Washington University in St. Louis
Office of Technology Management

COMPUTER VISION TRAINING SYSTEM FOR ROBUST ARTIFICIAL INTELLIGENCE IMAGE CLASSIFICATION

Tong, Liang, Vorobeychik, Yevgeniy, Wu, Tong

Maland, Brett

T-019303

Technology Description

Researchers in Prof. Yevgeniy Vorobeychik's laboratory have developed a new adversarial model and training methods to defend deep neural networks against physical attacks that corrupt image classifications. This system outperforms previous state-of-the-art techniques and has end-user applications in computer vision such as surveillance, facial recognition and autonomous driving.

Currently, machine learning with deep neural networks provides a valuable tool for image recognition. However, it is extremely vulnerable to malicious physical attacks that make unsuspicious changes to fool the image classifier. For example, small stickers could be used to intentionally deface a stop sign and trick computer vision into classifying it as a speed limit sign. This technology addresses this safety and security issue with a new attack model (rectangular occlusion attack) and training methods that effectively eliminate vulnerability to this type of physically realizable attack. The system provides a direct, generic defense that yields image classification models that are highly robust against the physically realizable attacks.

Examples of Physically Realizable Attacks

Facial recognition *– Adversarial eyeglasses superimposed on image of face leads to the predicted individual on the right.*
Autonomous driving *– Stop sign with an adversarial mask may be classified as a speed limit sign.*

Stage of Research

The inventors developed a new adversarial model (rectangular occlusion attacks) and developed two methods for efficiently computing the resulting adversarial examples. They experimentally demonstrated that this approach is significantly more robust against physically realizable attacks on deep neural networks than previous methods. (Publication)

Applications

- **Computer vision** – deep neural network training for image classification to defend against physically realizable attacks, with end-user applications in:
  - facial recognition
  - surveillance

- autonomous driving (e.g., traffic sign recognition)

Key Advantages

- **Highly robust**:
  - first effective method developed to specifically defend against physical attacks (instead of digital attacks) on deep neural networks
  - dramatically more robust than state-of-the-art alternatives (e.g., adversarial training with PGD attacks and randomized smoothing)

Patents: Application pending

Publications: Wu, T., Tong, L., & Vorobeychik, Y. (2019). Defending against physically realizable attacks on image classification. arXiv preprint arXiv:1909.09552.

Related Web Links: Vorobeychik Lab