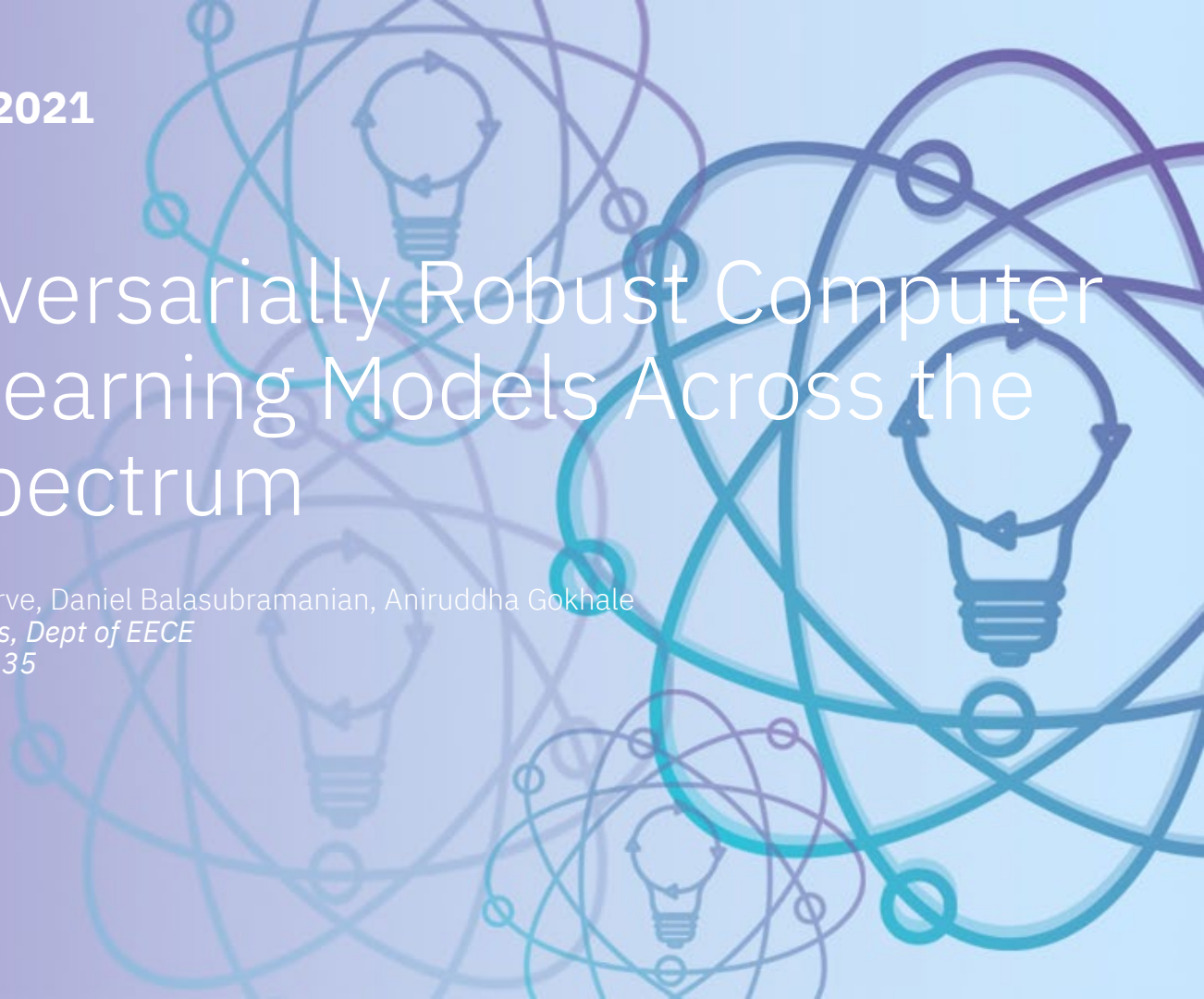
IBM IEEE CAS/EDS
**AI Compute Symposium 2021**
*October 13-14*

# Deploying Adversarially Robust Computer Vision Deep Learning Models Across the Computing Spectrum

Robert Canady, Xingyu Zhou, Yogesh Barve, Daniel Balasubramanian, Aniruddha Gokhale
*Institute for Software Integrated Systems, Dept of EECE*
*Vanderbilt University, Nashville, TN, 37235*



# Challenges in Computer Vision (CV) at the Edge

- Which model to choose
  - Speed vs Accuracy vs Model Size.
- CV models are typically very large.
  - Difficult if not impossible to deploy at edge.
  - Smaller models are often less accurate on clean data.
- Device Heterogeneity.
- Adversarial Machine Learning (AML).
  - Image Classification, Object Detection, NLP, etc.

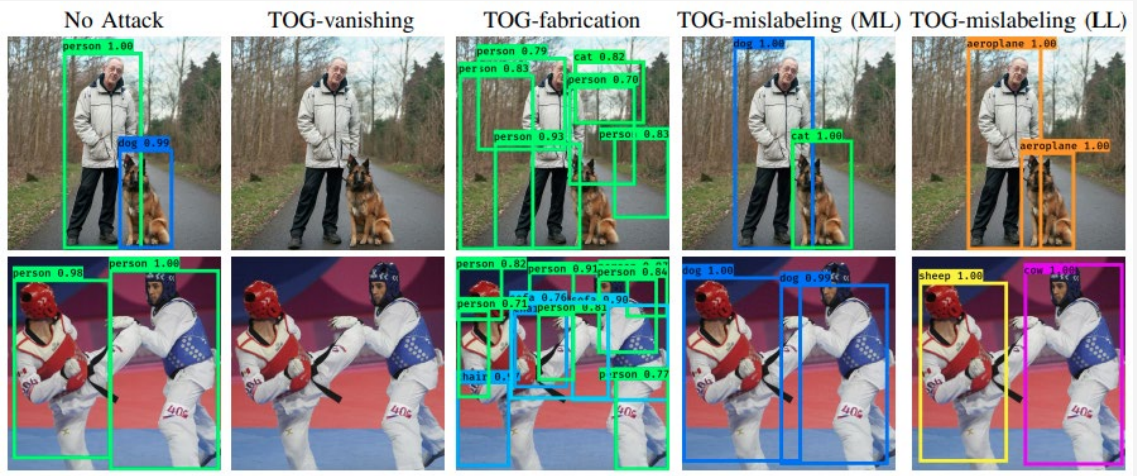


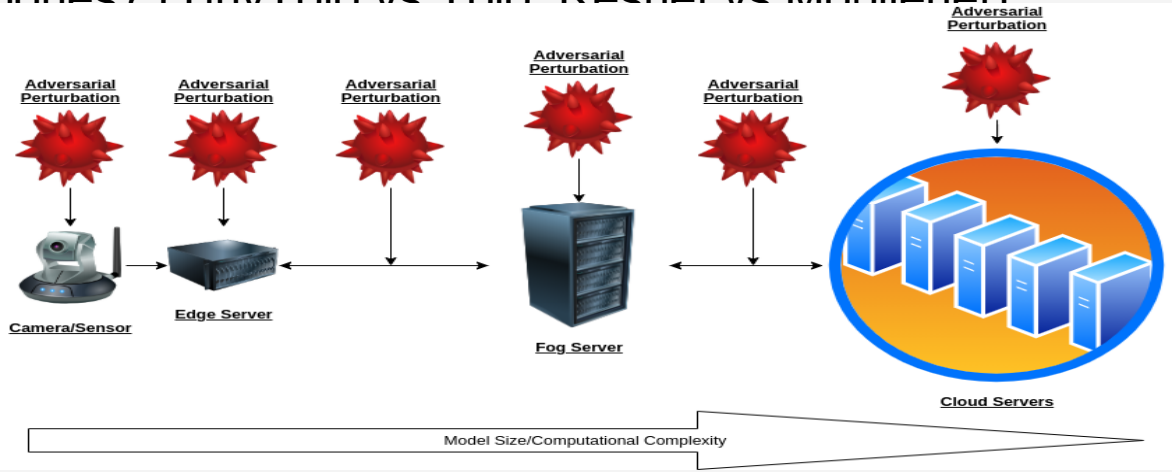

Fig.1. Examples of Object Detection Attacks from the paper, 'Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems'

# More Robust CV Models

- Adversarial Training
  - Typically focuses on one type of attack and perturbation bound.
    - Perturbation bound is essentially the strength of the attack (8/255, 16/255, etc).
  - Mostly explored in image classification domain.
  - Typically doesn't generalize well to other attacks and bounds.
  - Computationally expensive.
- Our Proposed Approach: Quartet Adversarial Training (QUAT).
  - Inject each type of TOG attack using different bounds into the training set, randomly.
  - Should help model generalize to other types of attacked data.
  - Initially added images to training set.
    - Moving towards setting the attacks as a pre-processing transformation.

# Adversarial Robustness During Edge Offloading

- Most AML focused on attacking or defending model with unlimited resources.
- We wish to explore the impact of AML after edge offloading.
  - Do perturbations affect both cloud, fog, and edge the same?
  - AML attacks typically base attack on one type of model.
    - What happens when you have different sizes of the same model or different backbones? (TinyYolo vs Yolo, Resnet vs Mobilenet)



# Initial Results

- Only tested QUAT on TinyYOLOv3 and YOLOv3.
  - Results were promising: 17% mAP increase on attacked data (YOLOv3).
  - Only 1-3% improvement with TinyYOLOv3.
- Plan to test inference on Jetson TX2, Jetson Nano for edge results.
  - FasterRCNN, YOLOv5, SSD, Retinanet.