Security for AI

Breakout Session Slides

Our detailed report is located here: https://cps-vo.org/node/87188

The Need for Secure AI

Physical Limitations

Ivan Evtimov et al. "Robust physical-world attacks on deep learning models." *arXiv preprint arXiv:1707.08945* (2017).

Imperceptible Perturbations

Ramanathan, Arvind, et al. "Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision." *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016.

Vulnerability to Single-Pixel Attacks

Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* (2019).

Susceptibility to Nature

Ozdag, Mesut, et al. *On the Susceptibility of Deep Neural Networks to Natural Perturbations.* Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2019.

Consequences for Autonomy

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." *Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 2017.

Summary: Avenues of Inquiry

1. What does security look like at the data collection, design, training, test, and inference phases?
    a) Successful attacks exist at most of these levels.
    b) Need new techniques for defense.
        i. Beyond adversarial training and defensive distillation
        ii. Theoretical guarantees beyond L2 and Lp norms

2. How do we define metrics for secure AI?
    a) Certified defenses in visual or other real-world norms e.g. malware must execute.
    b) Mutual information for membership inference attacks

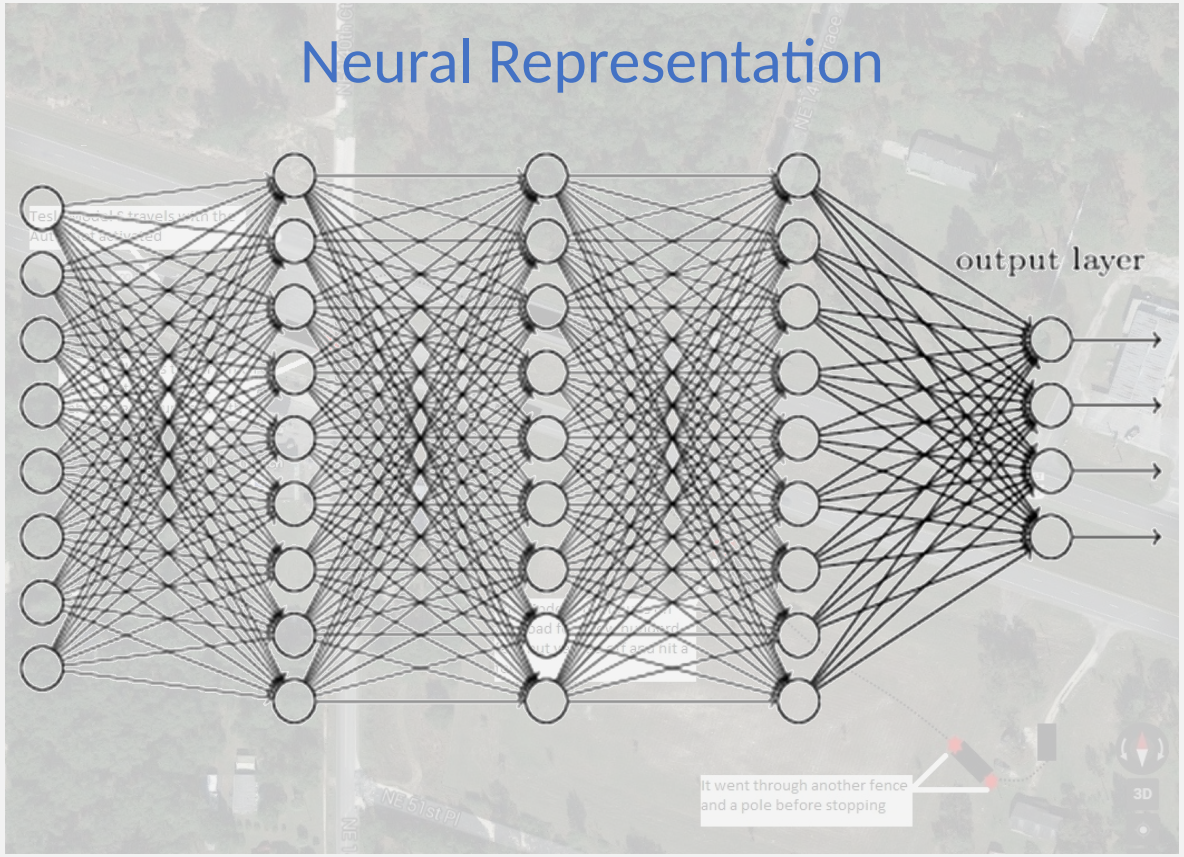
3. What can we formally prove about the security of AI?
    a) Non-linear function approximation beyond ReLUs.
    b) Beyond direct translation to Satisfiability Modulo Theories and Convex Optimization
    c) Neuro-symbolic AI
    d) Autonomy vs. data analytics

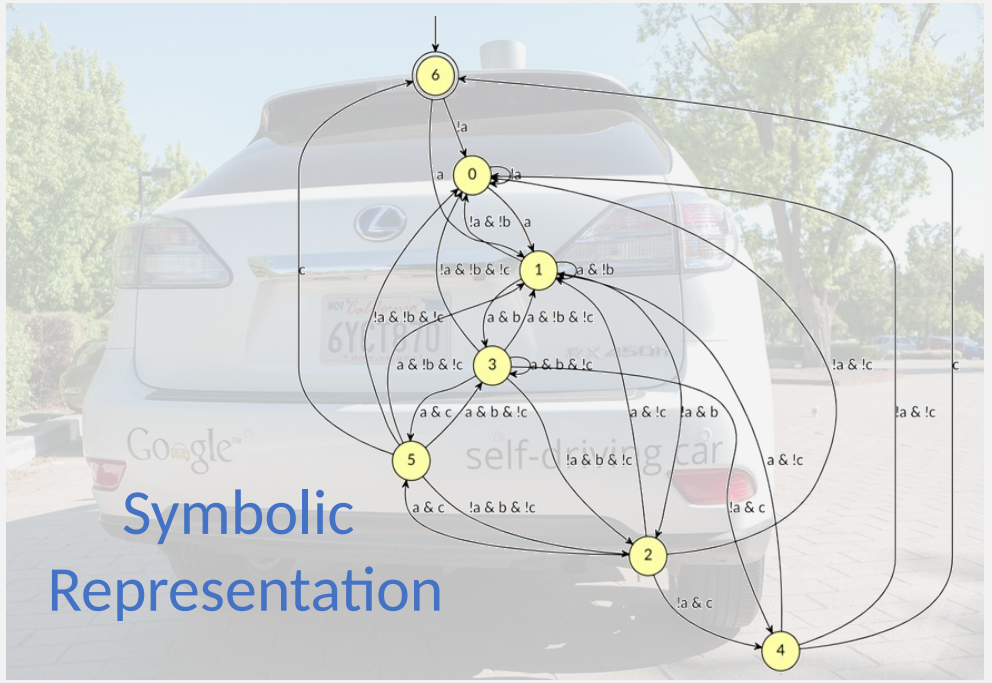
4. Is AI security different from traditional software and hardware security?
    a) Silent errors and the need for explainability in AI e.g. methods for time series.

A Promising Direction: Secure Neurosymbolic AI

"The car assumed that the bus would yield when it attempted to merge back into traffic"

[1] A Google self-driving car caused a crash for the first time. http://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report. (2016).





"The camera failed to recognize the white truck against a bright sky“

[2] Understanding the fatal Tesla accident on Autopilotand the NHTSA probe. https://electrek.co/2016/07/01/understanding-fatal-tesla-accident-autopilot-nhtsa-probe/. (2016).