UNDERSTANDING A CNN ATTACK: AS CRUCIAL AS SUCCEEDING IT

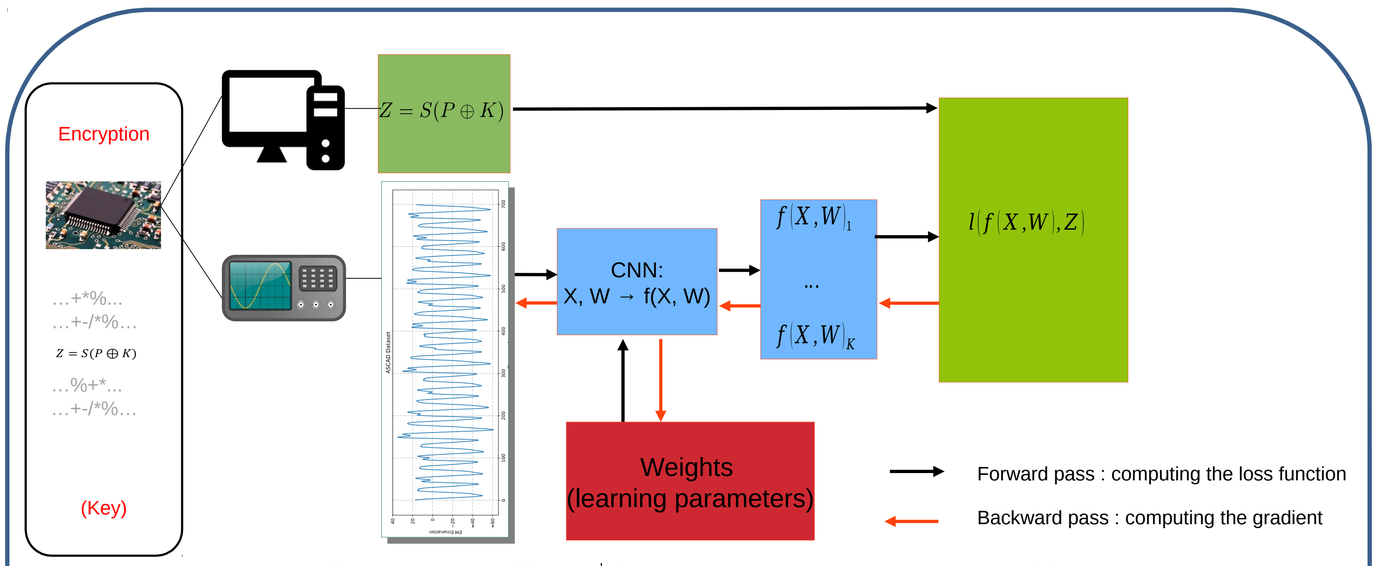
Today : CNNs are used to extract sensitive information in Side Channel Analysis. How do they extract relevant information from the trace? This well known problem in Machine Learning implies a lack of expertise to design further counter measures.
New idea : use CNNs to localize sensitive information and PoIs.

CONTACT :

Loïc Masure (1, 2, 3)

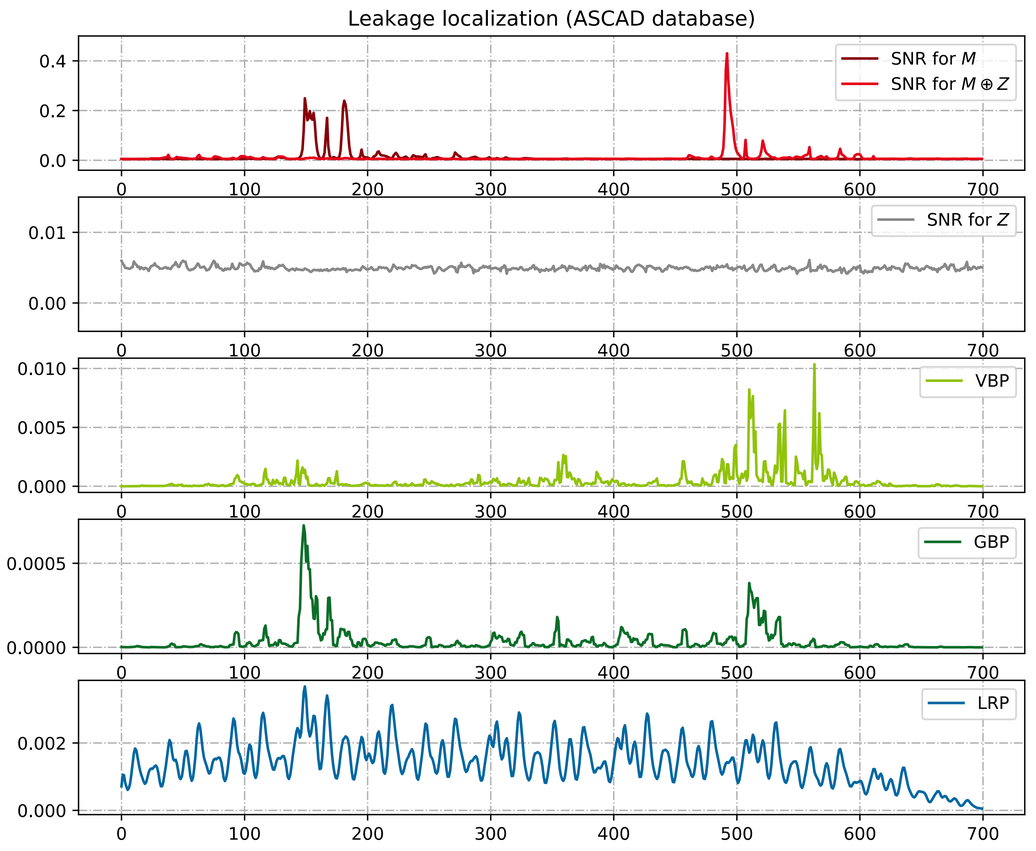
Cécile Dumas (1, 2)

Emmanuel Prouff (3)



Parameter	Value	Metrics	Value
n_{blocks}	8	Cross Entropy	7.914
n_{conv}	1	Accuracy	0.570 %
n_{dense}	2	Guessing Entropy (one-shot)	110.418
$n_{filters,1}$	32	Min. Traces Required	86.000
W_{conv}	5		
N_{param}	$\approx 5.10^6$		

Table 1: CNN architecture and performance metrics



Vanilla Backprop (VBP)

- Computes the loss function and its gradient via backprop through the layers to the input trace.
- The loss function is more likely to be impacted by small perturbations at relevant PoIs. Therefore the gradient should be higher at those points.
- Problem: the loss function for complex CNNs is prone to be less smooth, leading to irrelevant peaks unless training with lots of data.

Guided Backprop (GBP)

- Same principle as VBP except for ReLU layers.
- Activation in a ReLU layer:

$$f_i^{l+1} = ReLU(f_i^l) = \max(f_i^l, 0)$$

- Regular backprop:

$$R_i^l = (f_i^l > 0).R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$$

- Guided Backprop:

$$R_i^l = (f_i^l > 0).(R_i^{l+1} > 0).R_i^{l+1}$$

- Gives better and sparser relevance maps.

Layerwise Relevance Prop (LRP)

- Each type of layer has its own rule of relevance propagation.
- More details at heatmapping.org
- Performance highly depends on an *explanable* architecture.

Results

- We introduced means to **interpret** the success of a CNN attack.
- It can help the developer to understand the vulnerability and therefore help for a better design against such threats.
- Sensitivity methods (VBP, GBP) perform a better **characterization** than SNRs against **masking**.
- Those methods are also robust against **desynchronization**, provided the CNN is trained on desynchronized data (see animated demo).
- Visualizing the characterization helps evaluating whether a CNN overfits, and therefore leads the choice for an **optimal architecture**.

Affiliations

1. CEA, LETI, MINATEC Campus, F-38054 Grenoble, France, name.surname@cea.fr
2. Univ. Grenoble Alpes, F-38000, Grenoble, France
3. Sorbonne Universités, UPMC Univ Paris 06, POLSYS, UMR 7606, LIP6, F-75005, Paris, France