Adversarial and Secure Learning

Summer Semester 2019, Chair I20, TUM

Bojan Kolosnjaji, TUM
Ching-Yu Kao, Fraunhofer AISEC

Machine learning is everywhere

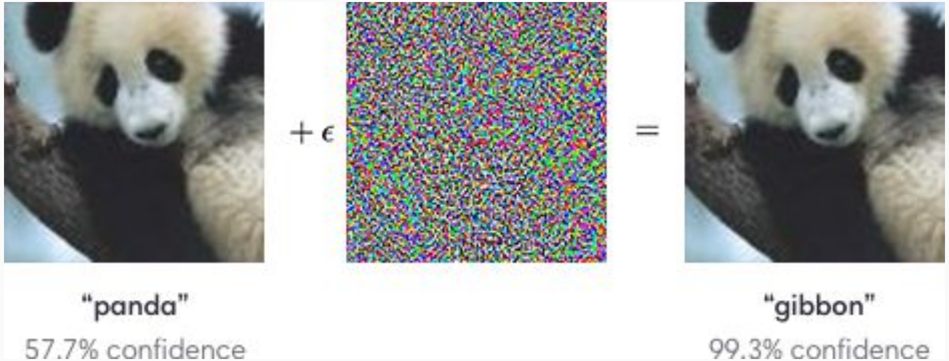
- Computer vision
- Speech recognition
- Biometrics
- Text processing
- Recommendation systems
- Spam detection
- Malware Detection
- ...

Learning in adversarial environment

- Problem considered in the research community at least since early 2000s
- With the hype over machine learning (deep learning) the problem gains importance
- Adversarial perturbations studied in vision, text, malware...

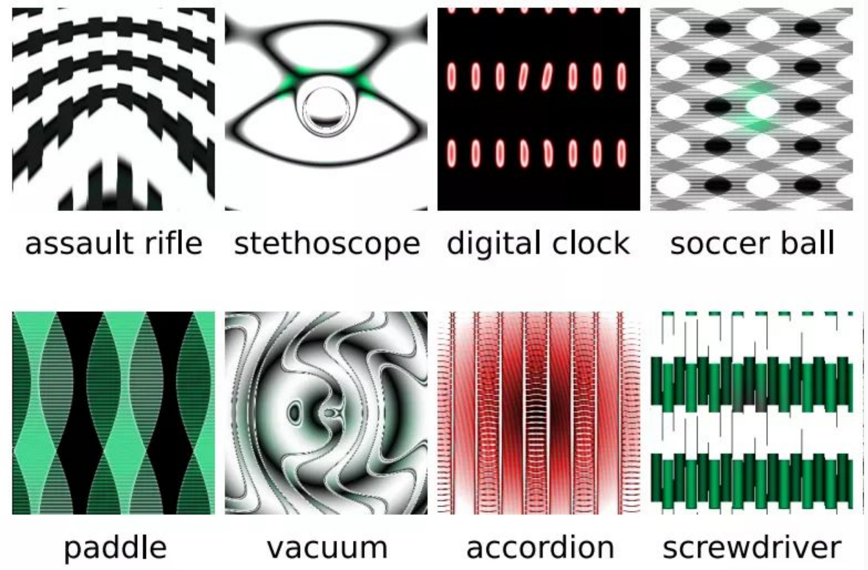
Danger - ML systems are vulnerable

- Easy to perturb data and cause misclassification

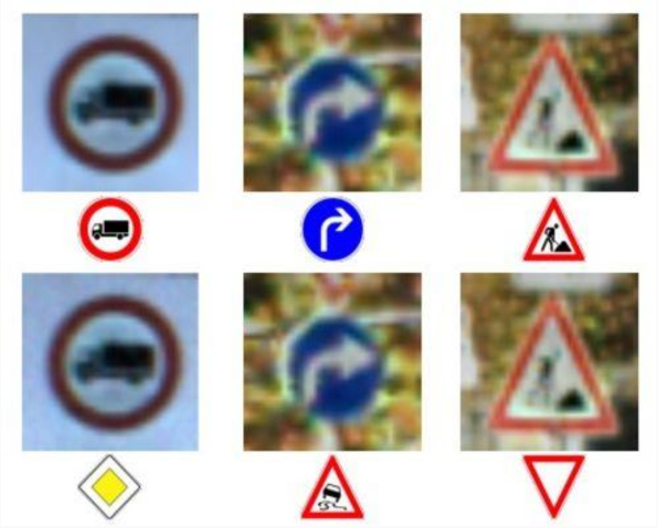


Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Danger - ML systems are vulnerable

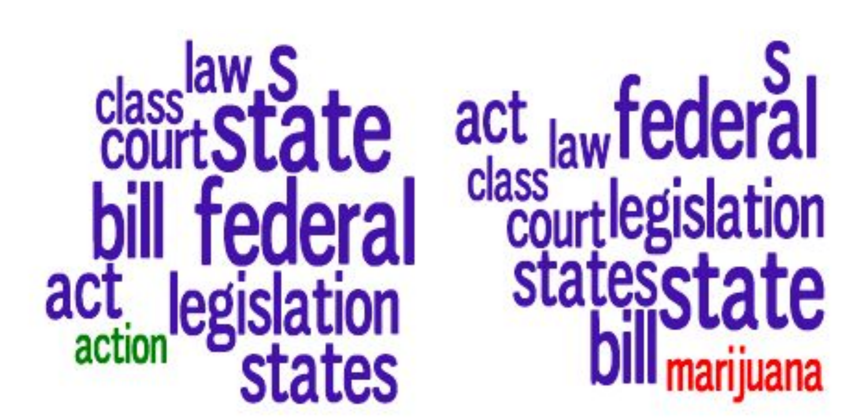


Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

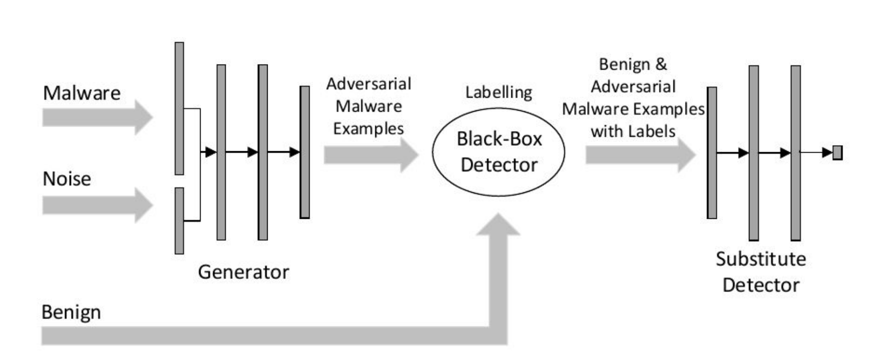


Papernot, Nicolas, et al. "Practical black-box attacks against deep learning systems using adversarial examples." *arXiv preprint* (2016).

Not only computer vision...



Mei, Shike, and Xiaojin Zhu. "The security of latent dirichlet allocation." *Artificial Intelligence and Statistics*. 2015.



Malware
Noise
Benign
Generator
Adversarial Malware Examples
Labelling
Black-Box Detector
Benign & Adversarial Malware Examples with Labels
Substitute Detector

Hu, Weiwei, and Ying Tan. "Generating adversarial malware examples for black-box attacks based on GAN." *arXiv preprint arXiv:1702.05983* (2017).

Rising interest

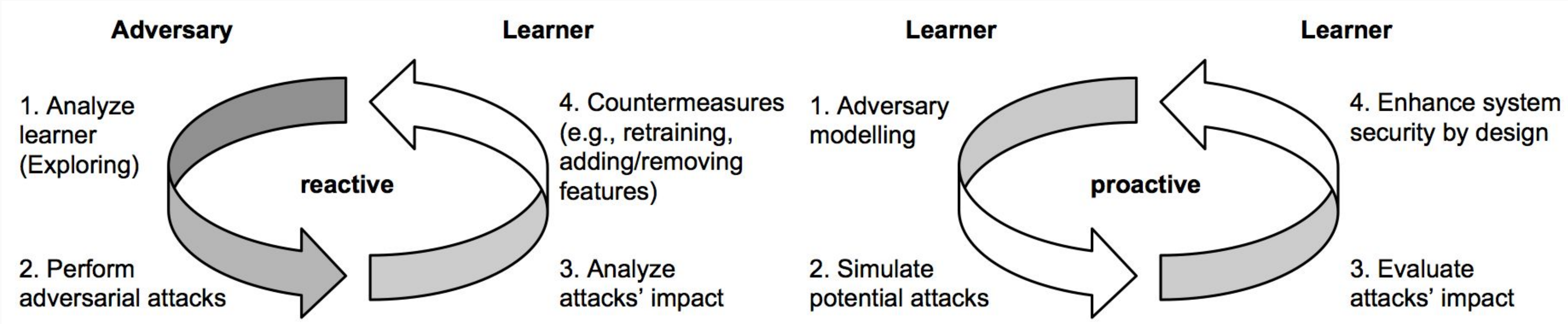
- In the research community
  - Many new papers in top level ML and security conferences, especially since 2015.
  - Still unsolved problems

- In the tech media and general public
  - AI unreliable?
  - What if AI is hacked, we are doomed...

We need to consider attacks (security)

- Potentially unreliable:
  - **Training** data - poisoning
  - **Test** data - evasion
- Evaluate security under **adversarial** environment
- Think about designing **robust** systems

Arms race



Adversary

Learner

1. Analyze learner (Exploring)

2. Perform adversarial attacks

reactive

3. Analyze attacks' impact

4. Countermeasures (e.g., retraining, adding/removing features)

Learner

Learner

1. Adversary modelling

2. Simulate potential attacks

proactive

3. Evaluate attacks' impact

4. Enhance system security by design

From Adversarial to Explainable Learning

- Behavior in adversarial conditions -> new information about learning algorithms

- Better understanding of algorithms -> possibly more robustness

Seminar goals

- Investigate inherent **vulnerabilities** of ML methods
- Special interest for: **SVM**, **Neural Networks**, **Random Forest**
- Consider **attack types** and **countermeasures**
- Study problems in various **application scenarios**
- Be **aware of security** when applying ML in the future
- Prepare for **further research** in this area

Some of the possible topics (1)

- **Evasion** of machine learning classification algorithms
- **Feature selection** in adversarial environment
- Attacks on **Support Vector Machines** (SVM)
- Connections of **Robustness** and **Regularization** in SVM
- Analysis of **adversarial examples** for **Neural Networks**
- Adversarial attacks on **reinforcement learning**, **sequence labeling**, **structured prediction**, **graphs**

...

Some of the possible topics (2)

- **Generative Adversarial Networks**, Adversarial Autoencoders
- Techniques for increasing **robustness** of **Neural Networks**
- Adversarial attacks on **spam detection**
- **Evading** and **Poisoning malware detection** systems
- Attacks on graph-based **anomaly detection**
- **Provably secure** learning and **verification**
- **Tree ensembles** under attack

...

Seminar plan

- 12 students, 12 topics, 6+1 seminar meetings
- Each student gets a **topic** with **2-4** highly regarded research **papers**
- Every student **presents** his topic on one seminar meeting (45 min)
- Students write a short **report** to summarize their topic (14 pages LNCS)
- **Grading** based on the presentation and report

Schedule

- Topics assigned after the matching (more info in a minute)
- Block-seminar - Tuesdays and Thursdays in May (mostly) at 4pm
  - 25.04. - Introductory Meeting - instructions about presentation and report
  - 14.05. - Student Presentation 1,2
  - 16.05. - Student Presentation 3,4
  - 21.05. - Student Presentation 5,6
  - 23.05. - Student Presentation 7,8
  - 28.05. - Student Presentation 9,10
  - 31.05. - Student Presentation 11,12

Prerequisites

- Student of Informatics or similar (advantage to Master students)
- Machine Learning - basic knowledge
- Interest in deeper knowledge of ML methods

How to apply?

- Send an e-mail to kolosnjaji@sec.in.tum.de until 08.02. with the following information:
  - Previous knowledge that qualifies you for the seminar (Machine Learning courses, internships, independent projects,...)
  - Optional: what topics are of your special interest, motivation...
- Apply through the matching system

Topic assignment

- Seminar Topics: published on 25.02.
- Pick and send three favorite topics (ordered list) until 03.03.
- We make final assignment on 04.03.
- Assignment: based on previous knowledge, motivation...

More information

- Follow the course website: https://www.sec.in.tum.de/i20/teaching/ss2019/adversarial-and-secure-machine-learning

- Ask course organizers:

  Bojan Kolosnjaji, TUM: kolosnjaji@sec.in.tum.de

  Ching-Yu Kao, Fraunhofer AISEC:ching-yu.kao@aisec.fraunhofer.de