ARO MURI: Robust Concept Learning and Lifelong Adaptation Against Adversarial Attacks

Insup Lee (PI)
PRECISE Center
Department of Computer and Information Science
University of Pennsylvania

ARO MURI W911NF2010080
Virtual Kickoff Meeting
30 June 2020

Penn Engineering
Children's Hospital of Philadelphia RESEARCH INSTITUTE
COGNITIVE COMPUTATION GROUP
GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab
PRECISE PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

Team Members & Expertise

INSUP LEE (PI)

BASTANI

KOSTAS DANIILIDIS

ERIC EATON

DAN ROTH

JAMES WEIMER

JULI IS (CHOP)

- Insup Lee: cyber-physical systems (CPS), high-assurance machine learning, security
- Osbert Bastani: machine learning, AI, programming language, security
- Kostas Daniilidis: computer vision, robotics, machine learning
- Eric Eaton: machine learning, life-long learning, interactive AI
- Dan Roth: machine leaning and inference methods, NLP
- James Weimer: learning-enabled CPS, autonomous vehicles
- Julia Parish-Morris (CHOP): developmental psychology, language development, children learning

Penn Engineering
Children's Hospital of Philadelphia RESEARCH INSTITUTE
COGNITIVE COMPUTATION GROUP
GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab
PRECISE

Team Members

INSUP LEE (PI)

BASTANI

KOSTAS DANIILIDIS

ERIC EATON

DAN ROTH

JAMES WEIMER

JUL IS
(CHOP)

- PhD Students:
  - Kaustubh Sridhar
  - Meghna Gummadi
  - Shuo Li
  - Stefanos Pertigkiozoglou
  - Soham Dan
  - Kimberly Tena (CHOP)
  - Vivian Lin (F'20)
- Postdocs:
  - Ivan Ruchkin
  - Souradeep Dutta (F'20)

Penn Engineering
Children's Hospital of Philadelphia RESEARCH INSTITUTE
COGNITIVE COMPUTATION GROUP
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab
PRECISE

Fragility of Deep Neural Networks

How a little electrical tape can trick a Tesla into speeding

Security researchers found an unsettling vulnerability in Tesla's intelligent cruise control.

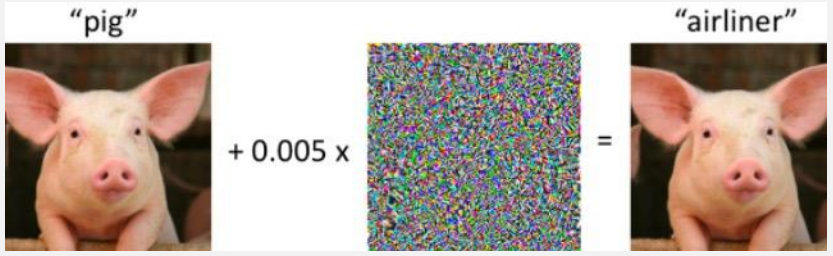
By Rebecca Heilweil | Feb 19, 2020, 2:10pm EST

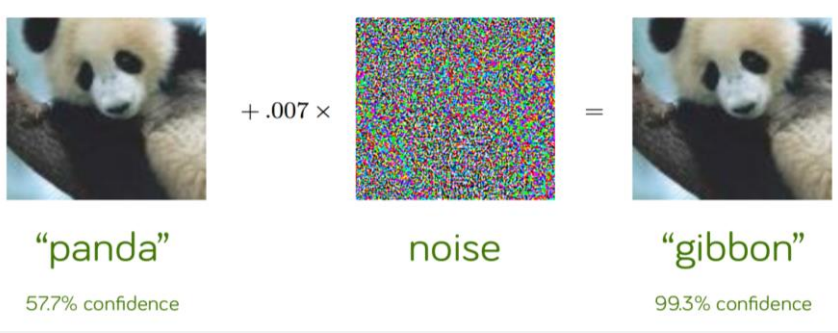
SHARE

Security researchers discovered a simple road sign hack that will trick Tesla's intelligent cruise control feature. | Jonathan Nackstrand/AFP via Getty Images

Digital vs Physical Attacks

Digital Attacks





Physical Attacks

How to protect neural networks from adversarial attacks?

Background

- Lots and lots of work on adversarial machine learning since 2017
  - Arms race: Offensive and defensive techniques
- State-of-the-art: offensive techniques are winning the battle
  - Given a trained model, it is still easy to find realizable adversarial examples
- Fundamental challenge: Adversarial learning has competing objectives
  - maximizing model accuracy vs. maximizing adversarial robustness

Project Goals

- We will develop the foundations for robust and adaptive concept learning in adversarial settings, building upon foundations from childhood development.
  - Employ semantic information and concepts in learning – as children do – so that adversarial attacks without understanding semantics cannot easily fool classifiers
  - Develop truly robust and adaptive learning tools that benefit—as children do—from experiences and interactions in the world
  - Demonstrate the robustness of our techniques to adversarial examples in real large-scale dynamic environments

Overview of the proposed research

Tasks are observational or inter-actable learning problems

System acquires concepts robust to adversarial attack over its lifetime

Concept acquisition mirrors robustness of children's learning process

Thrust II: Adaptive Learning
II.1: Adaptive concept learning
II.2: Lifelong learning

composable knowledge: general knowledge, cross-paradigm representations, specialized knowledge, domain-specific transformations

Concepts
Predictions
Revisions
Confidence
Learned Models
Monitoring Events

Thrust I: Concept-based Learning
I.1: Geometric reasoning
I.2: Inductive reasoning
I.3: Compositional reasoning

Thrust III: Verification & Monitoring
III.1: Verified learning
III.2: Monitoring learning

robustness specification

Verified Robust Concept Learner

Concept Learning Theory

- Prioritize shapes/ geometry, resilient to small noise
- Constant hierarchical lifelong learning in new settings
- Interactive exploration, self-adversarial to develop robustness

Penn Engineering
Children's Hospital of Philadelphia RESEARCH INSTITUTE
COGNITIVE COMPUTATION GROUP
GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab
PRECISE

Child Development and the Future of Adaptive Machine Learning

Julia Parish-Morris, PhD

Children's Hospital of Philadelphia Research Institute

ARO MURI W911NF2010080

JULI IS (CHOP)

Penn Engineering | Children's Hospital of Philadelphia RESEARCH INSTITUTE | COGNITIVE COMPUTATION GROUP

GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab | PRECISE PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

Proposed Thrusts and Tasks

- Thrust 1: Concept-based Learning Robust to Adversarial Examples (Lead: Bastani)
  - Task I.1: Robust learning of visual object and scene representations (Daniilidis, Eaton, Bastani, Parish- Morris)
  - Task I.2: Concept-based deep learning with inductive biases (Roth, Bastani, Daniilidis, Parish- Morris)
  - Task I.3: Compositional inference and reasoning for adversarial learning (Weimer, Lee, Parish- Morris)
  - **Connection with child learning: concept selection and representation**
- Thrust II: Adaptive Learning in Dynamic Environments (Lead: Eaton)
  - Task II.1: Leveraging Inductive Biases for Adaptive Concept Learning (Bastani, Roth, Parish- Morris)
  - Task II.2: Lifelong Learning (Eaton, Daniilidis, Parish-Morris)
  - **Connection with child learning: hierarchical and continual learning in new settings**
- Thrust III: Verification and Monitoring of Learning (Lead: Weimer)
  - Task III.1: Verified learning (Weimer, Lee, Parish-Morris)
  - Task III.2: Monitoring learning (Weimer, Lee, Parish-Morris)
  - **Connection with child learning: trust building, validation by probing, self-adversarial for robustness**
- Thrust IV: Integration and Evaluation (Lead: Lee)
  - Toolset and dataset development
  - Evaluation platform and scenarios

Schedule

- ***9:00 am:*** Kickoff Message, *Purush Iyer*
- ***9:10 am:*** Overview and Child Development/Adaptive Machine Learning, *Insup Lee & Julia Parish-Morris*
- ***10:00 am:*** Thrust I. Concept-based Learning Robust to Adversarial Examples, *Osbert Bastani*
- ***10:30 am:*** Thrust II. Adaptive Learning in Dynamic Environments, *Eric Eaton*
- ***11:00 am:*** Thrust III. Verification and Monitoring of Learning, *James Weimer*
- ***11:30 am:*** Thrust IV: Integration, Platform, and Evaluation, *Insup Lee*
- ***11:45 am:*** Government Caucus
- ***12:15 pm:*** Feedback and Open Discussions