CAREER: Robustifying Machine Learning for Cyber-Physical Systems

Soumik Sarkar, PhD. Department of Mechanical Engineering, Iowa State University, Ames, IA

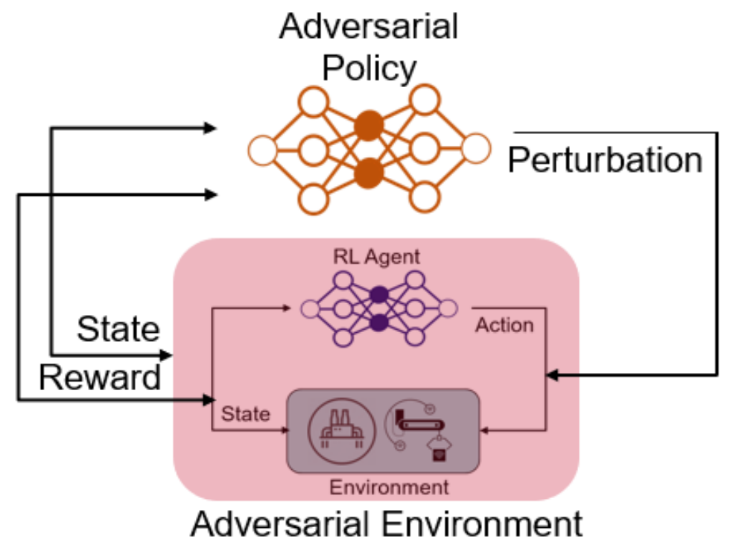
Introduction:

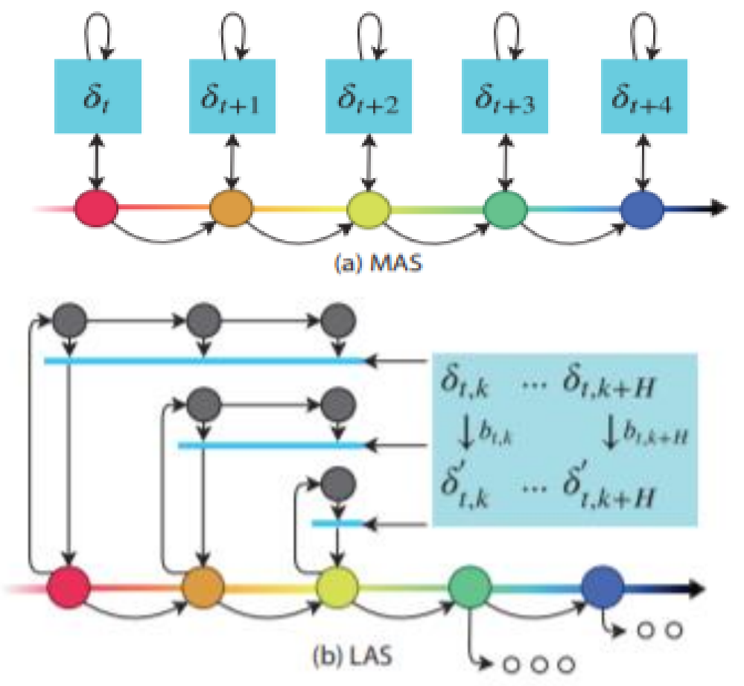
Deep learning frameworks are vulnerable to adversarial attacks in multiple forms, including imperceptible attacks. Attacks can: **1)** occur during training (data poisoning) or inference, **2)** be white (full model knowledge), grey (partial model knowledge) or black (no model knowledge) box, **3)** untargeted (general failure) or targeted (specific failure mode), **4)** learning-based or optimization-based.

Challenges and objectives

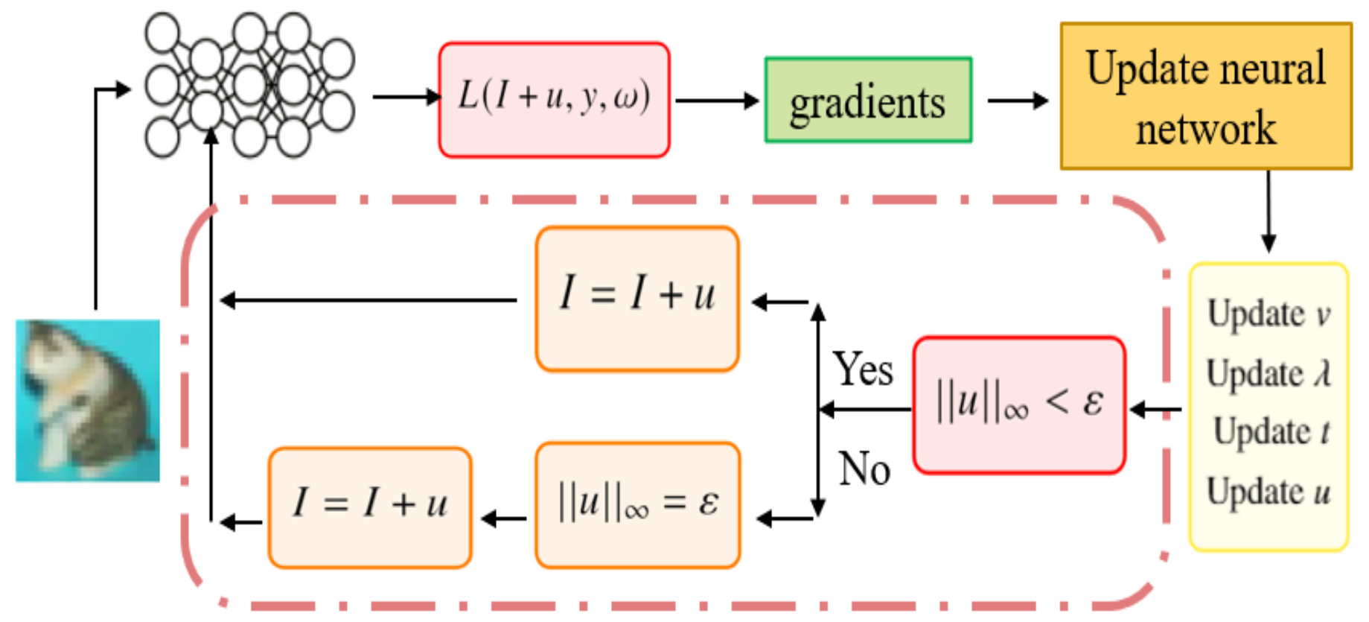
- Understand adversarial attacks on deep learning models for perception (e.g. CNN) & decision-making (e.g. RL)
- Study different forms of threat models to understand vulnerabilities of RL-based controllers in CPS
- Investigate adversarial training schemes to robustify machine learning algorithms
- Develop methodologies to improve efficiency of generating adversarial examples
- Leverage adversarial examples to improve down-stream applications such as robust traffic control & synthetic data generation

Technical Approach / Key Contributions

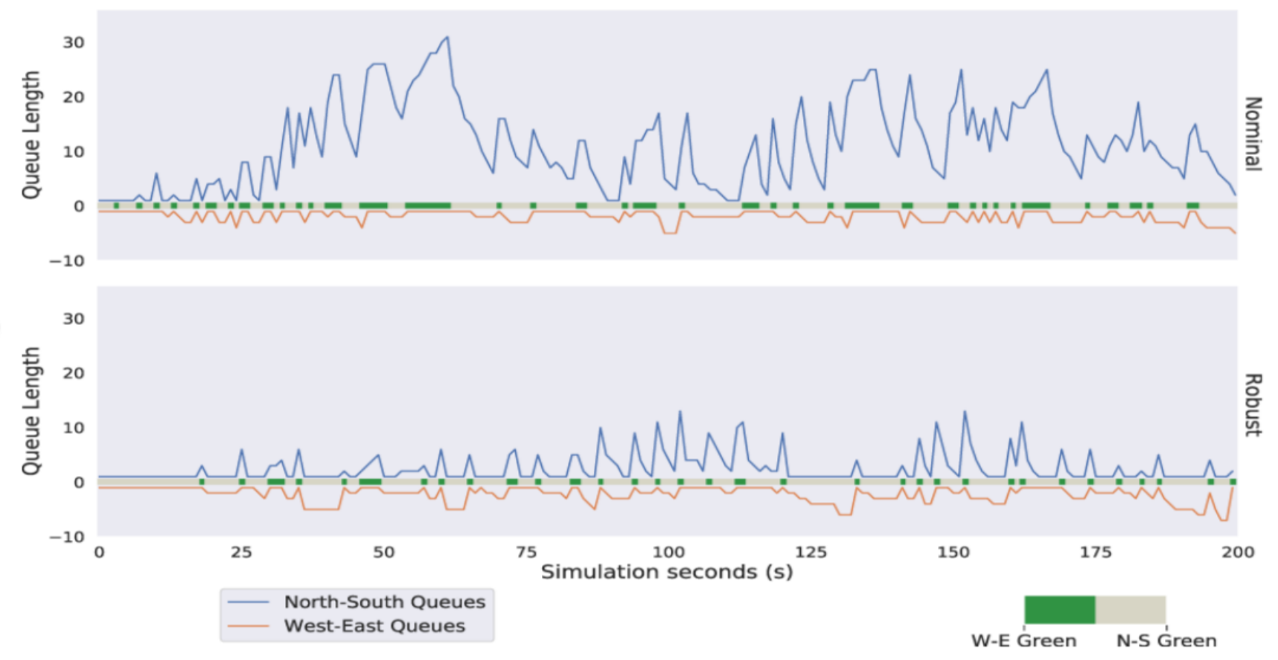




- White-box optimization attacks leverages spatiotemporal information to craft attacks on RL
- RL learns a black-box targeted action space attack on another RL policy
- These methods identify weaknesses in RL policies & enables adversarial training



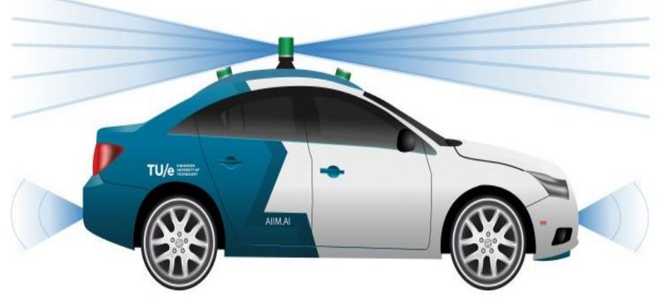
- Method to compute attacks and train the neural networks in iterative steps of ascent/descent. This saves computational time & achieves a robust model compared to state-of-the-art methods.



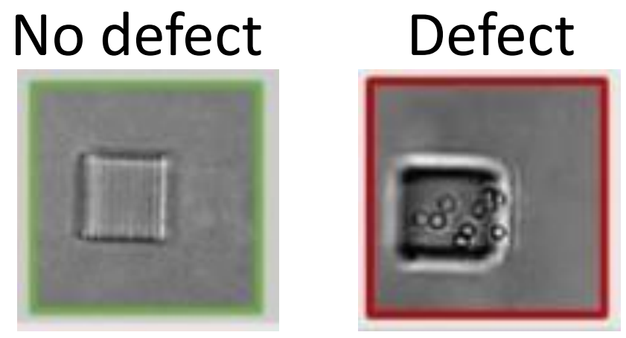
- Robust Deep RL for traffic signal control proposes a robust training framework that robustify agents against noisy approximations of traffic states

Scientific Impact

- Attack models studied are generic and applicable to any commonly used vision and RL-frameworks
- Understanding of model vulnerabilities can lead to development of defense methods
- Robust models can be deployed in various CPS applications that leverage ML modules

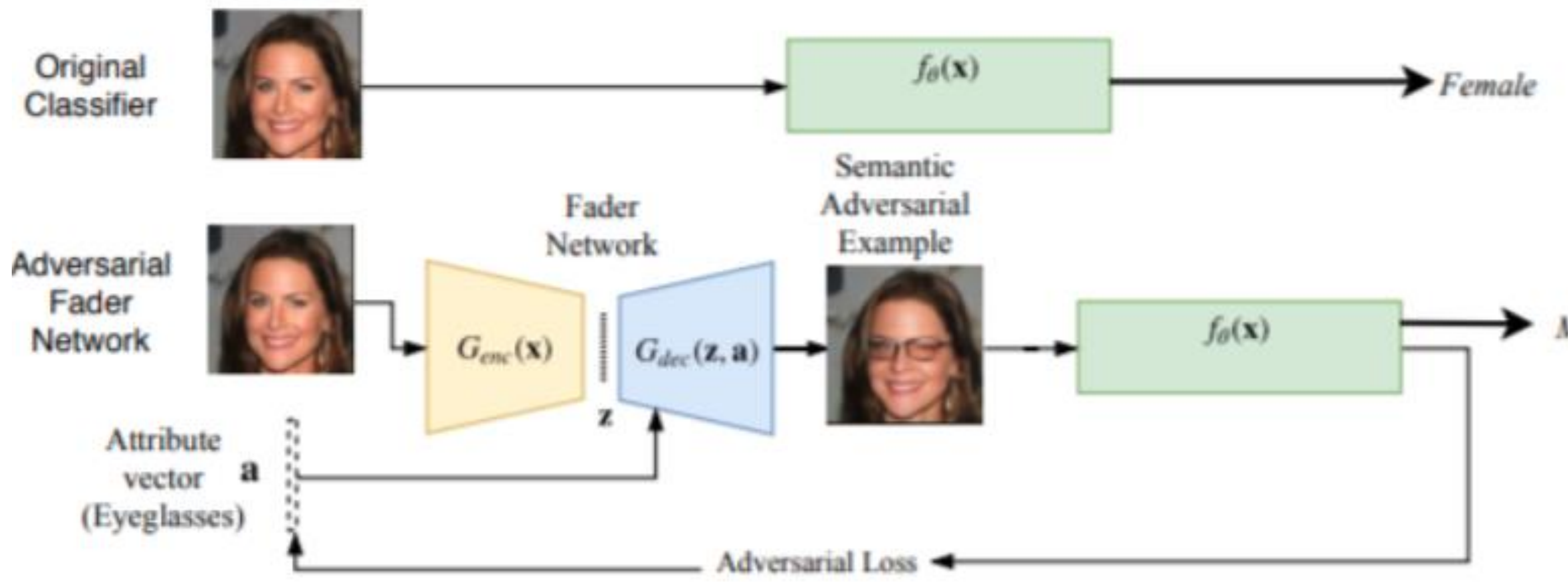

Self-driving

Robotics



Manufacturing

Data Generation



- An algorithm to generate ``semantically'' meaningful adversarial images for classifiers, through attribute conditioned generative models.

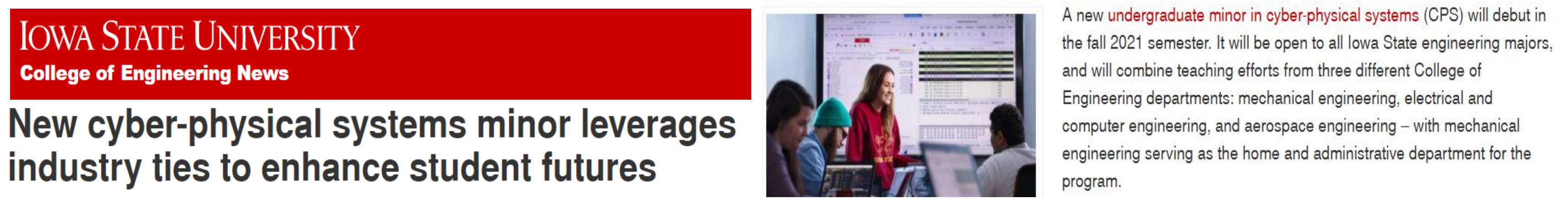
Broader Impacts

Society

1. Proposed robust ML frameworks will be a reliable, safe and efficient technological option for decision-making in safety-critical CPSs such as self-driving vehicles, manufacturing plants & smart biomedical devices.
2. **Certification** - Although many industries are becoming inclined to leverage recent advancements in ML, its tremendous potential cannot be realized without certification that needs safety and robustness guarantees.
3. **Adoption** - the outcomes of the proposed research will significantly improve the adoption of ML by traditional industries.

Education & Outreach

1. PI Sarkar led the development of an Undergraduate minor on Cyber-Physical Systems at Iowa State that will begin in Fall 2021.


IOWA STATE UNIVERSITY
College of Engineering News

New cyber-physical systems minor leverages industry ties to enhance student futures

A new undergraduate minor in cyber-physical systems (CPS) will debut in the fall 2021 semester. It will be open to all Iowa State engineering majors, and will combine teaching efforts from three different College of Engineering departments: mechanical engineering, electrical and computer engineering, and aerospace engineering – with mechanical engineering serving as the home and administrative department for the program.

2. PI Sarkar continues to offer his successful graduate course: “Data Analytics and Machine Learning for Cyber-Physical Systems Applications“ that is improved based on outcomes of this project.
3. PI Sarkar continues his efforts in K-12 outreach activities for STEM education

Impact Quantification

1. The project partially supports 2 female graduate students and 1 Hispanic student in the area of CPS. The goal is to support a few more underrepresented minority graduate students.
2. PI Sarkar is mentoring 2 undergraduate research assistants in the area of CPS. The goal is to support a few more underrepresented minority undergraduate students.
3. PI Sarkar runs deep learning workshops under the NVIDIA Deep Learning Institute program and Midwest Big Data Summer School. These events impact more than 100 participants each year.

2021 NSF Cyber-Physical Systems Principal Investigators' Meeting
June 2-4, 2021

Award ID#: CNS-1845969