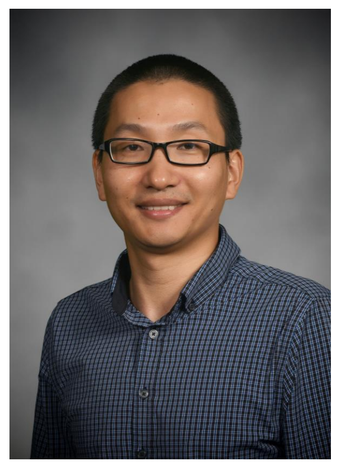
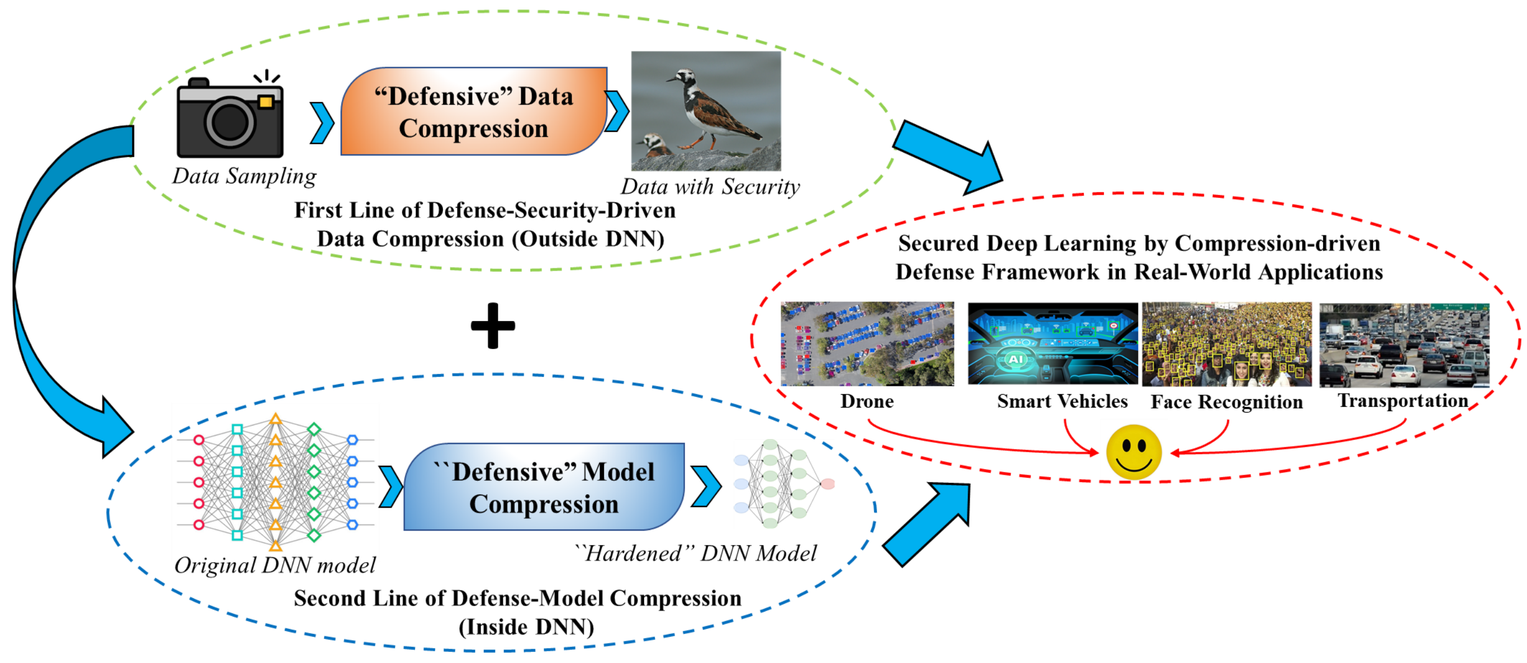
EAGER: Invisible Shield: Can Compression Harden Deep Neural Networks Universally Against Adversarial Attacks?



PI: Wujie Wen, Lehigh University & Florida International University

https://www.lehigh.edu/~wuw219/research.html, CNS-1840813

Deep Neural Networks (DNN) suffer from a security threat: decisions can be misled by adversarial inputs crafted by adding human-imperceptible perturbations into normal inputs. This project investigates a compression based defense strategy to protect DNNs against the attack, with low cost and high accuracy guarantee.



Defense Challenges:

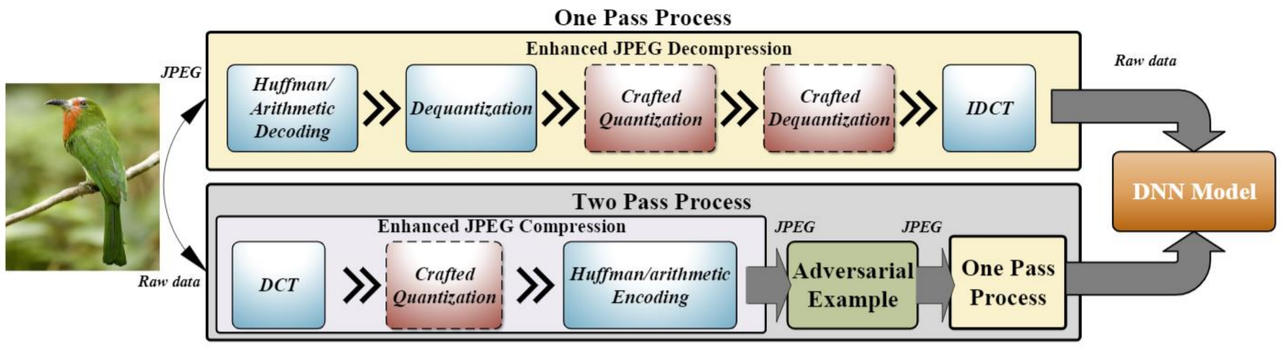
1. Diversified Attack Natures;
2. Unknow Adversary's Strategies;
3. High Implementation Cost;
4. Difficult to Guarantee Accuracy.

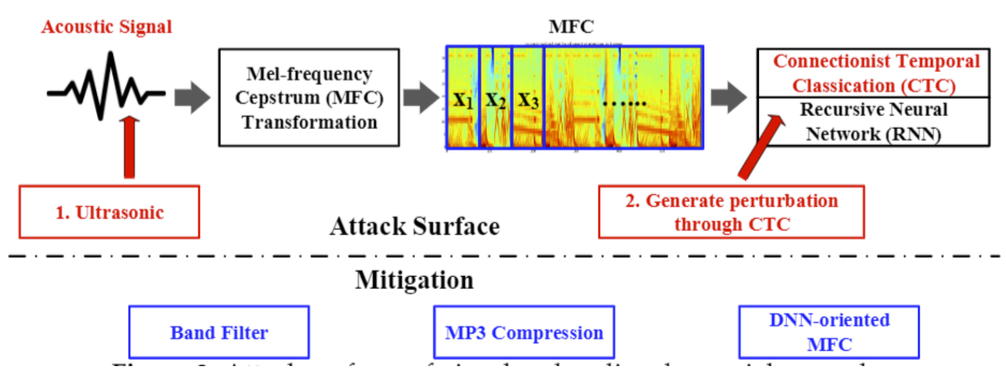
Scientific Impacts:

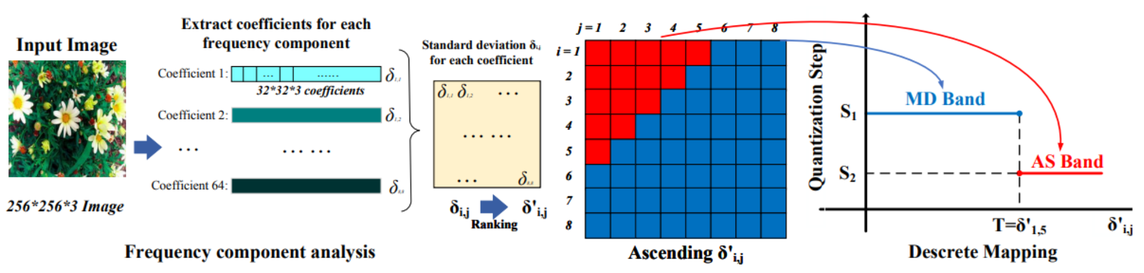
1. A New Paradigm to Secure Deep Learning : Integrating Defense into Compression-Essential Component for Volume Reduction Deployed in Any Practical System;
2. Beneficial to Various Communities: Data Science, Cyber- and Hardware- Security, Computer Vision and Hardware Architecture.

Approaches:

1. Security-aware input compression tailored for DNNs by jointly considering defense efficiency, accuracy and compression [1];
2. Security-driven model compression to harden DNN model [2] [3].

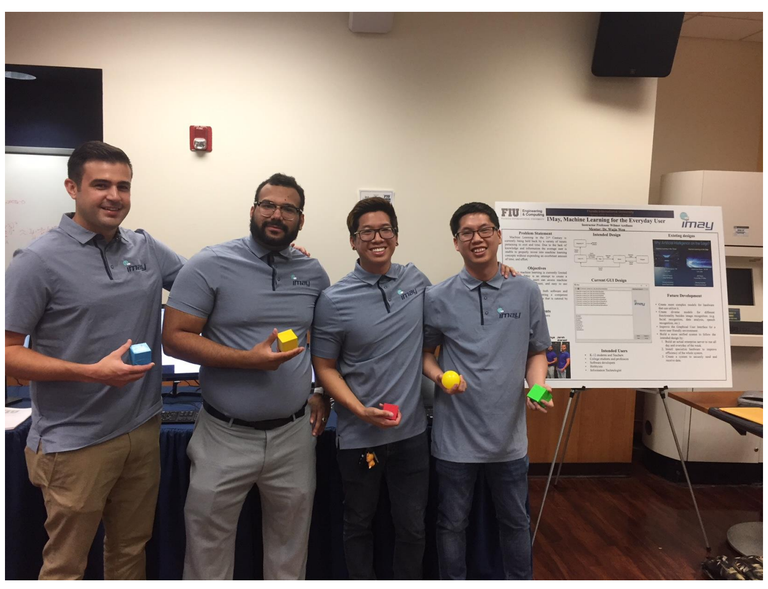






Broad Impacts:

The project enhances economic opportunities by promoting wider applications of deep learning into realistic systems with security guarantee, and gives special attention to educating women and students from under-represented/under-served groups.



Reference:

[1] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and **Wujie Wen**, "Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), June 2019, Long Beach, CA.

[2] S. Wang, X. Wang, P. Zhao, **W. Wen**, D. Kaeli, P. Chin, and X. Lin, "Defensive dropout for hardening deep neural networks under adversarial attacks", IEEE/ACM International Conference On Computer Aided Design (ICCAD), Nov. 2018.

[3] Q. Liu, T. Liu, Z. Liu, Y. Wang, Y. Jin and **W. Wen**, "Security Analysis and Enhancement of Model Compressed Deep Learning Systems under Adversarial Attacks," Proc. ACM/IEEE 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 2018, pp. 721-726. (**Best Paper Award Nomination**)

NSF-CNS 1840813
Wuije Wen, Lehigh University & Florida International University
Email: wuw219@lehigh.edu