Security of Embedded AI
Side-channel attacks for input extraction

Maria Méndez Real
Chaire Professeur Junior
Meutes de Drones Maritimes ***Autonomes*** *et de* ***Confiance***
maria.mendez-real@univ-ubs.fr

ARCHI 2025

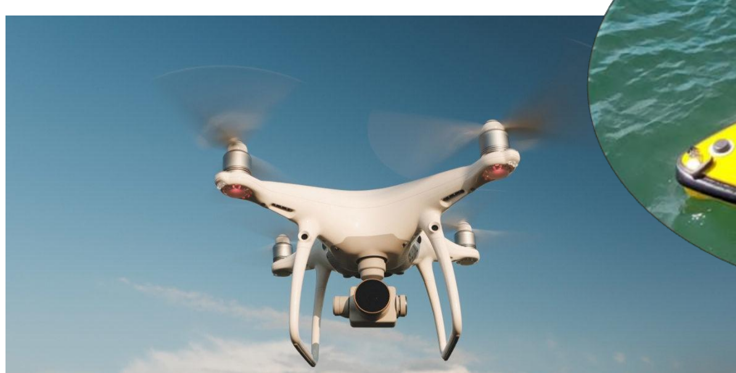
Contents

- Motivating security of embedded Neural Network

AI in todays (critical) systems

- Autonomous navigation
- Cable detection/following
- Mine detection
- Surveillance, traffic monitoring
- Intrusion detection systems
- …

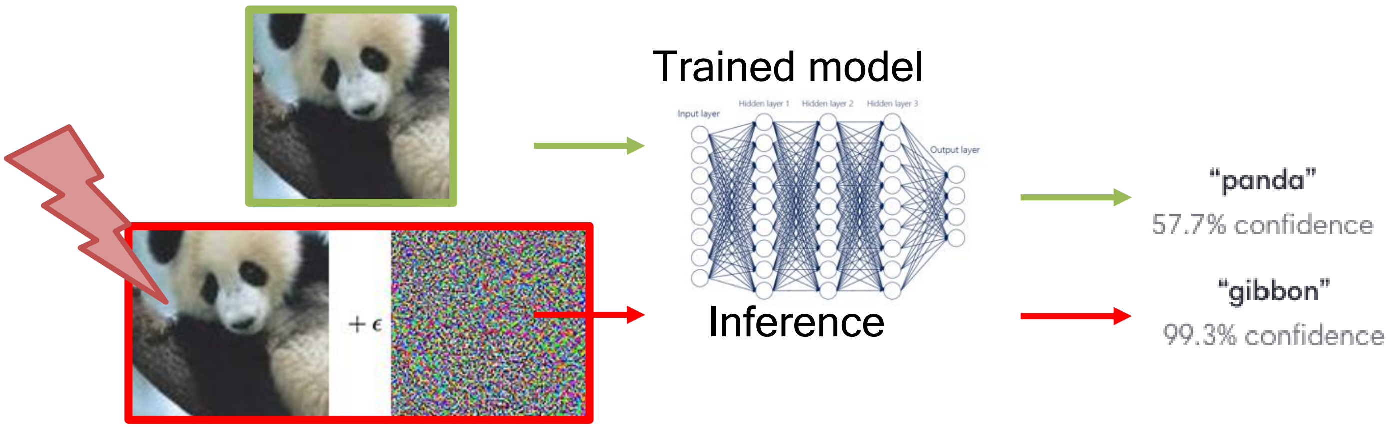
=> but what can go wrong?



Lab-STICC

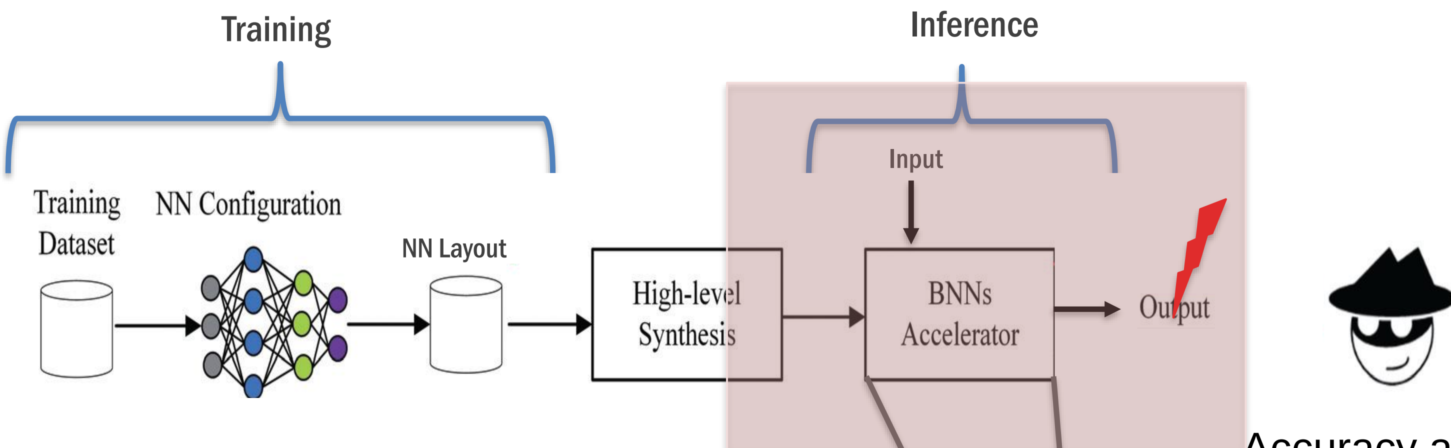
What can go wrong?

- Misclassification



Adversarial example

Y. Zhong, et al., Adversarial Learning with Margin-based Triplet Embedding Regularization, CCV'19

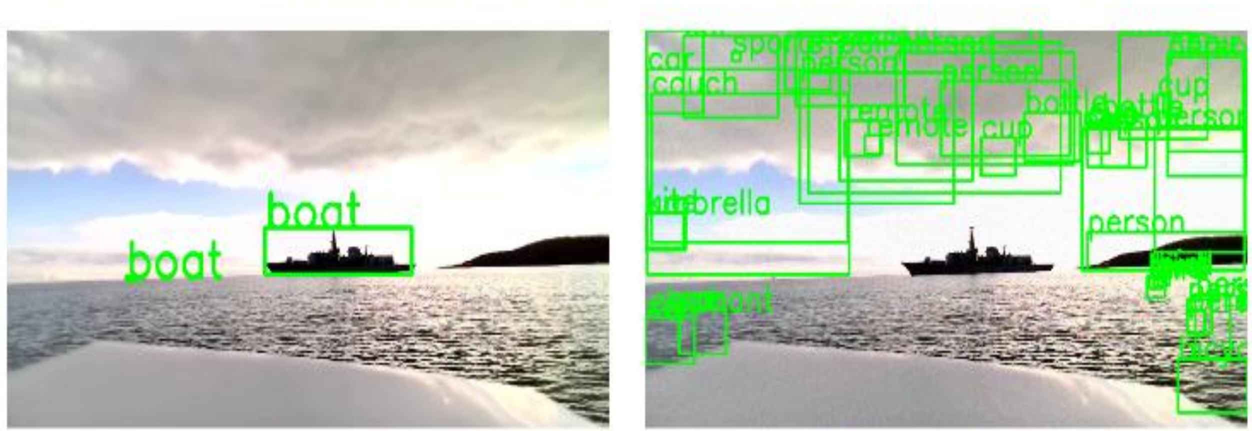


- Accuracy abuse
- Performance abuse: energy/delay response abuse

Lab-STICC

What can go wrong?

- Missclassification
  - Adversarial examples in maritime autonomous systems
  - Real physical scenarios?



(a) Original Input

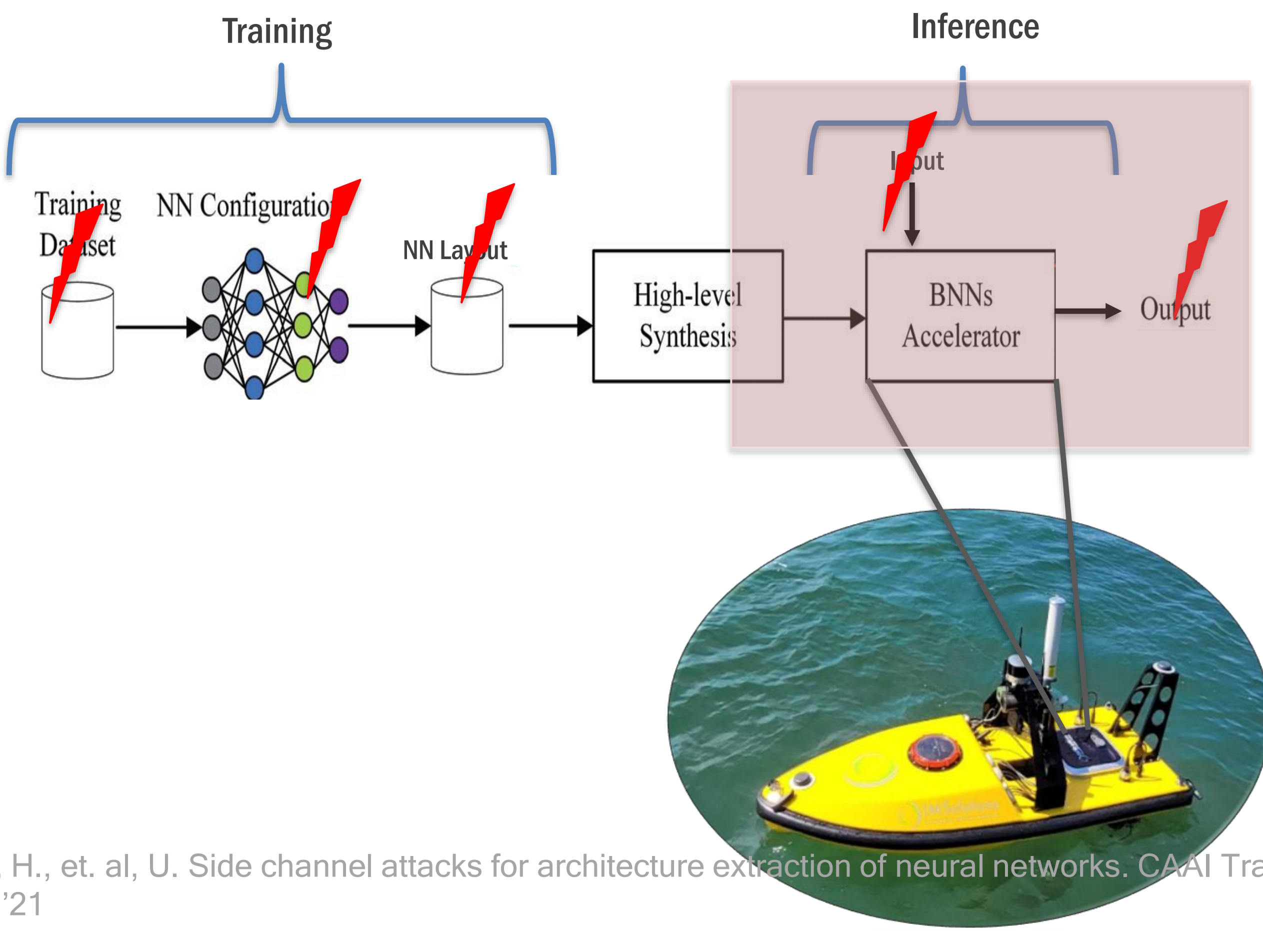
(b) Adversarial Image

Wang, Y., et. al., Towards a physical-world adversarial patch for blinding object detection models. Information Sciences, 556, 459-471.'21

Tam, K, et. al.,. 'Adversarial AI Testcases for Maritime Autonomous Systems',AI, Computer Science and Robotics Technology,'23

What can go wrong?

- Since, a panoply of attacks on **different vulnerable assets**



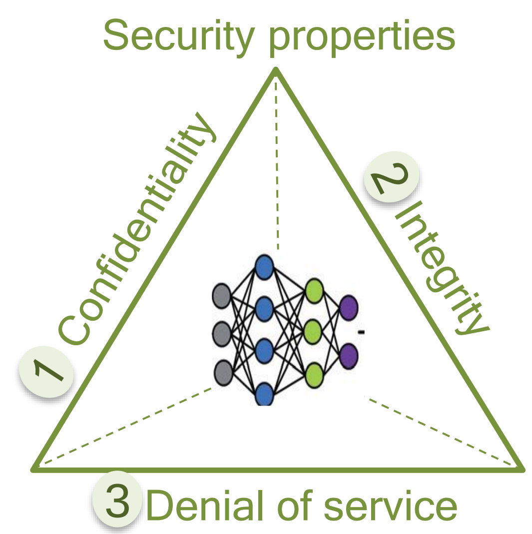
- Chabanne, H., et. al, U. Side channel attacks for architecture extraction of neural networks. CAAI Transactions on Intelligence Technology,'21
- M. Méndez Real, et al., Physical Side-Channel Attacks on Embedded Neural Networks: A survey, AS, Side-Channel Attacks Special Issue'21

Lab-STICC

What can go wrong?

- Misclassification 3 2
- No response ontime 3
- IP theft 1
- **Private data theft/disclosure** 1

$\Rightarrow$ unreliable AI,
$\Rightarrow$ mission failure,
$\Rightarrow$ collateral damage,
$\Rightarrow$ data disclosure,
$\Rightarrow$ money loss



Lab-STICC

- ANSSI, Sensibilisation et initiation à la cybersécurité. CyberEdu, notions de base'15
- P-A. Moellic, et al. Security of Software Embedded Neural Network Models: State of the Art and Threat Modeling. Tech. Report'21

Objective (long term)

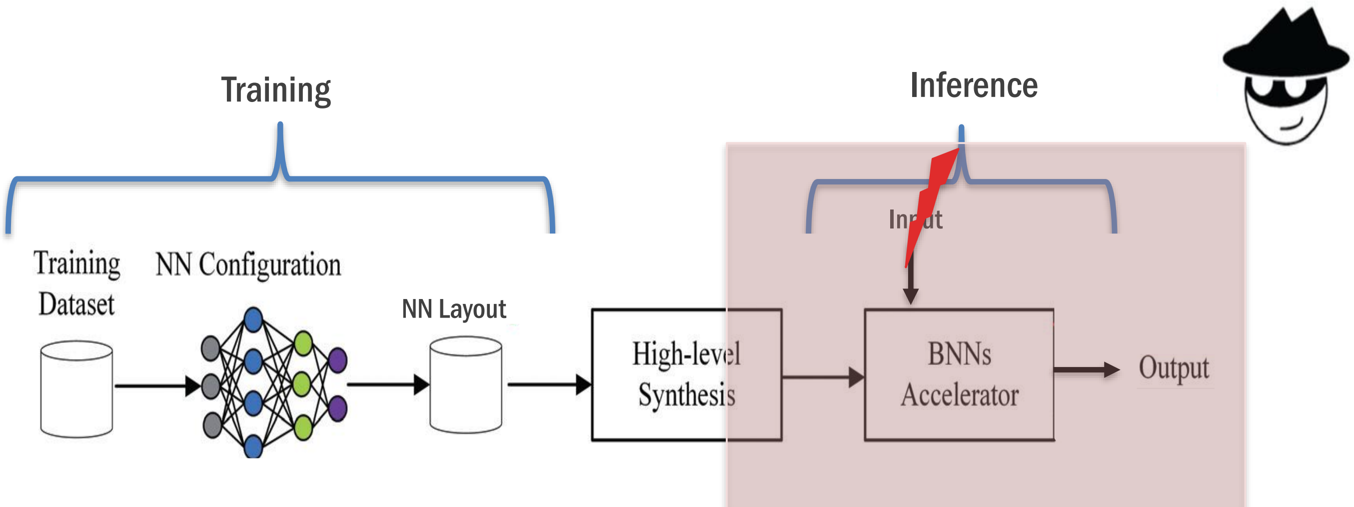
Are CNN models intrinsically different/vulnerable/robust to SCA vulnerabilities?

How can the target and implementation choices impact CNN security vulnerabilities?

Can CNN security vulnerabilites be evaluated/measured?

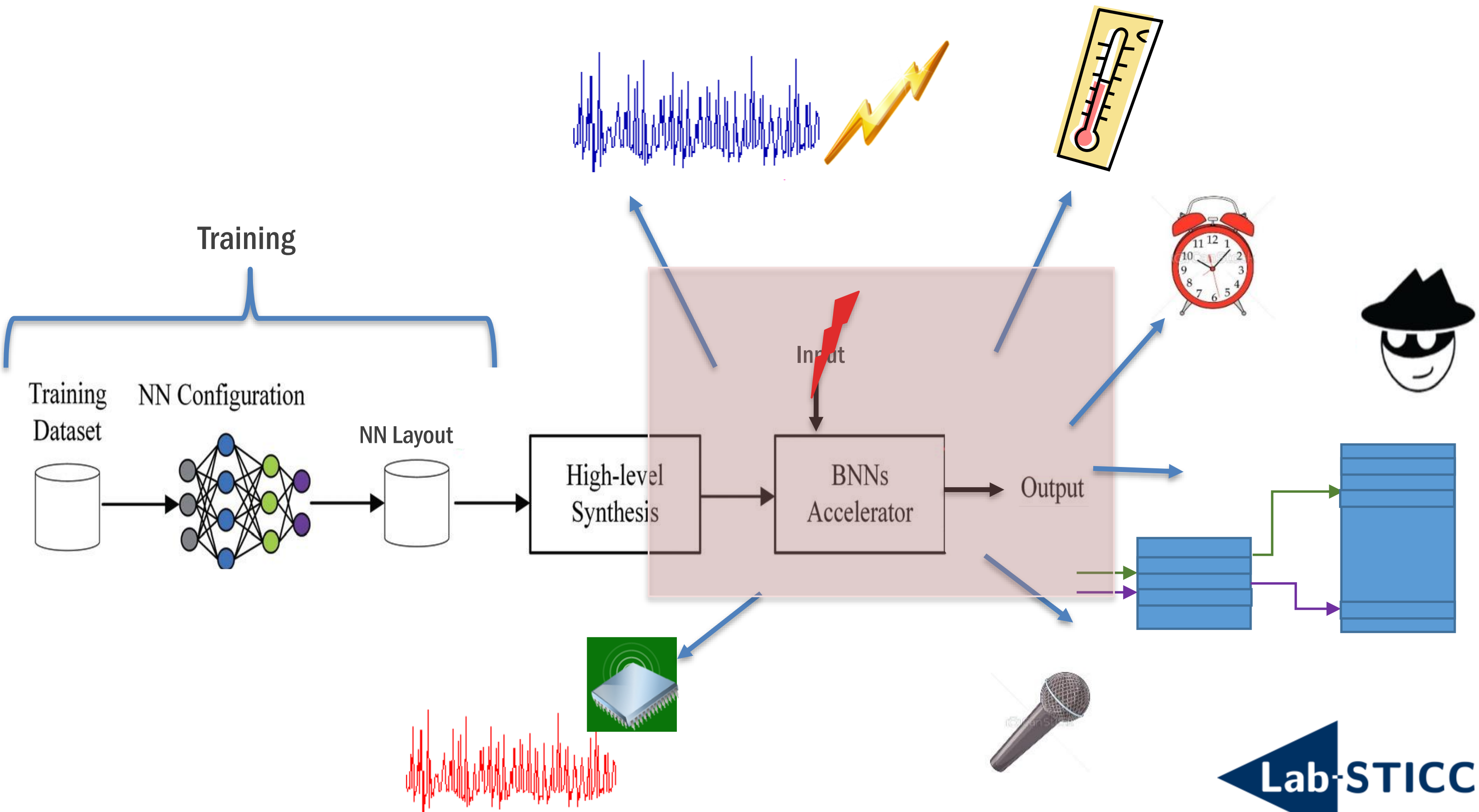
Focus on privacy attacks

- By observing **side-channel information**, can secret/private information be deduced?
  - Private information: inference inputs



Side-Channel Information

- Electromagnetic emissions, power consumption
- Are NN just as vulnerable as crypto?



AI vs Crypto

- AI vs Crypto
  - Secret asset: Secret key (256 bits) vs images
  - Leakage assessment metrics

- Threat model specificities
  - Crypto: public crypto algorithms

    $\Rightarrow$ Possible to hypothesize on intermediate results …

Disclaimer

- This talk is not about AI
- This talk is about (some) security vulnerabilities of AI accelerators

Contents

- Motivating security of embedded Neural Network
- Input extraction-vulnerability at the SoC level?

Input extraction - vulnerabilities at the SoC level?

- First work on black box scenario

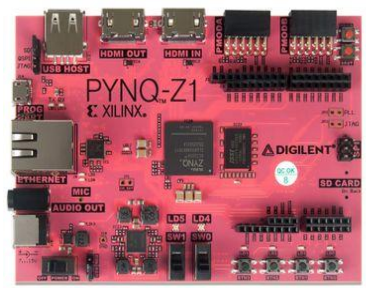
- Threat model
  - Black box, no interaction with the victim NN
  - Physical proximity to the target
  - **EM traces available to the attacker**

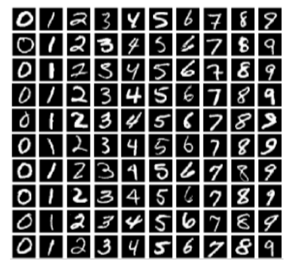
- Deducing secret/private information from the victim EM signature

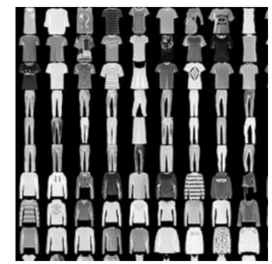
Thu, May Myat, M. Méndez Real et al. You only get one-shot: Eavesdropping input images to neural network by spying soc-FPGA internal bus. Conference on Availability, Reliability and Security'23.

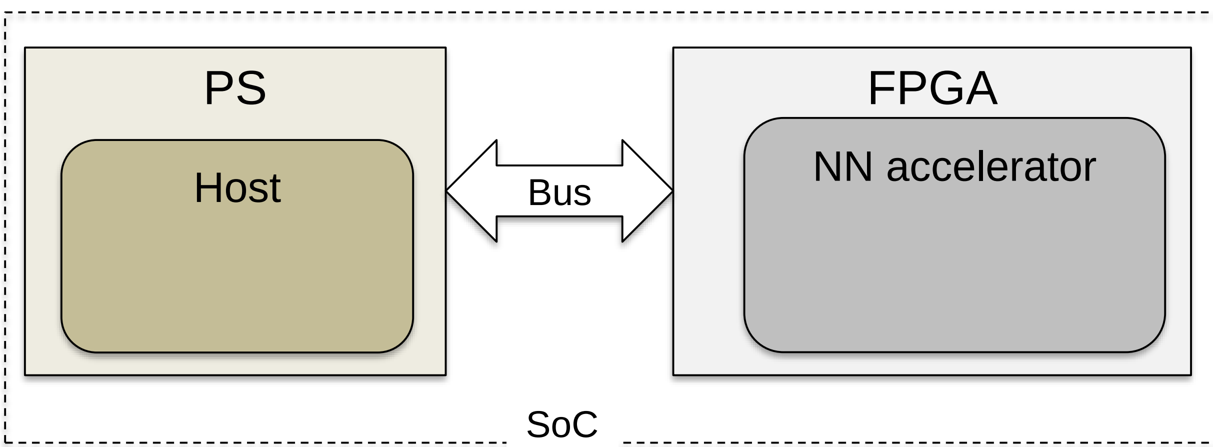
System considered and setup

- Xilinx FINN to implement LeNet trained on MNIST dataset on Zynq-7000 SoC (A9+FPGA)

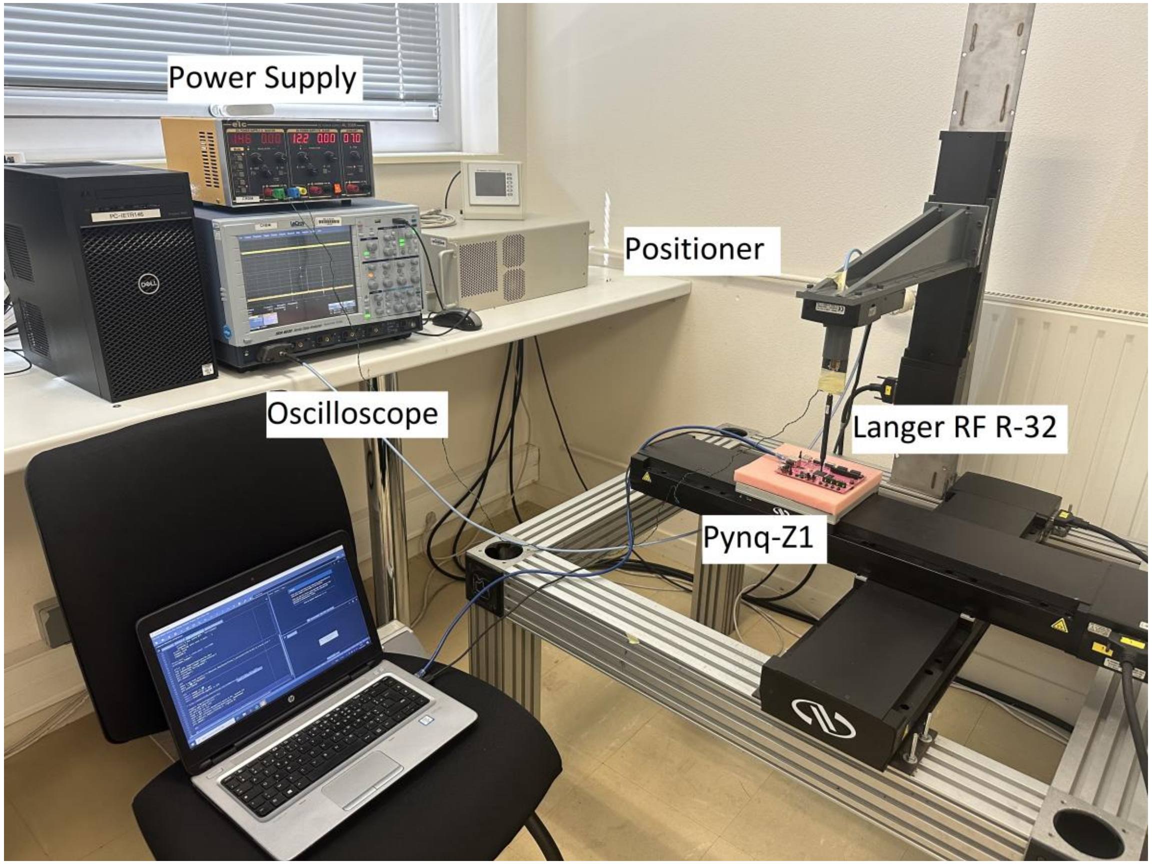


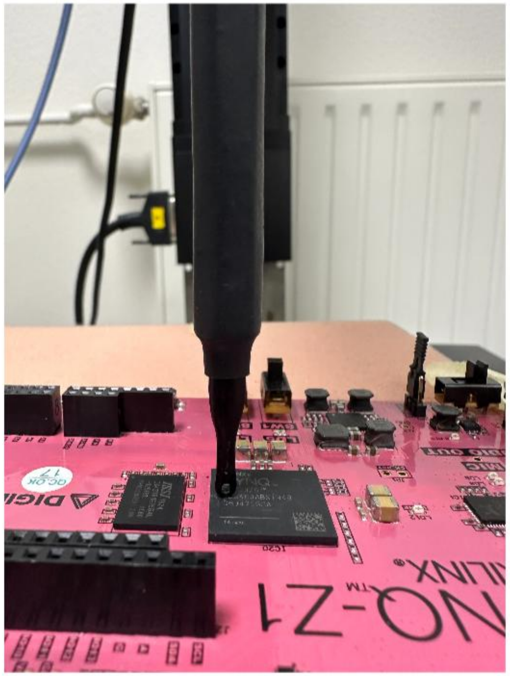




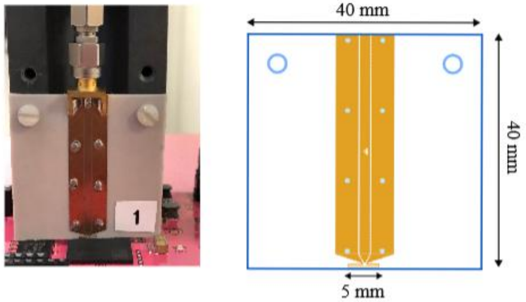


System considered and setup





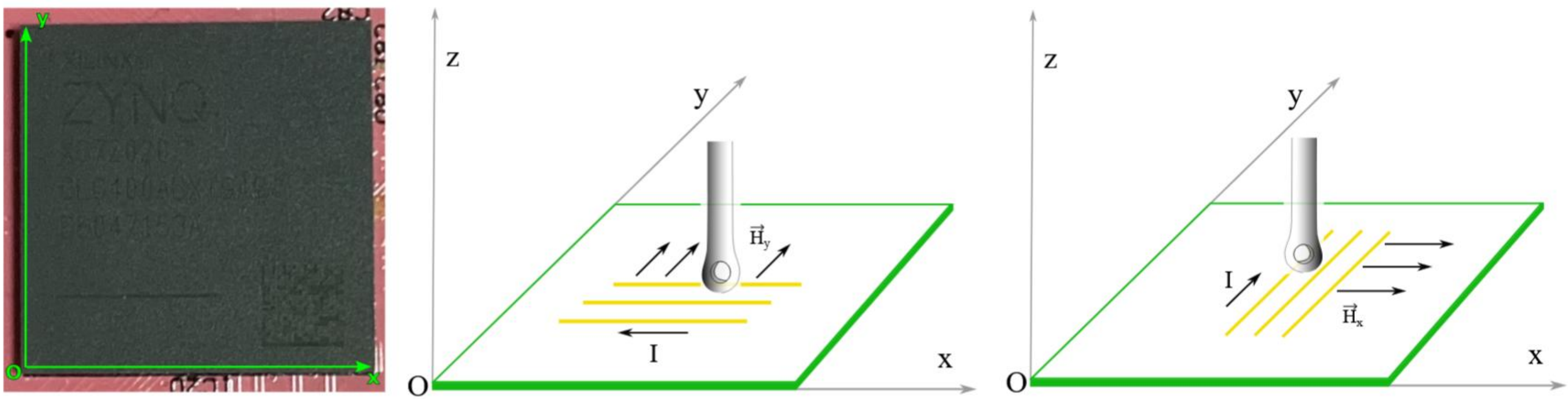
H-field probe



(a) (b)

Leakage localization

- EM cartography
- Data transfer on current bus wires

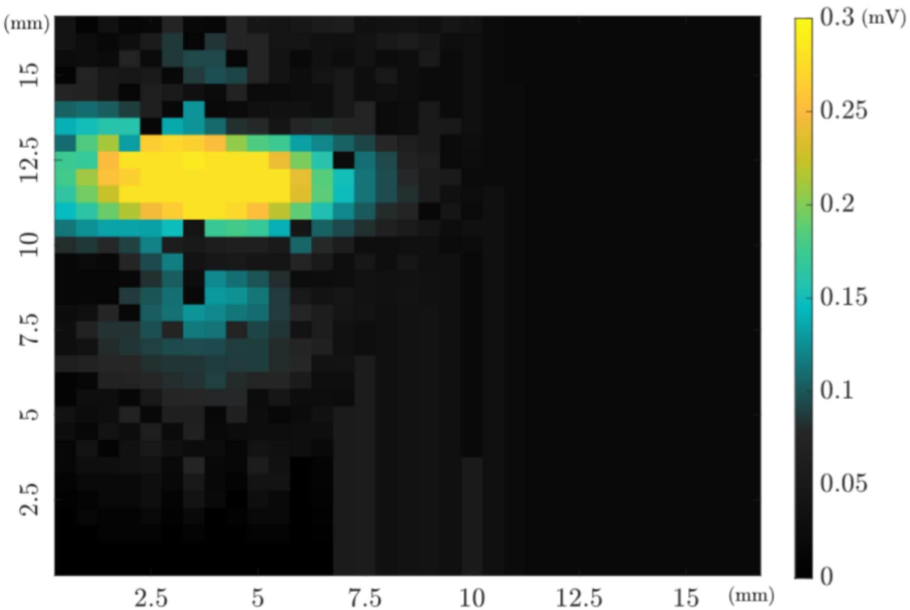


0.1 mm distance, with a spatial resolution of 0.5 mm

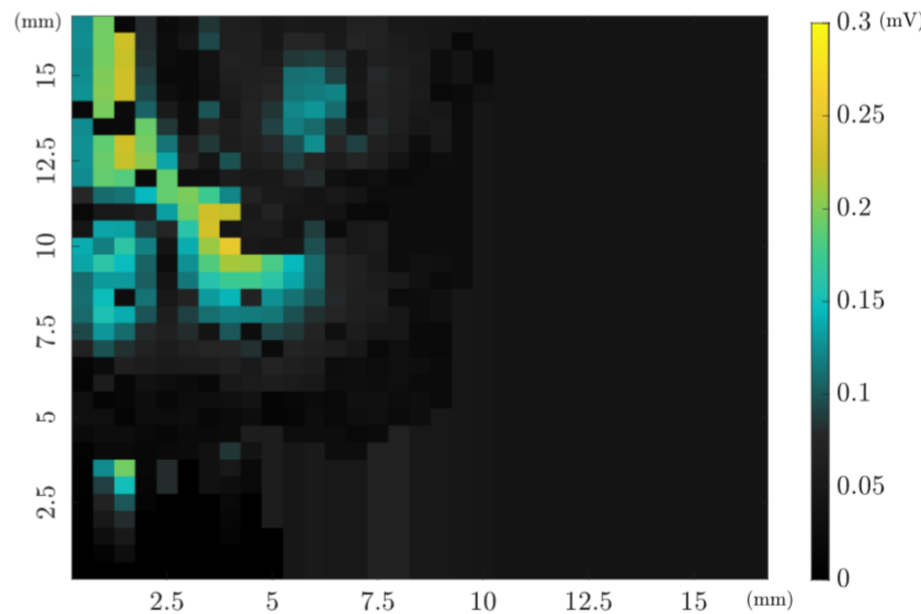
Thu, May Myat, M. Méndez Real et al. Bus electrocardiogram: Vulnerability of SoC-FPGA internal AXI bus to electromagnetic side-channel analysis. IEEE Compatibility-EMC Europe'23.

Leakage localization

- EM cartography
- Data transfer on current bus wires, AXI 32-bit bus

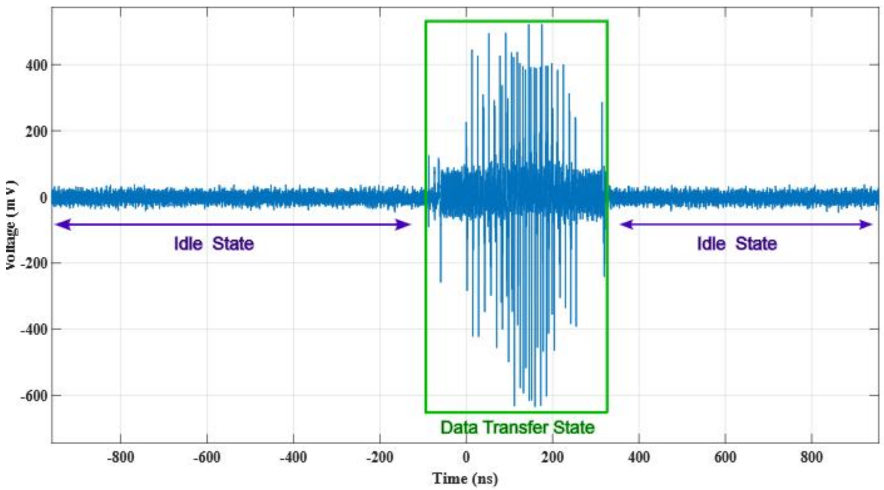


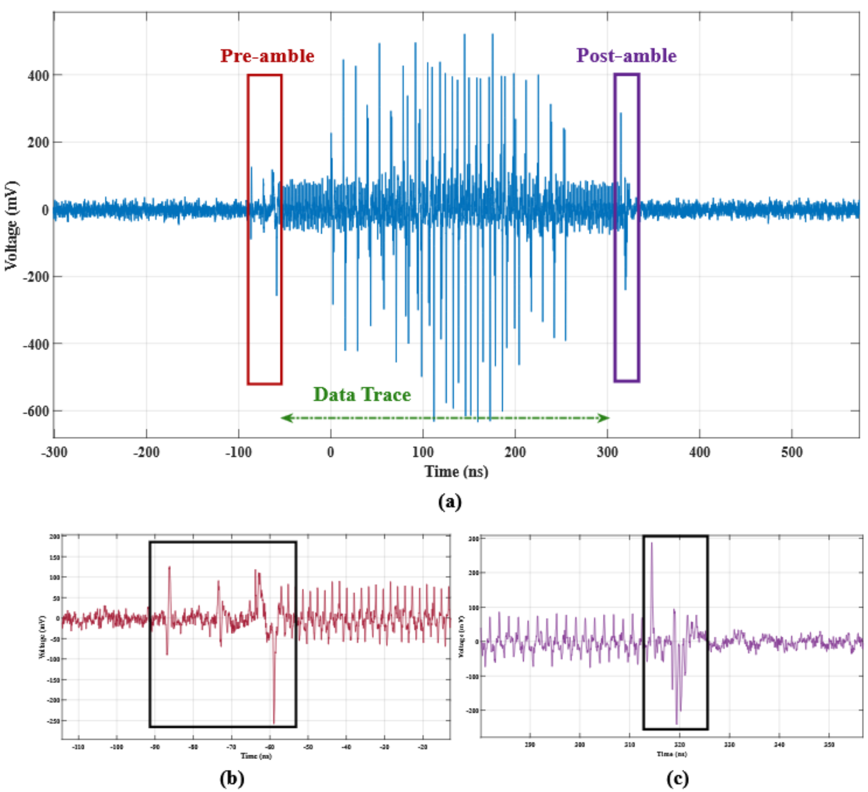
Component Hy



Component Hx

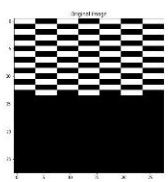
- Bus protocol easies synchronization



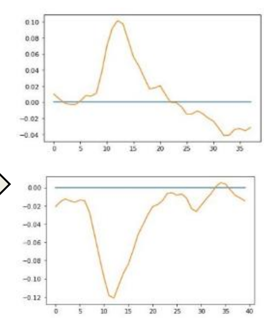


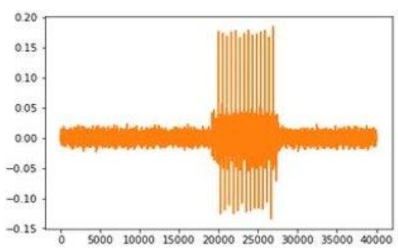
Intuition and leakage model

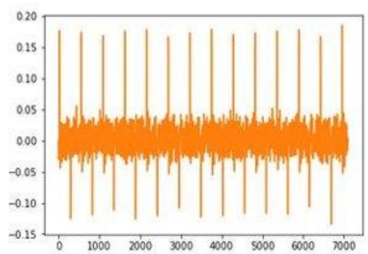
- The bus activity revealed by the EM emanations is proportional to the hamming distance HD
  - *signed HD* (upper/lower bit transitions)



Signed transitions







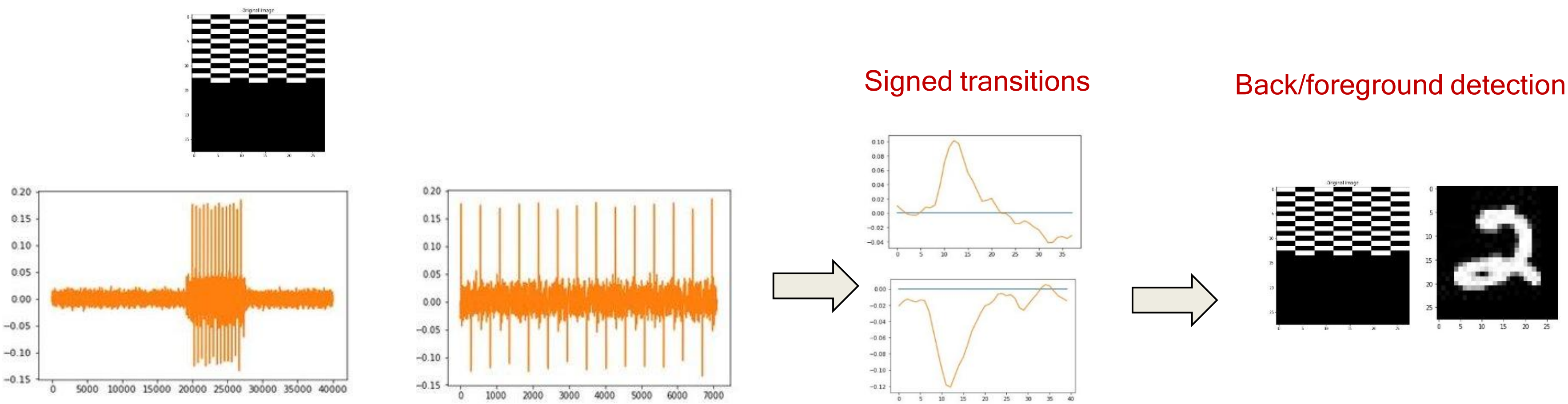
EM traces -> activity on the bus

Intuition and leakage model

- Simply define a threshold to differentiate the set of the majority of the pixels in a X-pixel group
- deducing the difference between neighbouring image pixels

  -> Back vs foreground pixels

  -> Single EM trace

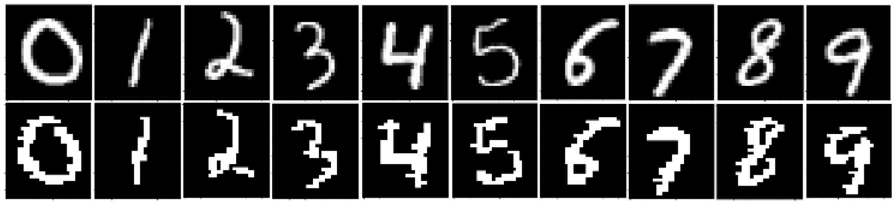


Signed transitions

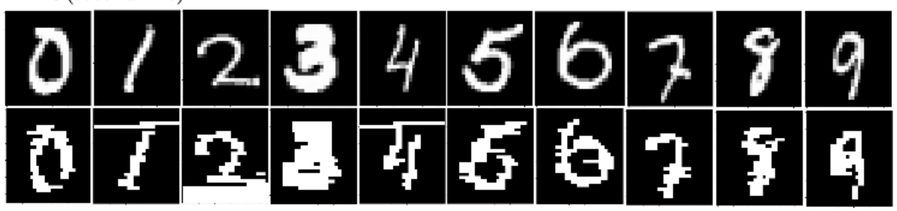
Back/foreground detection

EM traces -> activity on the bus

Some results



(a) Example of original MNIST images (first line) with highly accurate recovered images through HBIR (second line)

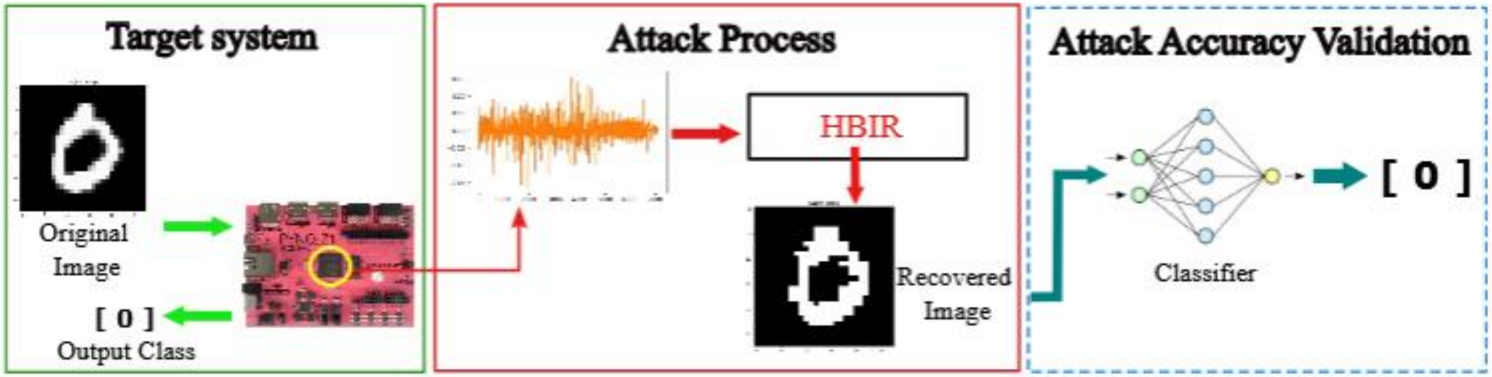


(b) Example of original MNIST images (first line) with less accurate recovered images through HBIR (second line)

- How to evaluate?

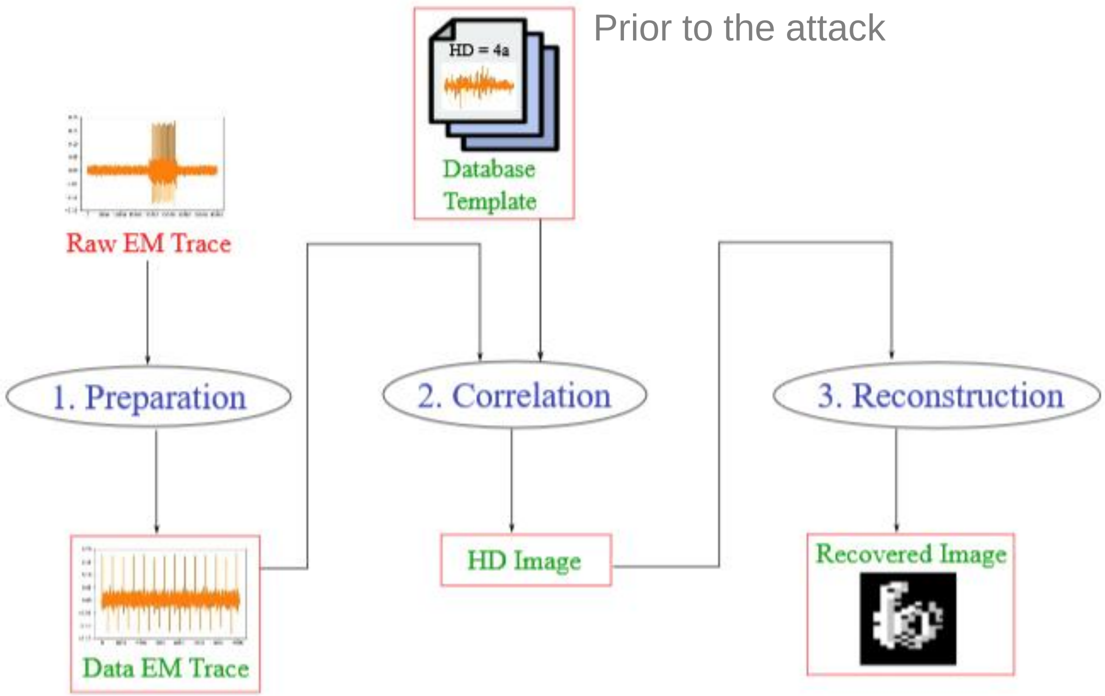
Some results

- 83,69% accuracy on the class of recovered images (vs 89%)



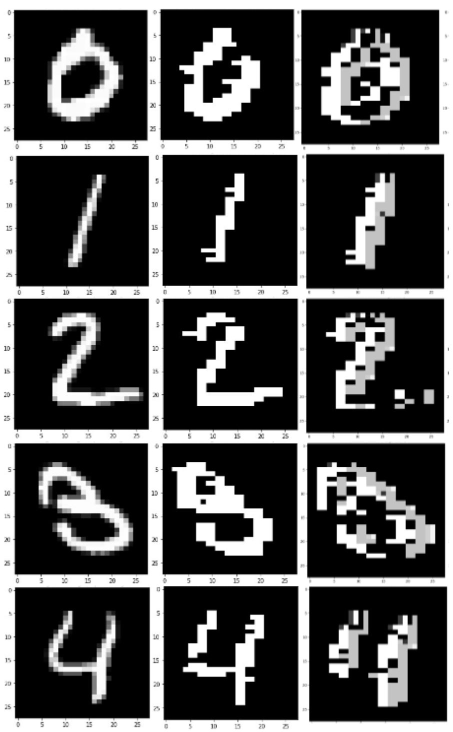
Can we go further?
-> template-based attack

- Can we deduce input data characteristics?
- (Oriented) Hamming distance between two consecutive cycles?
- Threat model
  - An access to a similar victim target is assumed prior to the attack
- Building a template attack? -> HD of 4/8/12/16

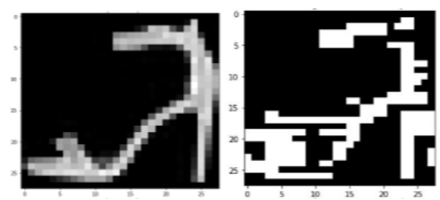


Some results

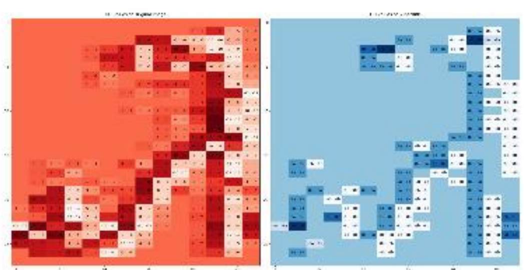
- Tests on MNIST and Fashion MNIST
- How to evaluate?
  - average HD difference: 12,5% MNIST, 30% Fashion MNIST
  - recognition accurary: similar for MNIST, better for fashion MNIST (73% vs 60%)



Original Image HBIR Template-based



Example of horizontal attack-based reconstruction on fashion MNIST



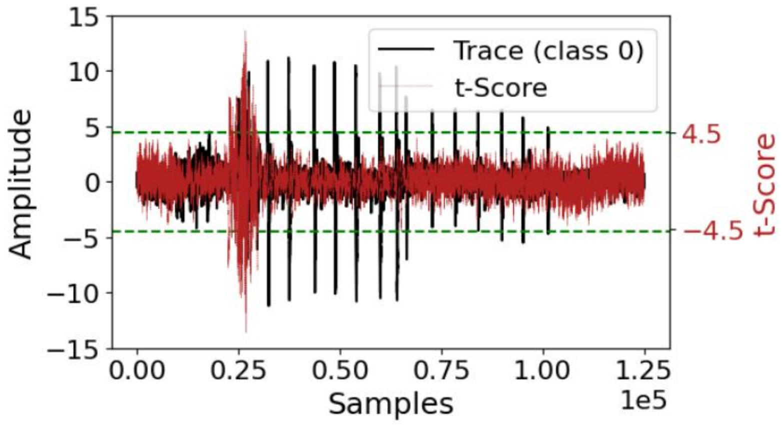
Example of template-based reconstruction on fashion MNIST

Some remarks

- EM activity on data transfer applied to NN accelerators
- Highly depends on the dataset
- Interesting to enhance the accuracy at the pixel value?
- Still a controlled environment
- What about power?

Similar but CNN-based attack

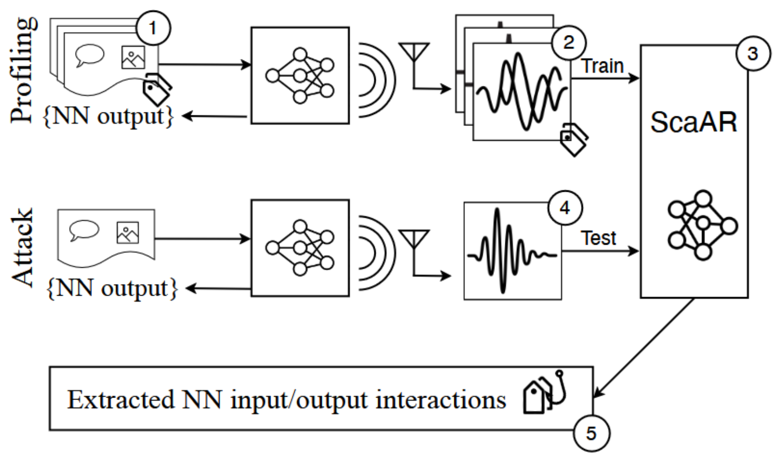
- Threat model and objective
  - Similar: EM, physical access
- Localizing the leakage
  - Statistical approach: Test-Vector Leakage Assessment (fixed vs random)



Liu, Zhuoran, et al. "Real-world Edge Neural Network Implementations Leak Private Interactions Through Physical Side Channel." arXiv preprint arXiv:2501.14512 (2025).

Similar but CNN-based attack

- Threat model and objective
  - Similar: EM, physical access
- Approach:
  - Test-Vector Leakage Assessment
  - Directly training a classifier on EM traces (4layer, 1dimensional CNN model)



Liu, Zhuoran, et al. "Real-world Edge Neural Network Implementations Leak Private Interactions Through Physical Side Channel." arXiv preprint arXiv:2501.14512 (2025).

Set up and some results

- Set up: ZCU104
- Results on LeNet5: 13,77% less accuracy (vs 6% in our simple image processing methodology)

Original accuracy

	MNIST	CIFAR-10	ImageNet-10
MLP	84.2	54.3	-
CNN3	98.8	-	-
LeNet5	$\boxed{86.4}$	72.4	-
SqueezeNet	-	91.1	58.1
ResNet18	90.8	$\boxed{94.1}$	69.8

Recognition accuracy on reconstructed images

Device	Implementation	MNIST	CIFAR-10	IMN-10
ZCU104	MLP	45.6	54.7	-
	CNN3	80.0	-	-
	LeNet5	$\boxed{74.5}$	-	-
	SqueezeNet	-	-	64.5
	ResNet18	92.8	$\boxed{89.4}$	-
RPi3B	LeNet5	-	96.3	-

Liu, Zhuoran, et al. "Real-world Edge Neural Network Implementations Leak Private Interactions Through Physical Side Channel." arXiv preprint arXiv:2501.14512 (2025).

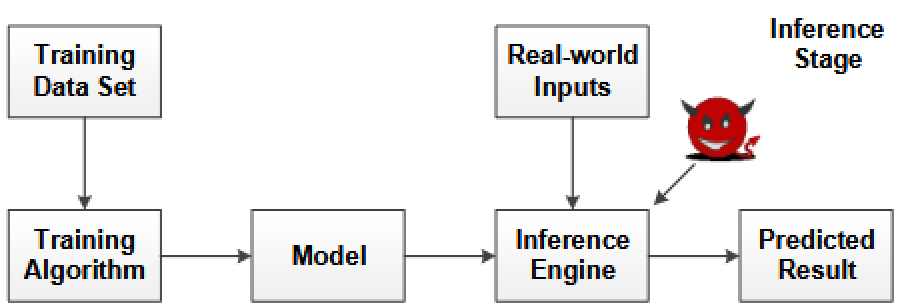
Lab-STICC

Contents

- Motivating security of embedded Neural Network
- Input extraction-vulnerabilites at the SoC level?
- Input extraction – within the acceletor -> exploiting an implementation choice

Input extraction – within the accelerator
-> exploiting an implementation choice

- Threat model and objective
  - Physical proximity to the target
  - **Power traces available**
  - No interaction with the victim NN
  - **Knowledge/hypothesis on NN implementation details**
  - **Assumes same input is infered several times (noise reduction)**



L. Wei et al. I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators. ACSAC'18

Intuition

- Deducing secret/private information from the dynamic power signature
- **Convolution unit drives power consumption**
- How convolution is implemented?
- Can we observe intermediate values in the convolution unit and correlate them to a power consumption model?

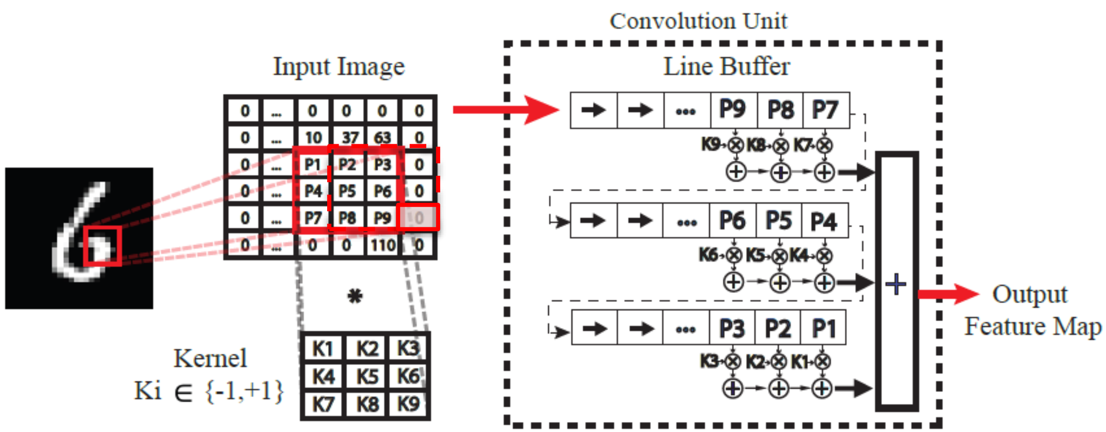


Fig. 7: Detailed view of the convolution unit. Output is generated from the 3×3 input image, shown in the red box, and the kernel.

Fig from: Moini, Shayan, et al. Remote power side-channel attacks on BNN accelerators in FPGAs, in IEEE DATE'21

Power model

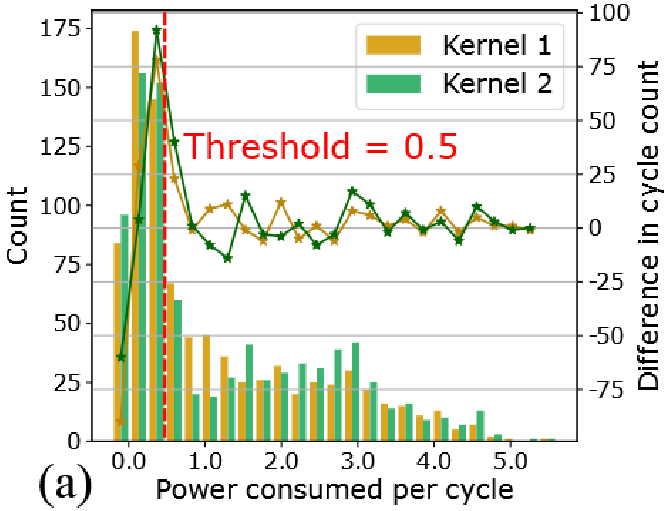
- The more internal activity (i.e., convolution unit), the higher the power consumed
- **If data remain unchanged between cycles, internal transitions induced are limited**
    - ⇒ Observing the magnitude of the power consumption in each cycle
    - ⇒ Deducing related pixels with similar values (e.g., background)

Experimental protocol

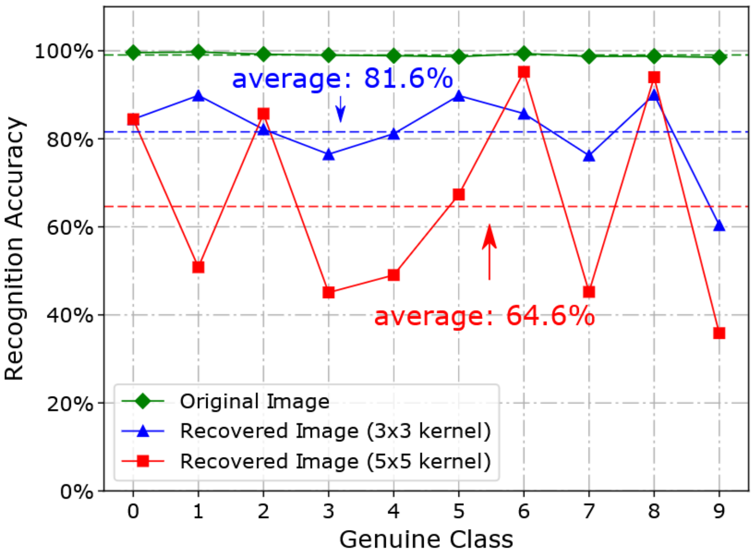
- Simply choosing a threshold to differentiate back from foreground pixels
- Experimental setup
  - BNN, kernel 3*3 and 5*5
  - MNIST 28*28
  - Line size buffer 28
  - Xilinxs Spartan-6 on the SAKURA G board designed for power measurements

Some results

- Metrics?
  - Pixel-level accuracy
  - Recognition accuracy (through MLP, vs 99%)



For background pixels



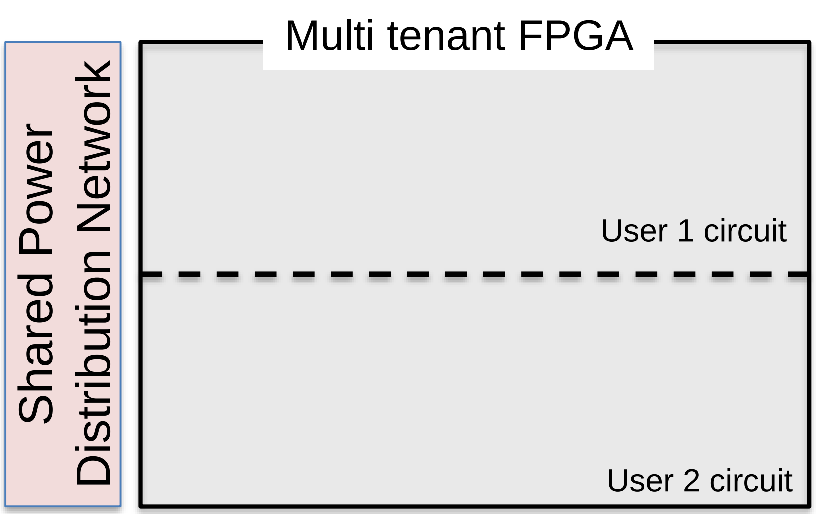
- Loss of information proportional to the size of the kernel
- Compared to EM?

Contents

- Motivating security of embedded Neural Network
- Input extraction-vulnerabilites at the SoC level?
- Input extraction – within the accelerator -> exploiting an implementation choice
- What if there is no physical access to the victim?

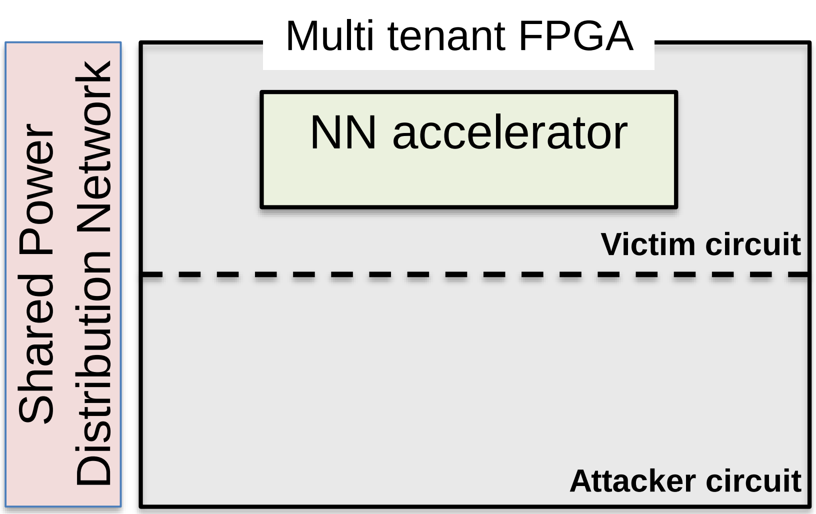
What if there is no physical access to the victim?

- Multi tenant environment on FPGA
- The power distribution model is shared among the entire FPGA



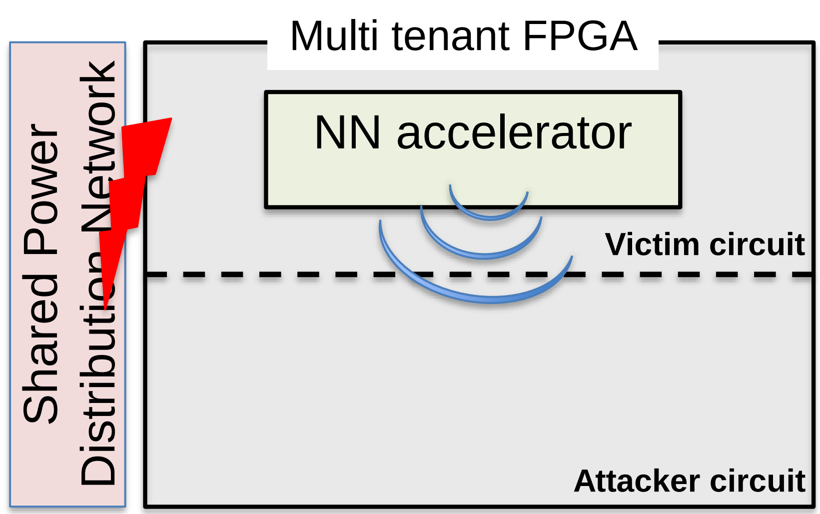
What if there is no physical access to the victim?

- Multi tenant environment on FPGA
- The power distribution model is shared among the entire FPGA

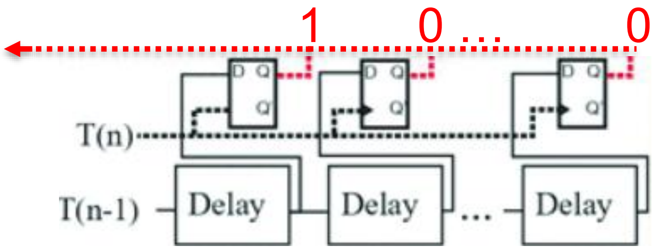


What if there is no physical access to the victim?

- Multi tenant environment on FPGA
- The power distribution model is shared among the entire FPGA
- Can a collocated attacker sense what the victim is processing?

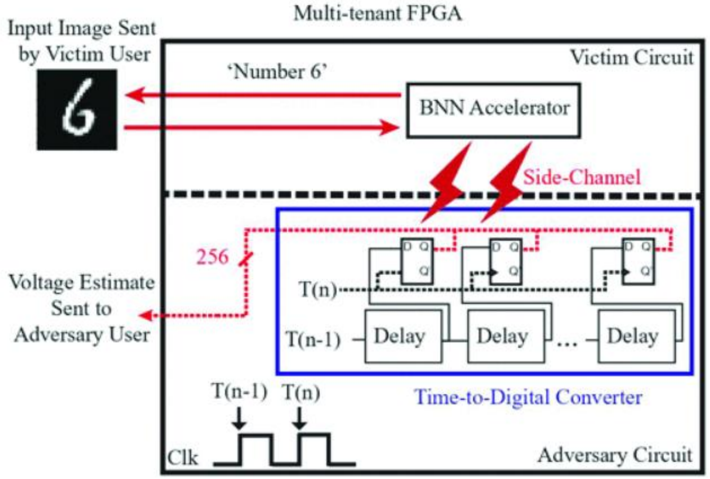


- Custom circuits can be designed as voltage sensors
  - Signal delay varies as supply voltage changes
  - An attacker circuit, near the victim circuit can sense voltage changes and deduce the victim activity!
  - Ex:
    - Time Delay Converter
    - Ring Oscillators



Zick, K. et. Al.,. Sensing nanosecond-scale voltage attacks and natural transients in FPGAs. In Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays'13

- Threat model
  - **Same, BUT no physical proximity is required**
  - No interaction with the NN
  - Attack and victim co-located on the same FPGA
  - Attack locates voltage sensors near the victim circuit
  - Based on line buffer architecture for convolution implementation



Moini, Shayan, et al. "Remote power side-channel attacks on BNN accelerators in FPGAs." 2021 Design, Automation *& Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021.

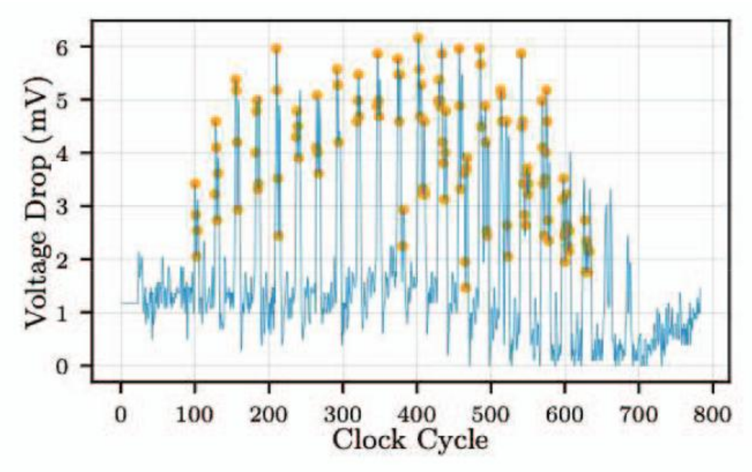
Experimental setup

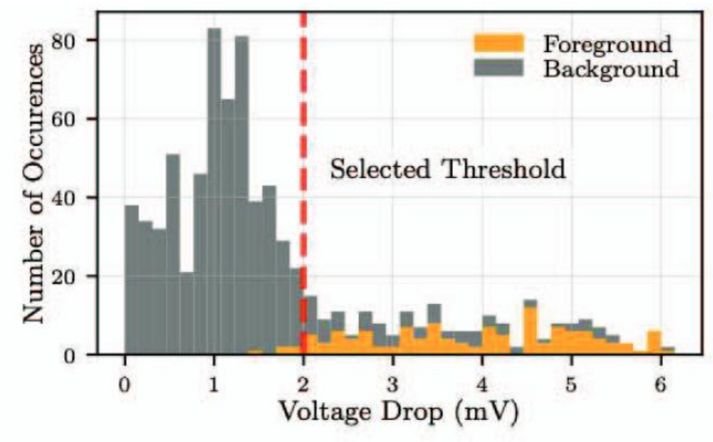
- 3 different Xilinx FPGA-based boards

TABLE I: Details of the evaluation boards used for the experiments. The system clock generates the clock for the BNN accelerator and TDC module.

Board Name	Device	FPGA Family	Clk (MHz)
ChipWhisperer	XC7A100T	Artix 7	50
ZCU104	XCZU7EV	Zynq UltraScale+	120
VCU118	XCVU9P	Virtex UltraScale+	100

Some results





Orange peaks are foreground pixels

Moini, Shayan, et al. "Remote power side-channel attacks on BNN accelerators in FPGAs." 2021 Design, Automation *& Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021.

Some results

- Examples on inputs recovered on different boards

(a) Input images

(b) Recovered images from ZCU104

(c) Recovered images from VCU118

- What can go wrong?

Some results

- Efficiency depends on the number of runs (same image), and TDC placement (3000 runs)
- Metric? -> cross-correlation, recognition accuracy (65%)

(a) ZCU104, cross-die (b) ZCU104, adjacent (c) VCU118, cross-die (d) VCU118, adjacent

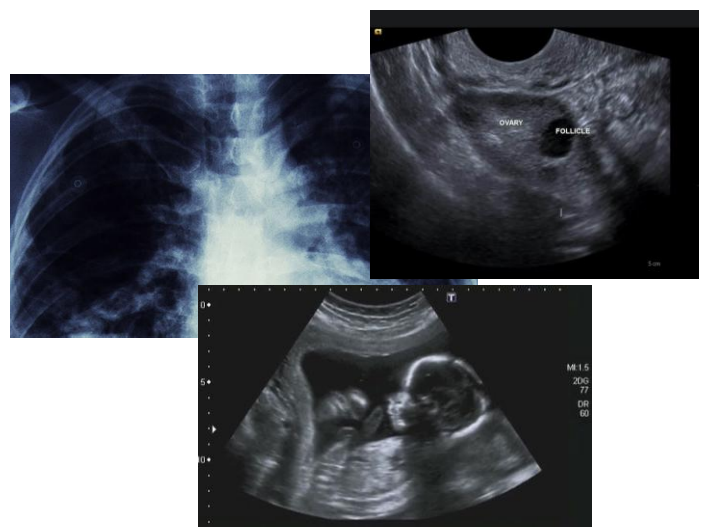
Fig. 9: Recovered images with adjacent and cross-die placement for 3,000 runs.

(a) 100, 0.19 (b) 500, 0.61 (c) 1,000, 0.65 (d) 3,000, 0.75

Fig. 10: Recovered images for the ZCU104 board for (number of runs, normalized cross-correlation with the original image).

Some conclusions

- Embedded AI are security/privacy vulnerable in many ways
- Attack vectors at the implementation, at the SoC
- Privacy attacks
  - Are these rather simple inputs interesting?/realistic?

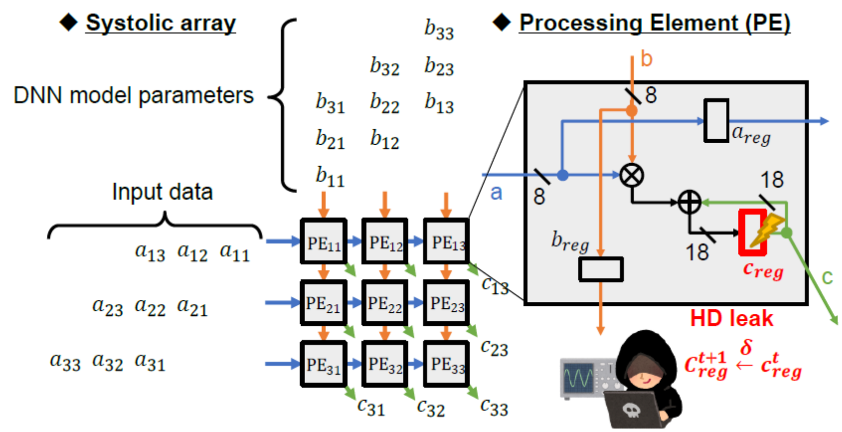


Can we extract anything else? weights?

- Many other attacks.

What if the only thing the attacker knows is actually the inputs?

- A long way to go …



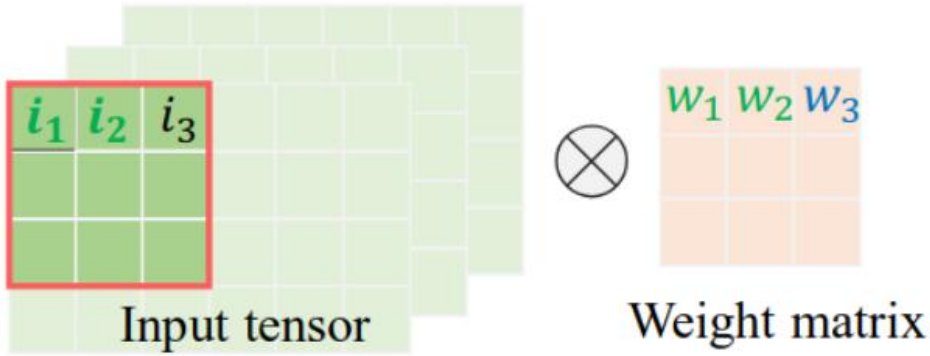
Credit: K. Yoshida et al., Model Reverse-Engineering Attack using Correlation Power Analysis against Systolic Array Based Neural Network Accelerator, ISCAS'20

Solutions?

- Vulnerabilities on the bus transfer
- Vulnerabilities on the implementation choices
- Vulnerabilities when sharing the PDN

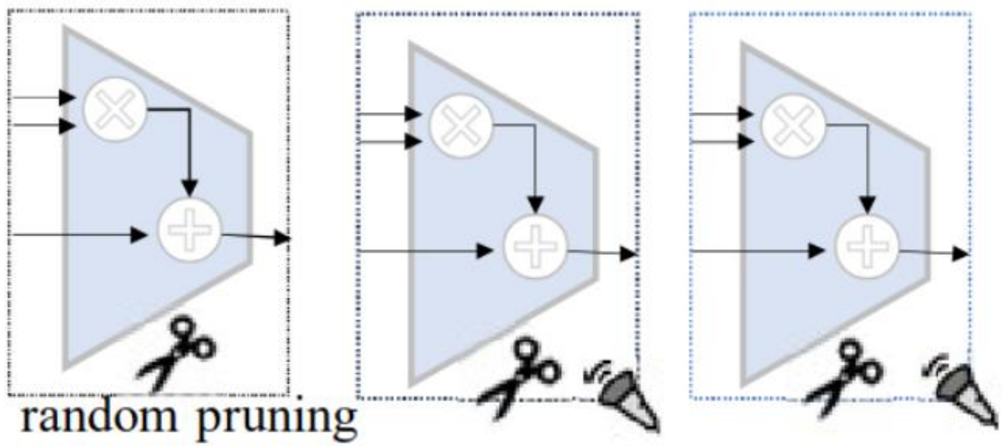
Illustration of MACPruning

Attacker lose a proportion of correlated EM emission signal to their attack model due to random pruned MAC operation.



Target weight: w_3 Attacker input: i_1, i_2, i_3

Attack model: $HW(a_3)$?



?

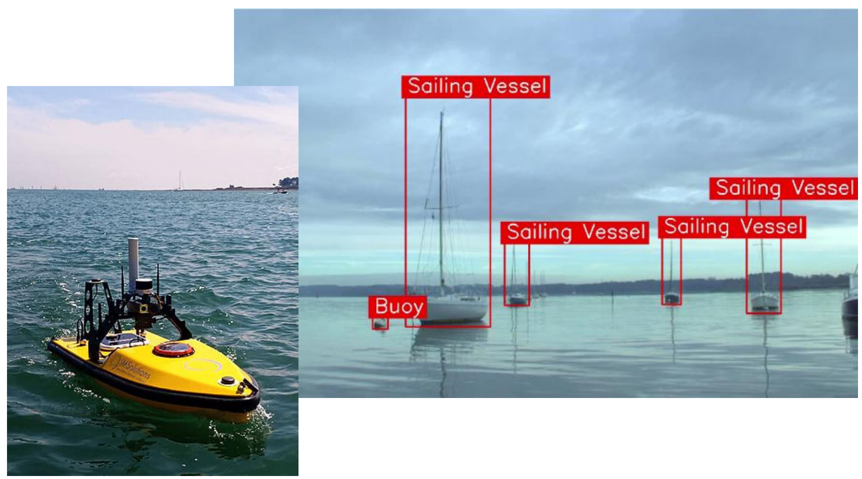
Which attack model should I use? Which MAC should I attack?

$HW(a_3) = (i_1w_1 + i_3w_3)? \ (i_2w_2 + i_3w_3)? \ (i_1w_1 + i_2w_2 + i_3w_3)?$

Lab-STICC

Thank you !

Interested in PostDoc positions?



Maria Méndez Real
Chaire Professeur Junior
Meutes de Drones Maritimes Autonomes et de Confiance
maria.mendez-real@univ-ubs.fr