Live demonstration: man-in-the-middle attack on edge artificial intelligence

Hu, Bowen; He, Weiyang; Wang, Si; Liu, Wenye; Chang, Chip Hong

2024

Hu, B., He, W., Wang, S., Liu, W. & Chang, C. H. (2024). Live demonstration: man-in-the-middle attack on edge artificial intelligence. 2024 IEEE International Symposium on Circuits and Systems (ISCAS). https://dx.doi.org/10.1109/ISCAS58744.2024.10558371

https://hdl.handle.net/10356/174146

https://doi.org/10.1109/ISCAS58744.2024.10558371

Live Demonstration: Man-in-the-Middle Attack on Edge Artificial Intelligence

Bowen Hu, Weiyang He, Si Wang, Wenye Liu, and Chip-Hong Chang
Centre for Integrated Circuits and Systems, School of Electrical and Electronic Engineering
Nanyang Technological University Singapore, Singapore
{bowen006, e210050, wliu015}@e.ntu.edu.sg, {si.wang, echchang}@ntu.edu.sg

***Abstract*—Deep neural networks (DNNs) are susceptible to evasion attacks. However, digital adversarial examples are typically applied to pre-captured static images. The perturbations are generated by loss optimization with knowledge of target model hyperparameters and are added offline. Physical adversarial examples, on the other hand, tamper with the physical target or use a realistically fabricated target to fool the DNN. A sufficient number of pristine target samples captured under different varying environmental conditions are required to create the physical adversarial perturbations. Both digital and physical input evasion attacks are not robust against dynamic object-scene variations and the adversarial effects are often weakened by model reduction and quantization when the DNNs are implemented on edge artificial intelligence (AI) accelerator platforms. This demonstration presents a practical man-in-the-middle (MITM) attack on an edge DNN first reported in [1]. A tiny MIPI FPGA chip with hardened CSI-2 and D-PHY blocks is attached between the camera and the edge AI accelerator to inject unobtrusive stripes onto the RAW image data. The attack is less influenced by dynamic context variations such as changes in viewing angle, illumination, and distance of the target from the camera.**

I. Introduction

To bring the inference closer to where the data is generated, dedicated hardware accelerators and development toolkits have been made commercially available to enable pre-trained deep neural network (DNN) models to be deployed on edge devices for real-time computer vision tasks such as face recognition and autonomous driving. These edge AI accelerators are usually connected directly to the camera to allow the inference to be made locally without being impacted by unreliable network connectivity. In most scenarios, the connection between the camera and the edge inference device is unprotected, which opens the door to new hardware-based attacks. This demonstration shows that through a discreetly mounted interface bridge, ephemeral adversarial patterns can be injected into the camera data lane between the image sensor and the AI processor. In this demonstration, the raw data of the camera is connected to the edge inference machine implemented by a Raspberry Pi 4B (RPi 4B) and an Intel Neural Computing Stick (NCS 2), through a high-speed mobile industrial processor interface (MIPI) CSI-2.

* This research is supported in part by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity Research & Development Programme (Cyber-Hardware Forensic & Assurance Evaluation R&D Programme NRF2018NCRNCR009-0001) and in part by the Ministry of Education, Singapore, under its AcRF Tier 2 Award No. MOET2EP50220-0003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

We use a Trion T20 MIPI FPGA chip that has built-in MIPI CSI-2 transmitter and receiver to intercept and modify the high-speed data stream. For rapid prototyping, a T20 development board is used as the MIPI bridge but a stealthier attack can be implemented with a tiny $3.6mm \times 4.5mm$ WLCSP80 packaged T20 chip. Using Algorithm 1 of [1], virtual adversarial light stripes are generated and injected onto the intercepted image data of the real target without having to project them directly onto the physical target to cause it to be misclassified by the DNN. This hardware-assisted attack can be performed online by tampering one scan line a time in less than $0.04ms$, making a real-time attack feasible without interrupting the image traffic. These adversarial stripe line patterns are sparse and require no additional off-chip memory to store.

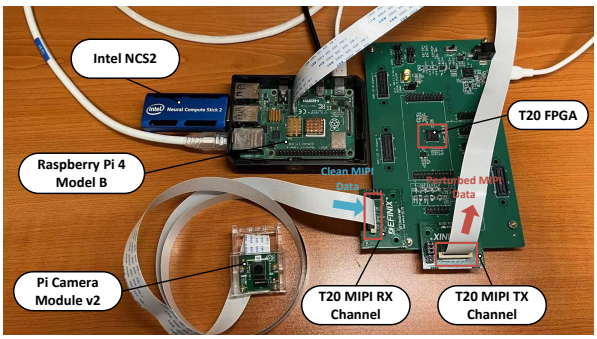

Fig. 1. Overview of the demonstration setup.

II. Demonstration Setup

The camera module is the Pi Cam V2, which is an 8-megapixel RGB camera. The raw image data is read by a RPi 4B device before it is forwarded to a DNN implemented on an Intel NCS2 for real-time inference. An Efinix Trion T20 MIPI FPGA is used as a MITM device to intercept the image data in transit, inject ephemeral adversarial patterns and relay the tainted image data to the edge DNN.

III. Visitors Experience

Visitors can place objects in front of the camera, and the edge device can classify the placed objects in real time and display the footage with the classification results on the monitor. The visitors can also participate as attackers by activating an attack pattern to the camera MIPI data lanes with a single button press to cause a misclassification.

References

[1] W. Liu, W. He, B. Hu and C. -H. Chang, "A Practical Man-in-the-Middle Attack on Deep Learning Edge Device by Sparse Light Strip Injection into Camera Data Lane," 2022 IEEE 35th International System-on-Chip Conference (SOCC), Belfast, United Kingdom, 2022, pp. 1-6.