SaTC: Small: Understanding and Taming Deterministic Model Bit Flip Attacks in Deep Neural Networks

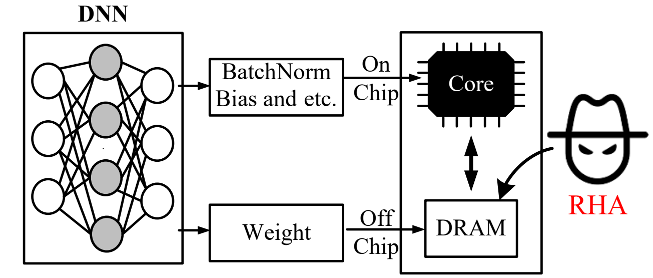
Challenges:

- How to model and characterize impacts of internal hardware fault attacks in modern machine learning (ML) models?
- Can we make ML models inherently robust to deterministic bit flip?
- How to combine algorithm-level mitigation with architecture/system protection mechanisms to offer holistic security against hardware-based ML model tampering?

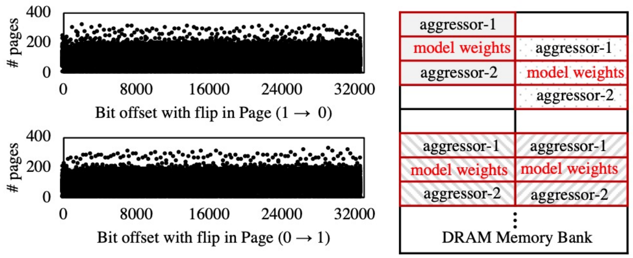
Solutions:

- An end-to-end model bit flip attack via rowhammer that hijacks inference of ML models (**USENIX Security' 20**).
- Novel algorithm to characterize the attack surface of ML model fault attacks with various adversarial objectives (**TPAMI' 21**).
- The first memory fault attacks that compromise neural machine translation in NLP models (**SEED'21**).
- A binarization-aware training method to enhance the robustness of DNN models against model bit flip (**CVPR'20**).

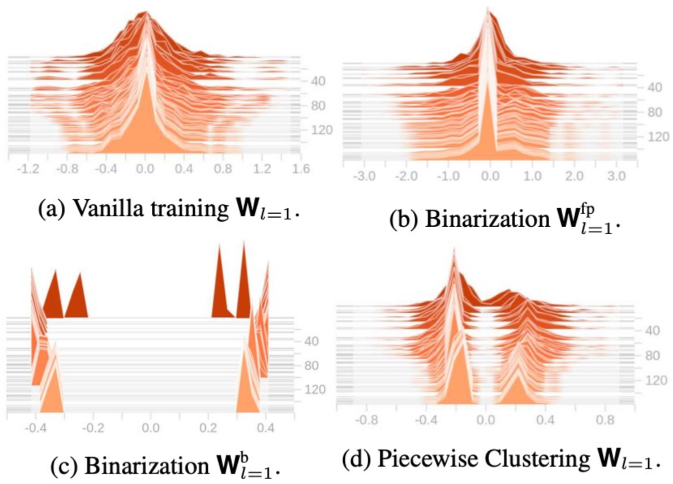
Projects: **SaTC-2019536** and **SaTC-2019548**.
PIs: Fan Yao (University of Central Florida), Deliang Fan (Arizona State University).



ML model bit flip attack vectors



DRAM Rowhammer fault injections in model weights



(a) Vanilla training $\mathbf{W}_{l=1}$.

(b) Binarization $\mathbf{W}^{\mathrm{fp}}_{l=1}$.

(c) Binarization $\mathbf{W}^{\mathrm{b}}_{l=1}$.

(d) Piecewise Clustering $\mathbf{W}_{l=1}$.

BFA-resistant model training

Scientific Impact:

- Several papers published in top ML, system/hardware security conferences and journals.
- Key research outcomes have led to a new and active research direction in adversarial machine learning, i.e., *adversarial weight attacks and defenses.*

Broader Impact and Participation:

- This project advances understandings of security of HW-based weight perturbations and proposes new SW/HW protections against them. This enables future research on secure and efficient ML systems against hardware-based model tampering.
- Research outcomes have been integrated into graduate courses related to hardware and ML security offered at UCF in both ECE and CS departments.
- Education and training on hardware security for AI have been offered to both masters and undergraduate students with internships and independent study.