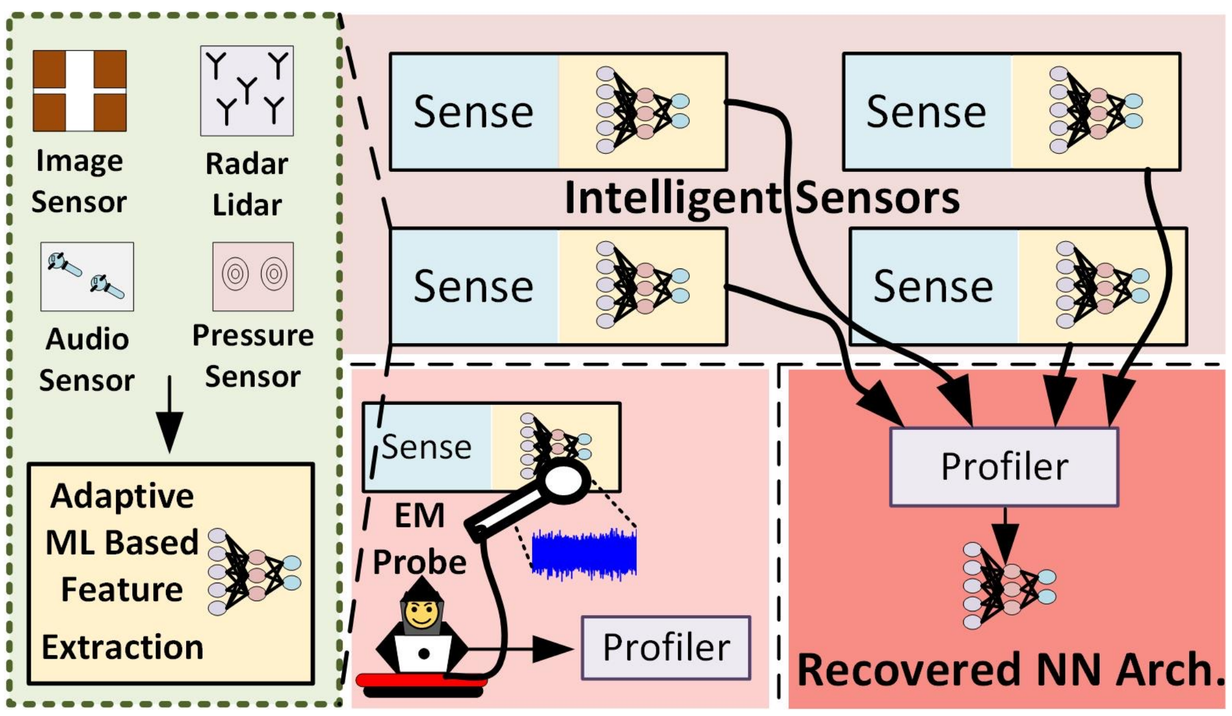
SNATCH: Stealing Neural Network Architecture From ML Accelerator in Intelligent Sensors

Sudarshan Sharma, Uday Kamal, Jianming Tong, Tushar Krishna and **Saibal Mukhopadhyay**

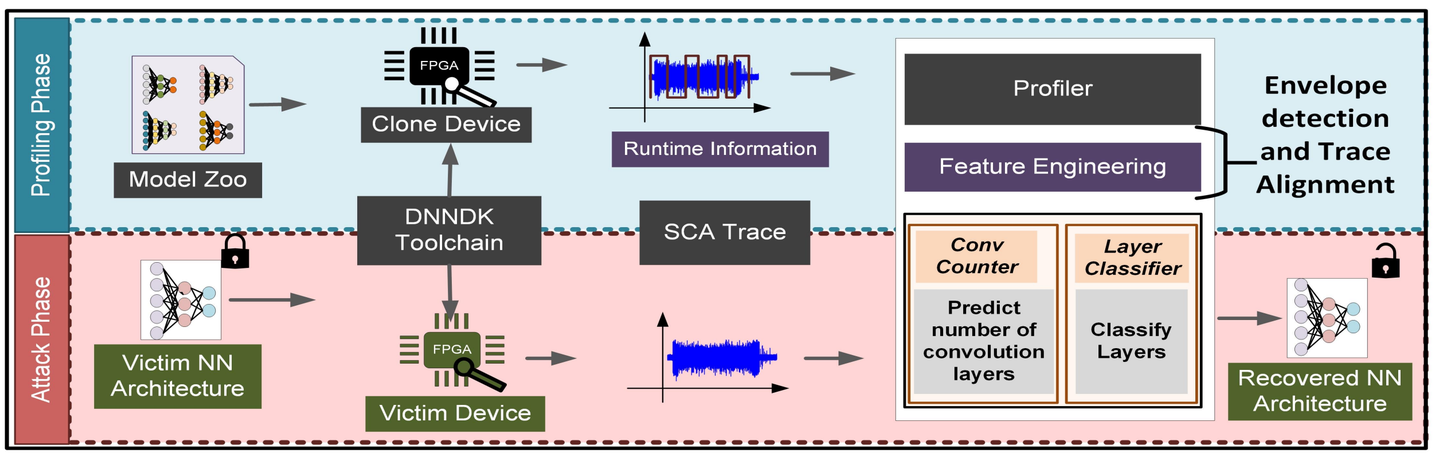
CogniSense
CENTER ON COGNITIVE MULTISPECTRAL SENSORS

I. Motivations

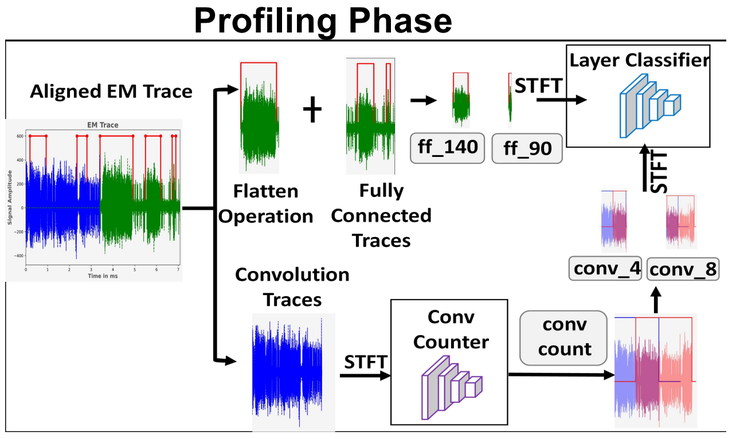


- Intelligent Sensors includes ML based feature extractors running on specialized ML Accelerators (MLA).
- The intended use of sensor data is encoded in the ML model running on these MLAs
- In general, HW implementation of ML model are believed to be more secure compared to the software-based-approaches.
- We demonstrate a profiling-based side channel attack (SNATCH) that can extract NN architecture even if they are directly implemented in a proprietary hardware

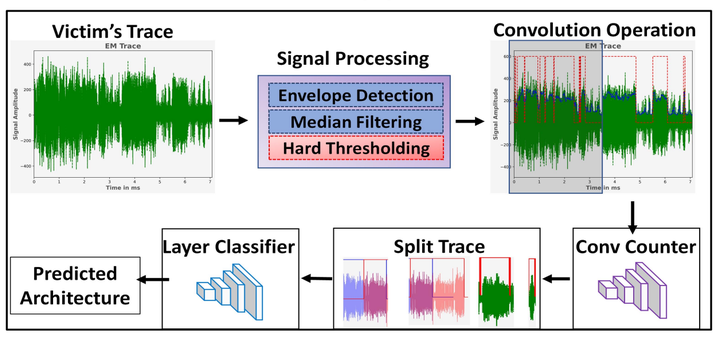
II. Attack Methodology



Profiling Phase



Attack Phase



- Attacker generates multiple NN architecture, compiles them using DNNDK toolchain and collects EM side-channel traces on the clone device
- Attacker uses the logged side channel traces to trained a profiler which consists of two CNN models ConvCounter and Layer Classifier
- In the attack phase, the attacker uses this profiler to reverse engineer the victim's NN architecture layer-wise based on the leaked EM side channel traces

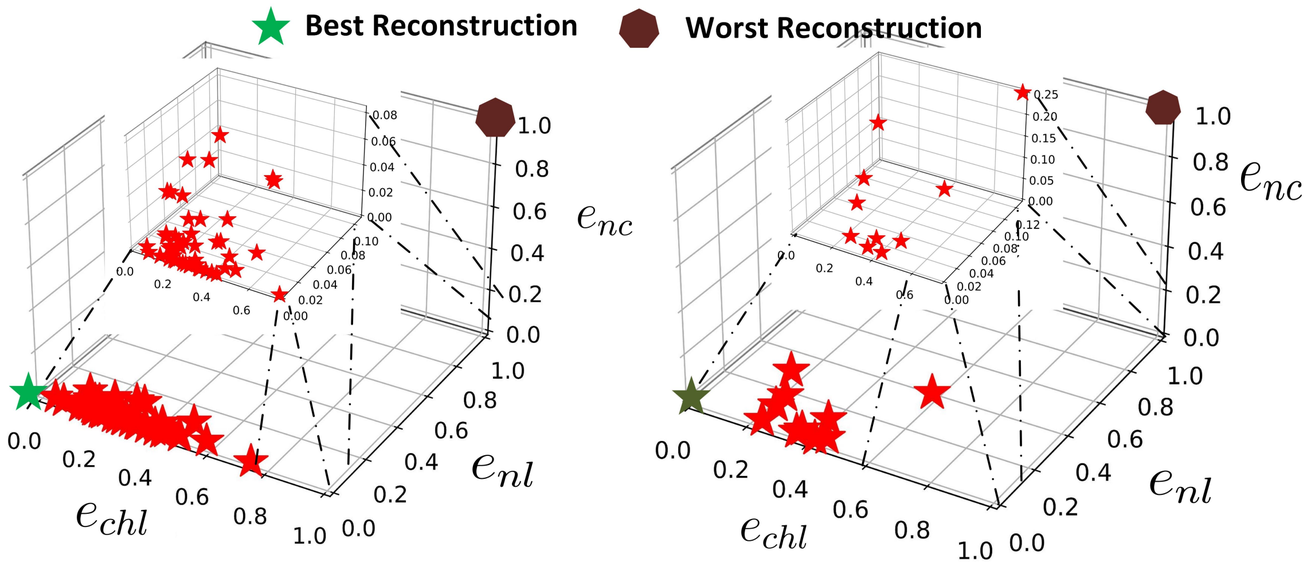
III. Profiler Accuracy

Conv Counter and Layer Classifier Prediction Accuracy

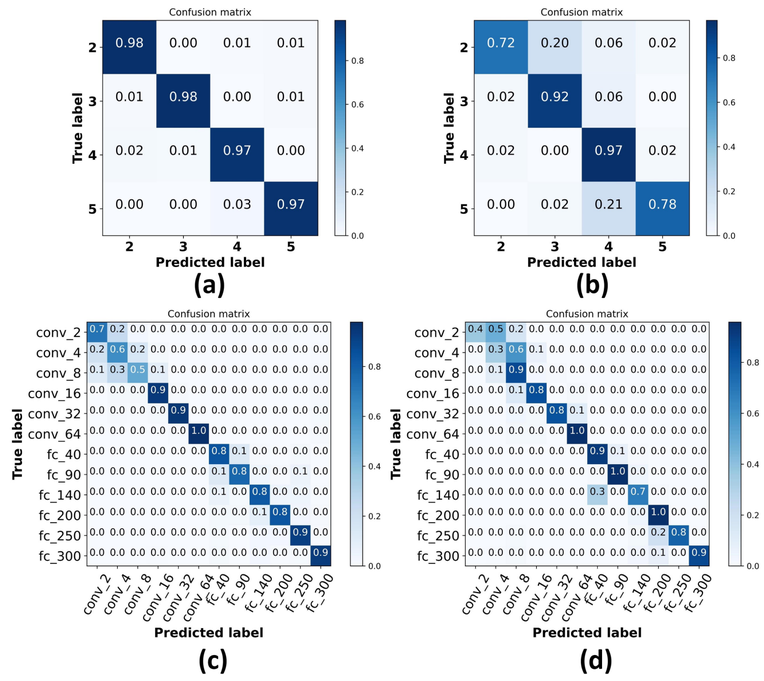
Dataset		Conv Counter Prediction Acc.		Layer Classifier Prediction Acc.	
		Profiling Acc. (Device-I) %	Attack Acc. (Device-II) %	Profiling Acc. (Device-I) %	Attack Acc. (Device-II) %
Same-Diff. Device	MNIST	95.3	97.4	78.4	72.0
	CIFAR10	95.3	96.5	87.9	80.6
		Profiling Acc. on Profiled Set %	Attack Acc. on Unseen Set %	Profiling Acc. on Profiled Set %	Attack Acc. on Unseen Set %
Seen-Unseen Arch.	MNIST	100.0	99.0	75.6	76.9
	CIFAR10	97.8	83.1	84.1	79.2

IV. Reconstruction Error

- e_{nc} measures error in predicting the number of convolution layer
- e_{chl} represents the error in predicting the number of convolution channels and hidden neurons
- e_{nl} quantifies the error in layer prediction



IV. Profiler Confusion Matrix



- Confusion Matrix of models for CIFAR10 (a) Conv Counter (c) Layer Classifier for Same-Different Device (b) Counv Counter (d)Layer Classifier for Seen-Unseen Architecture

V. Conclusions

- We demonstrate that the EM activity associated with the execution of a layer on the MLA reveals information about the layer, which can be exploited to steal the NN architecture.
- The attack can mount Denial of Service and various misuse attack on the Intelligent sensor using the stolen NN architecture.

This work was supported in part by CogniSense, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Semiconductor Research Corporation