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

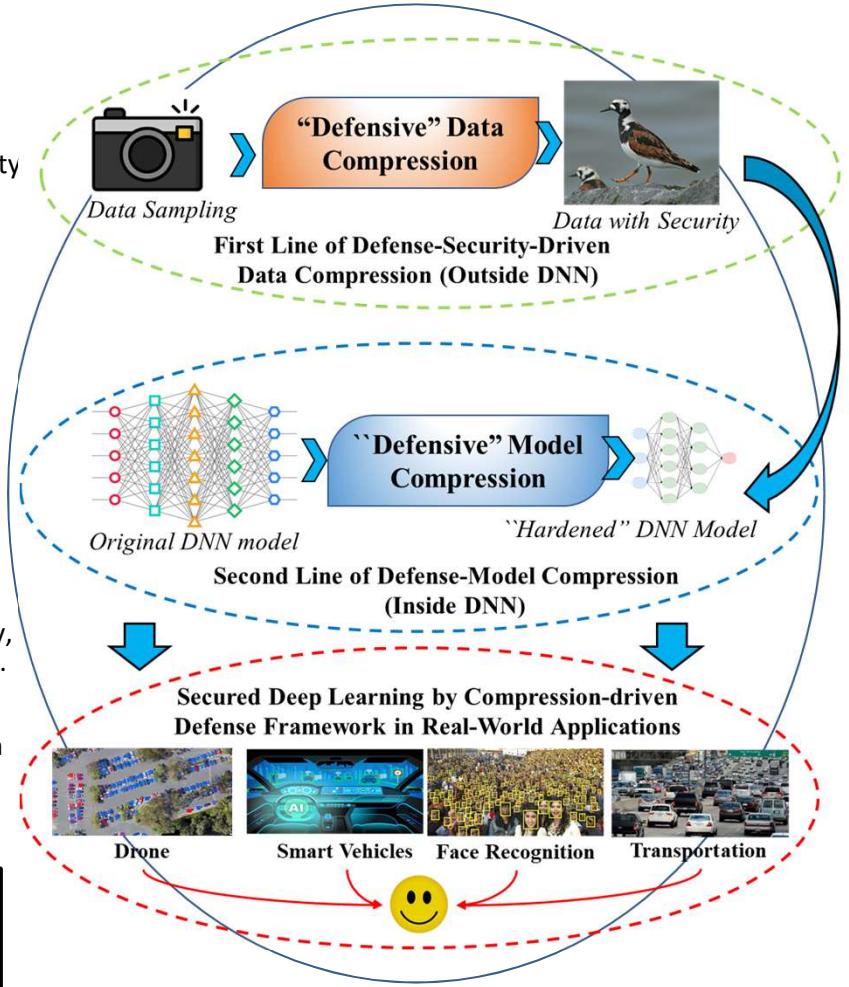
PI: Wujie Wen

Challenge:

- Deep Neural Network (DNN) is highly vulnerable to input data with imperceptible perturbations (adversarial examples) for decision making both theoretically and physically, causing a significant security concern for security-critical tasks.
- Defense becomes urgent but faces following difficulties: 1) Diversified attack natures; 2) Various attack strategies; 3) Low overhead and accuracy guarantee of benign data.

Solution:

- Integrating defense into low-cost compression, an essential component existing in both input data side and DNN model.
- 1) New concept-based input compression tailored for DNNs by jointly considering defense efficiency, accuracy and compression efficiency. 2) Security-driven model compression to harden DNN model; 3) New analytical models, evaluation metrics to measure deep learning security.



Scientific Impact:

- Create a new paradigm of safeguarding deep learning from a radically different perspective with a focus on integrating defenses into compression of the inputs and models—``Invisible Shield".
- Results will be useful to research community interested in data science, hardware- and cyber- security, and multimedia.

Broader Impact:

- 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.
- The project envisions technology transfer of the proposed techniques in the industrial development of secure deep learning systems.

NSF-CNS 1840813

Wuije Wen, Lehigh University & Florida International University
Email: wuw219@lehigh.edu