CAREER: Protecting Deep Learning Systems against Hardware-Oriented Vulnerabilities

Yingjie Lao; Clemson University

http://ylao.people.clemson.edu/hardware_AI_security

Artificial intelligence (AI) has recently approached or even surpassed human-level performance in many applications. However, the successful deployment of AI requires sufficient robustness against adversarial attacks of all types and in all phases of the model life cycle. Given this pressing need, this project aims at exploring novel hardware-oriented adversarial AI concepts and developing fundamental defensive strategies against such vulnerabilities to protect next-generation AI systems. The intellectual merits include (i) exploiting new algorithm-hardware adversarial attacks on deep neural network systems; (ii) developing methodologies that incorporate the hardware aspect into defense for enhancing adversarial robustness; and (iii) developing novel signature embedding and model recovery frameworks to protect the deep neural network models against hardware-oriented attacks.

Challenge:

- Although much progress has been made in enhancing the robustness of AI algorithms, there is a lack of systematic studies on hardware-oriented vulnerabilities and countermeasures, which also opens up demand for AI security education.

Scientific Impact:

- Advances the understanding of hardware-oriented vulnerabilities and countermeasures for AI systems.
- Further validates the need to protect both algorithm and hardware for next-generation AI applications.

Solutions:

1. DL Hardware Watermarking:

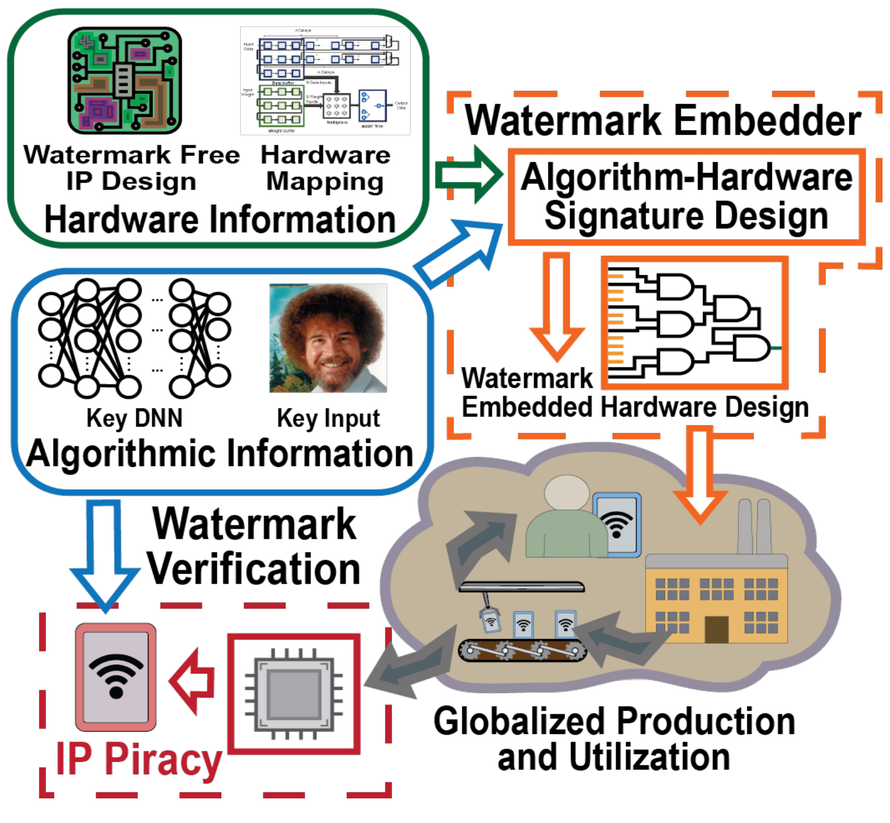
Developed the first hardware watermarking framework for protecting the hardware intellectual property (IP) of deep learning accelerators. (AAAI-22)

2. Clean-Label Poisoning Availability Attack:

Developed a novel attack methodology by using clean-label poisoned samples, which reveals a more serious threat to machine learning models compared to dirty-label attacks. (AAAI-22)

3. Other Aspects:

Studied the vulnerabilities of deep neural networks against more stealthy adversarial examples and poisoning attacks. (WACV-22, ICASSP-22)



DL Hardware Watermarking [1]

Broader Impact:

This project yields novel methodologies for ensuring trust in AI systems from both the algorithm and hardware perspectives to meet the future needs of commercial products and national defense.

Education:

- Results are integrated into my Creative Inquiry course “AI Security and Privacy”, which I teach in both spring and fall semesters.
- This project also participates the Clemson EUREKA! Program, a summer program for new Clemson honors students.

[1] J. Clements and Y. Lao, "DeepHardMark: Towards Watermarking Neural Network Hardware," AAAI-22.

[2] B. Zhao and Y. Lao, "CLPA: Clean-Label Poisoning Availability Attacks Using Generative Adversarial Nets," AAAI-22.

[3] B. Zhao and Y. Lao, "Towards Class-Oriented Poisoning Attacks Against Neural Networks," WACV-22.

[4] J. Clements and Y. Lao, "In Pursuit of Preserving the Fidelity of Adversarial Images," ICASSP-22.

Award ID#: 2047384