Imperial College London

18/09/2018

Security in Adversarial Machine Learning

Giulio Zizzo

Imperial College London

Machine Learning Applications

Google DeepMind: Mastering Go via self play in 2017

Imperial College London

Machine Learning Applications

Darktrace cyber security solutions

Imperial College London

Machine Learning Applications

Art Generation

Imperial College London

Machine Learning Applications

Stock Market predictions

Imperial College London

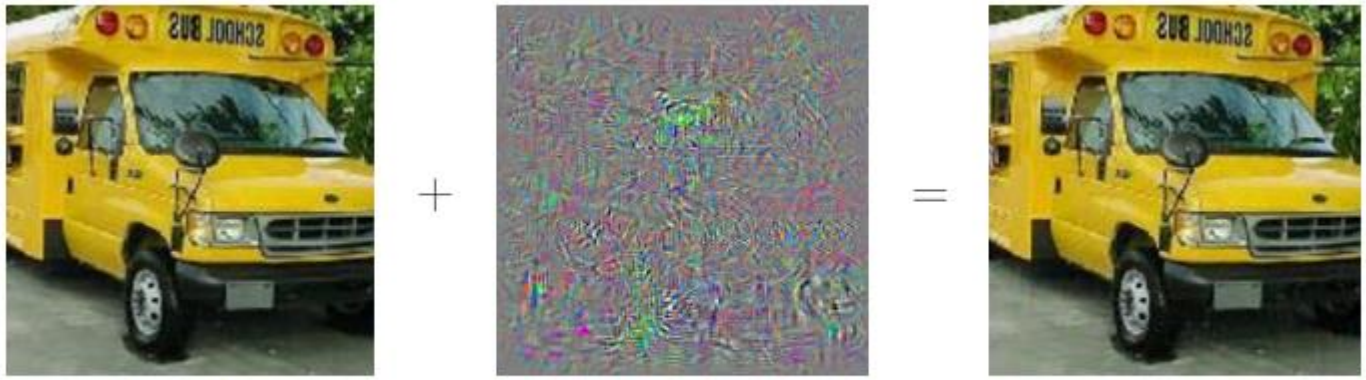
Machine Learning Applications

Autonomous Cars

Imperial College London

Evasion Attacks

- Originally discovered by researchers when trying to better interpret neural networks.



Imperial College London

Industrial Control Systems

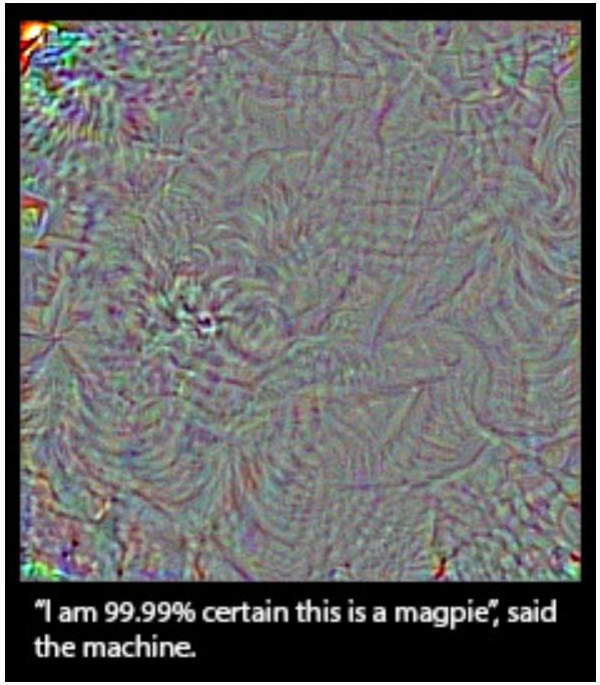
- Sensor data can be fabricated so that a process runs sub-optimally.
- Conducted in such a manner to remain undetected.

Feng, et al. "A Deep Learning Framework for Conducting Stealthy Attacks in Industrial Control Systems." (2017).

Imperial College London

Combating Stealthy Attackers

- Better explore the uncertainty in neural network predictions.
- Networks should be more uncertain about adversarial samples.
- Difficulties:
  - Bayesian approaches don’t scale well to neural networks of any real size like ones used in intrusion detection systems.
  - Off data manifold samples represent an intractable dataset size.



https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html

Imperial College London

Early Results

- High detection rate (95%+) with low false positives (<5%) on simple datasets.

- Caveats:
  - working on adaptive attacker detection.
  - Datasets are simple! (MNIST/CIFAR10) How well does this carry over to more complex multi-stage time-series data such as what is found in industrial control systems?

Imperial College London

Conclusions

- Adversarial Machine Learning is a new and dynamic research area.
- Many challenges remain and must be addressed for machine learning systems to be entrusted with critical decisions.
- Can lead to a deeper understanding of such system and heavily overlaps with neural network interpretation.

Imperial College London

Questions?