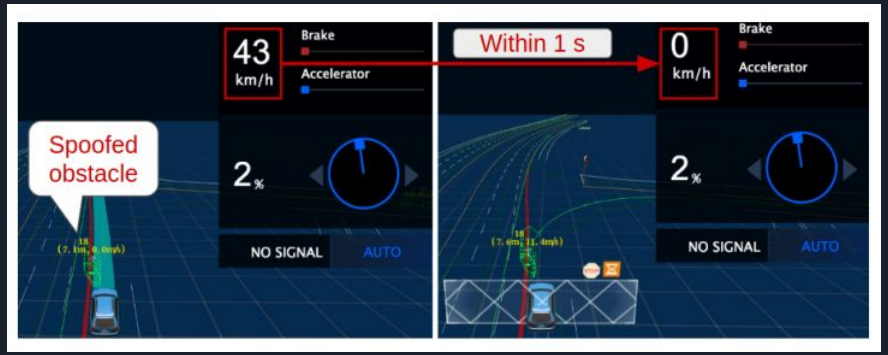
- Initial 3D Point Cloud data
- Perturbed input
- Merged and Transformed



Cao et al [1]

Scenario

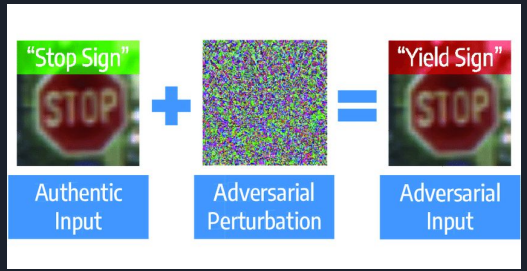
- On the road with target
- Inject spoof data
- Fool the machine learning model
- 75% success rate against Baidu Apollo's ML model



Cao et al [1]

Other scenarios and Impact

- At a stop sign
- Cause “accidental” injuries
- Harder to detect



Kunz [7]

Attacks against other image classifiers

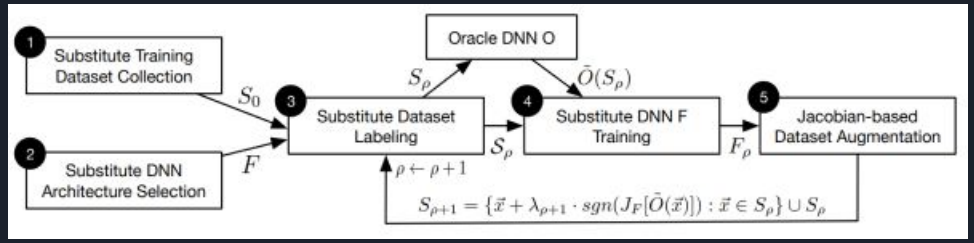
Attacks on MetaMind, Google, and Amazon

- Black Box Scenario
- Number Image Recognition
- Substitute DNN

Papernot et al [3]

Substitute DNN

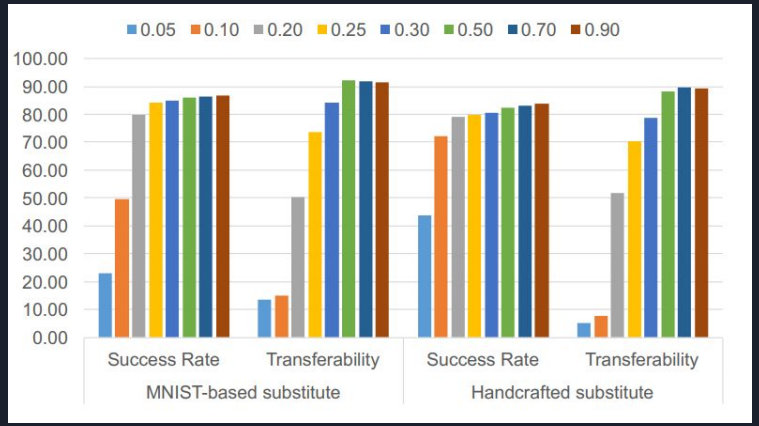
- As minimal queries possible to target DNN O for labels
- Label initial dataset
- Train for similar decision boundaries



Papernot et al [3]

Experiment

- Use MNIST dataset for target DNN training
- Create and train substitute DNN
- Perturb inputs
- Cause misclassifications
- 84% success rate



Papernot et al [3]

Defense Strategies

- Adversarial training
- Manually searching for adversary attacks

Conclusion

- No fool proof method for defense
- Machine Learning models can have security risks
- Adapt and create robust models

References

- Figure 1: Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z. Morley Mao. 2019. Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, New York, NY, USA, 2267–2281. DOI:https://doi.org/10.1145/3319535.3339815

- Figure 2: Loon, R. V. (2018, January 23). Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning. Retrieved October 30, 2020, from https://datafloq.com/read/machine-learning-explained-understanding-learning/4478

- Figure 3: Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17). Association for Computing Machinery, New York, NY, USA, 506–519. DOI:https://doi.org/10.1145/3052973.3053009

References

- Figure 4: Tanay, T. (2018, October). Adversarial Examples, Explained. Retrieved October 30, 2020, from https://www.kdnuggets.com/2018/10/adversarial-examples-explained.html
- Figure 5: Kemp, Roger & Macaulay, Calum & Palcic, Branko. (1997). Opening the Black Box: the Relationship between Neural Networks and Linear Discriminant Functions. Analytical cellular pathology : the journal of the European Society for Analytical Cellular Pathology. 14. 19-30. 10.1155/1997/646081.
- Figure 6: Polyakov, A. (2019, August 06). How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors). Retrieved October 31, 2020, from https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c

References

- Figure 7: Kunz, P. (2019, January 22). Subscription. Retrieved October 31, 2020, from https://ercim-news.ercim.eu/en116/special/detecting-adversarial-inputs-by-looking-in-the-black-box