AI to the Rescue: Where AI Meets Cryptography

Stjepan Picek

SAC 2025, August 13, 2025

Outline

Outline

What is Intelligence?

- Lat. intelligere – understand, comprehend.
- Intelligence is a descriptive concept – it describes certain properties of an individual or a group of individuals.
- There is no consensus on the definition of intelligence.
- Most definitions include concepts such as abstract reasoning, understanding, self-consciousness, communication, learning, planning, and problem solving.

What is Artificial Intelligence?

- A branch of computer science: Technical Sciences $\rightarrow$ Computer Science $\rightarrow$ Artificial Intelligence.
- The branches of Artificial Intelligence (according to ACM):
  1. General AI (cognitive modeling, philosophical foundations)
  2. Expert systems and applications
  3. Automated programming
  4. Deduction and theorem proving
  5. Formalisms and methods for knowledge representation
  6. Machine learning
  7. Understanding and processing of natural and artificial languages
  8. Problem solving, control methods, and state space search
  9. Robotics
  10. Computer vision, pattern recognition, and scene analysis
  11. Distributed artificial intelligence

The name "Artificial Intelligence"

- AI as an independent research area was established in 1956 at the Dartmouth Conference (10 scientists, 2 months)
- "...The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." (McCarthy et al. 1955)
- "We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."

Artificial Intelligence

- AI is the new electricity. (Andrew Ng)

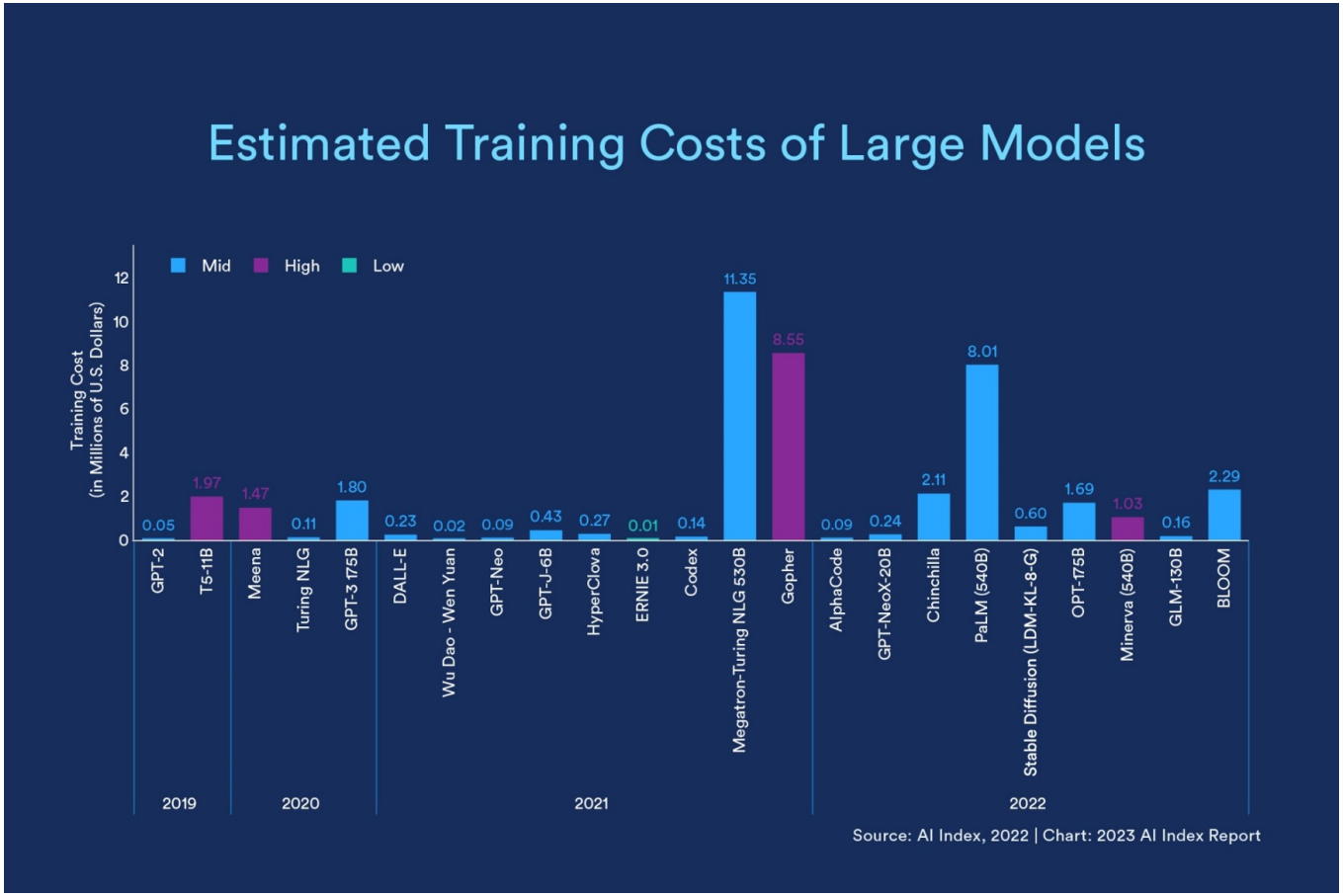
- Computer vision.
- Healthcare.
- Speech recognition.
- Natural Language Processing.
- Robotics.
- **Security**.
- ...

Artificial Intelligence

- Powerful hardware.
- Big data.
- Novel applications.

AI is Becoming Better

AI is Becoming More Expensive



History

- Communication between Norbert Wiener and Warren Weaver (1947).
- "A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples..."
- "Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.""

Learnability

- A learning machine consists of a learning protocol together with a deduction procedure. The former specifies the manner in which information is obtained from the outside. The latter is the mechanism by which a correct recognition algorithm for the concept to be learned is deduced.
- There is circumstantial evidence from cryptography, however, that the whole class of functions computable by polynomial size circuits is not learnable.
- The existence of good cryptographic functions that are easy to compute therefore implies that some easy-to-compute functions are not learnable.
- "A Theory of the Learnable" (Leslie Valiant, 1984).

Learnability

- "Cryptography and Machine Learning" (Ronald Rivest, 1991).

- "The techniques used demonstrate an interesting duality between learning and cryptography."
- "Cryptographic limitations on learning Boolean formulae and finite automata" (Michael Kearns, Leslie Valiant, 1994).

AI and Security

- Implementation attacks.
- Hardware Trojans.
- Modeling attacks on PUFs.
- Design of cryptographic primitives.
- Cryptanalysis.
- Intrusion detection.
- Malware and spam identification/detection.
- Fuzzing.
- Privacy-preserving machine learning.
- Adversarial machine learning.
- Steganography and steganalysis.
- LLMs as covert channels.
- ...

AI and Security

- AI for security and security of AI (ML).
- In AI for security, we can use AI either in the defense/design or attack phase.
- Attacks seems more explored since it is easier to validate that the attack works.

- In security of AI, we can use cryptographic techniques to either attack AI systems or to improve their privacy/security.

Outline

Heuristic Design of Cryptographically Strong Balanced Boolean Functions

- Eurocrypt 98.
- Experiments for $n = 8$.
- Genetic algorithm capable of generating highly nonlinear balanced Boolean functions.
- Hill climbing techniques are adapted to locate balanced, highly nonlinear Boolean functions that also almost satisfy correlation immunity.

Search for Boolean Functions With Excellent Profiles in the Rotation Symmetric Class

- Modified steepest-descent-based iterative heuristic search.
- Boolean functions on 9 variables having nonlinearity 241.
- 10 variable functions having first-order resiliency and nonlinearity 492.

Physically Unclonable Functions

- Physically Unclonable Functions (PUFs) are embedded or standalone devices used as a means to generate either a source of randomness or to obtain an instance-specific uniqueness for secure identification.
- This is achieved by relying on inherent uncontrollable manufacturing process variations, which results in each chip having a unique response.
- No two PUFs will give the same response when supplied with the same challenge.

Physically Unclonable Functions

- Two types of PUFs: strong and weak.
- The difference concerning the number of challenge-response pairs (CRPs) the attacker can obtain.
- The number of unique challenges c scales polynomially with the circuit area of a weak PUF.
- The number of unique challenges c scales exponentially with the circuit area of a strong PUF.

Physically Unclonable Functions

- Weak PUF has a limited number (typically, one or few) of responses to challenges.
- Strong PUFs have a large number of responses (concerning different challenges).
- Strong PUFs have a virtually unlimited number of challenges c, but their CRPs are highly correlated.
- Given enough (often a small amount) of CRPs, it is possible to build a predictive model of a strong PUF (in a way, we build a mathematical clone since it is not feasible to make an analog physical clone).
- There exists no validated design of a strong PUF that is fully resilient against modeling attacks.

Physically Unclonable Functions

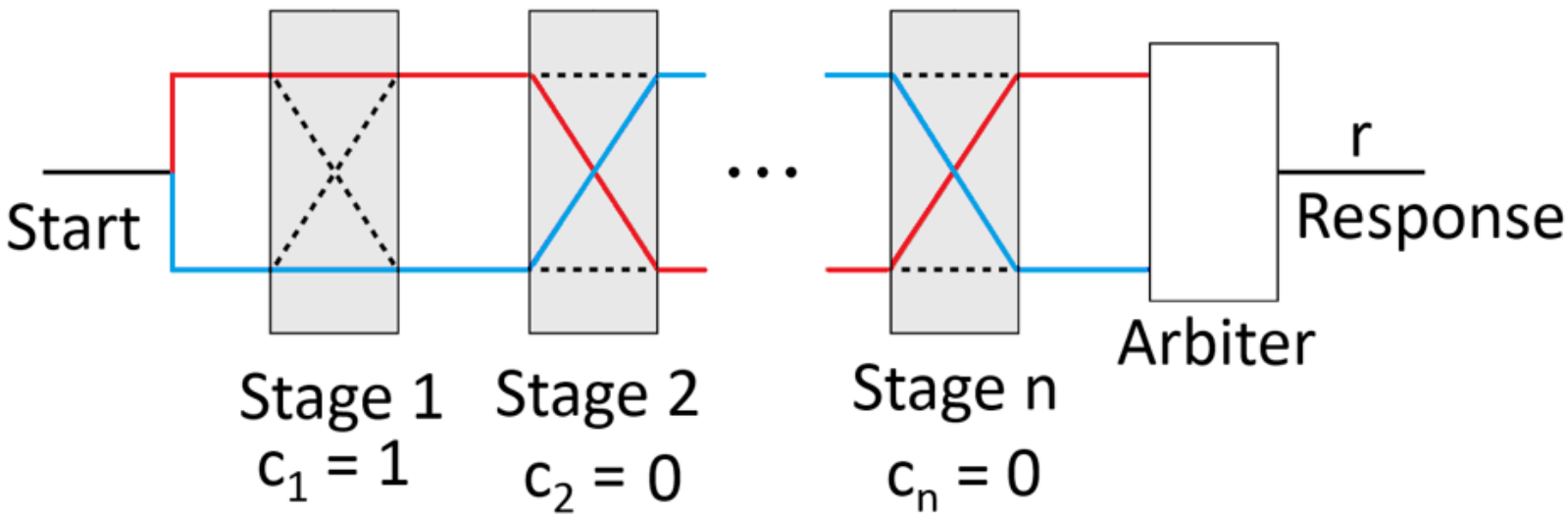


Figure: An example of a strong PUF - Arbiter PUF with n stages.

Physically Unclonable Functions

- Several techniques are commonly used to break strong PUFs.
- From ML domain, logistic regression, and from EC, evolution strategy.
- This domain is very interesting as AI provided results that were not possible to obtain with any other technique.
- What is more, even simple AI techniques can easily break strong PUFs.
- This also means there is not much development in the domain as attacks are easy to do, so there is no clear benefit of using more complex techniques, e.g., deep learning.

Cryptographic Theory vs. Physical Reality

- Cryptographic algorithms are (supposed to be) theoretically secure.
- Implementations leak in the physical world.

Implementation attacks

Implementation attacks do not aim at the weaknesses of the algorithm but at its implementation.

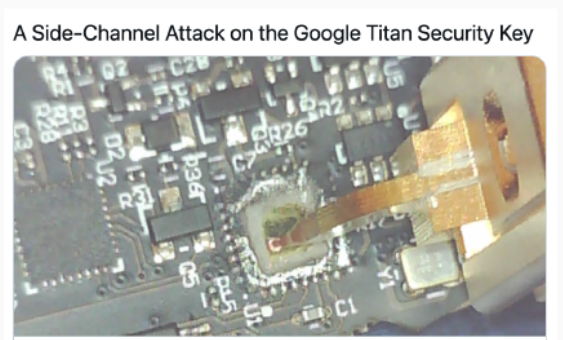
Relevance

November 13, 2019

October 3, 2019

May 28, 2020

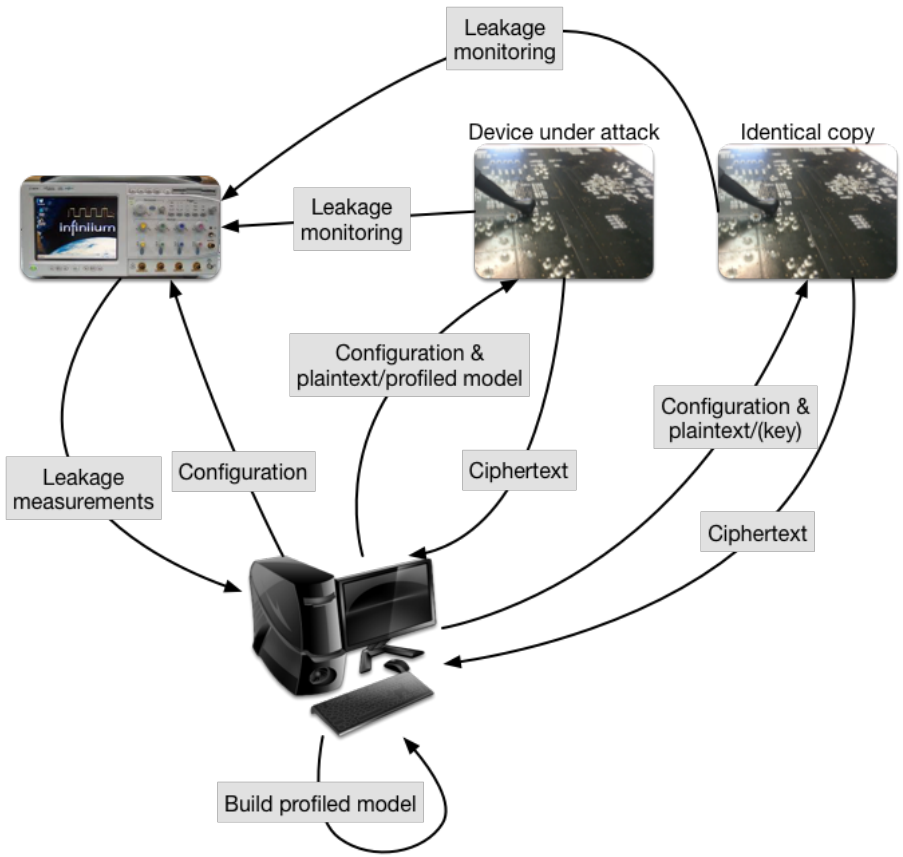
January 7, 2021



Profiling Attacks

- Profiling attacks have a prominent place as the most powerful among side-channel attacks.
- Within the profiling phase, the adversary estimates leakage models for targeted intermediate computations, which are then exploited to extract secret information in the actual attack phase.
- Some **machine learning** (ML) techniques also belong to the profiling attacks.

Profiling Attacks



- Profiling attacks are more complicated than direct attacks.
- The attacker must have a copy of the device to be attacked.

State-of-the-art Results with DLSCA

Table: Points of interest, the minimum number of attack traces to get guessing entropy equal to 1, model search success (when GE=1), and number of trainable parameters for all datasets and feature selection scenarios.

Dataset	Neural Network Model	Feature Selection Scenario	Amount of POIs (HW/ID)	Attack Traces (HW/ID)	Search Success (%) (HW/ID)	Trainable Parameters (HW/ID)
ASCADf	MLP	RPOI	200/100	5/**1**	99.22%/96.86%	82 209/429 256
ASCADf	CNN	RPOI	400/200	5/**1**	99.23%/99.08%	499 533/158 108
ASCADf	MLP	OPOI	700/700	480/104	82.80%/68.80%	16 309/10 266
ASCADf	CNN	OPOI	700/700	744/87	55.53%/35.33%	594 305/62 396
ASCADf	MLP	NOPOI	2 500/2 500	7/**1**	74.50%/39.00%	2 203 009/5 379 256
ASCADf	CNN	NOPOI	10 000/10 000	7/ **1**	15.40%/2.45%	545 693/439 348
ASCADf	CNN	NOPOI desync	10 000/10 000	532/36	2.44%/2.64%	268 433/64 002
ASCADr	MLP	RPOI	200/20	3/**1**	99.23%/100%	565 209/639 756
ASCADr	CNN	RPOI	400/30	5/**1**	100%/100%	575 369/636 224
ASCADr	MLP	OPOI	1 400/1 400	328/129	71.40%/37.25%	31 149/34 236
ASCADr	CNN	OPOI	1 400/1 400	538/78	47.92%/23.95%	270 953/87 632
ASCADr	MLP	NOPOI	25 000/25 000	6/ **1**	44.39%/7.02%	5 243 209/12 628 756
ASCADr	CNN	NOPOI	25 000/25 000	7/ **1**	19.17%/4.35%	369 109/721 012
ASCADr	CNN	NOPOI desync	25 000/25 000	305/73	0.71%/1.04%	22 889/90 368

ATM Data

- Collected the environmental audio (exploiting the webcam microphone) and the keylogs of the PIN pad through the USB interface during the experiment.

Figure: Our experimental setup.

Results

(a) *True digit = 7 Pred = 7 (0.999), 4 (0.000), 8 (0.000)*

(b) *True digit = 3 Pred = 3 (0.979), 2 (0.012), 6 (0.005)*

(c) *True digit = 6 Pred = 6 (0.819), 9 (0.170), 8 (0.009)*

(d) *True digit = 3 Pred = 3 (0.809), 2 (0.092), 5 (0.069)*

(e) *True digit = 3 Pred = 2 (0.329), 3 (0.315), 6 (0.185)*

Figure: PIN 73633 entered by a user in our test set in the *Single PIN pad* scenario. Our algorithm suggests 73632 as the most probable PIN (probability = 21.32%), 73633 as the second most probable PIN (probability = 20.43%), and 73636 as the third most probable PIN (probability = 11.96%). The algorithm predicts the correct PIN in the second attempt.

Fault Injection

- A fault injection (FI) attack is successful if, after exposing the device to a specially crafted external interference, it shows an unexpected behavior exploitable by the attacker.
- FI can be divided into the characterization phase (finding faults) and using those faults for a successful attack.
- Insertion of signals has to be precisely tuned for the fault injection to succeed.
- Finding the correct parameters for a successful FI can be considered a search problem where one aims to find, within a minimum time, the parameter configurations that result in a successful fault injection.

Fault Injection

- Depending on the source of the fault, the search space of possible parameters changes significantly.
- Generally, the search space is too big to conduct an exhaustive search.
- Commonly, one defines several possible classes for classifying a single measurement:
  1. NORMAL: smart card behaves as expected, and the glitch is ignored
  2. RESET: smart card resets as a result of the glitch
  3. CHANGING: the response is changing when repeating measurements.
  4. SUCCESS: smart card response is a specific, predetermined value that does not happen under normal operation

Fault Injection

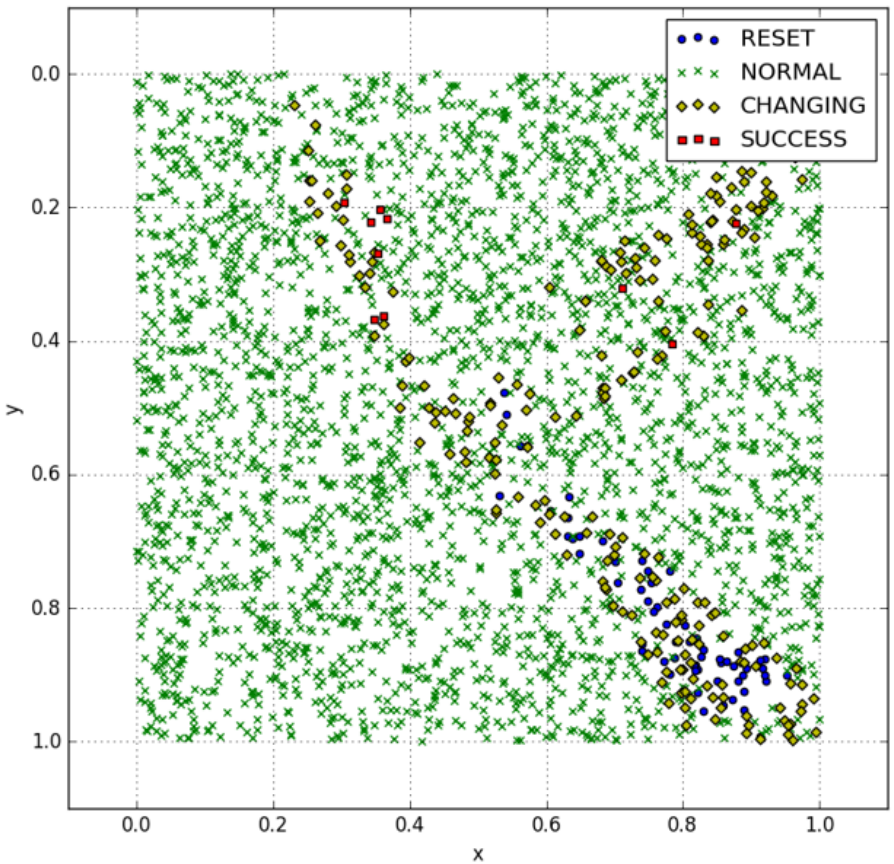
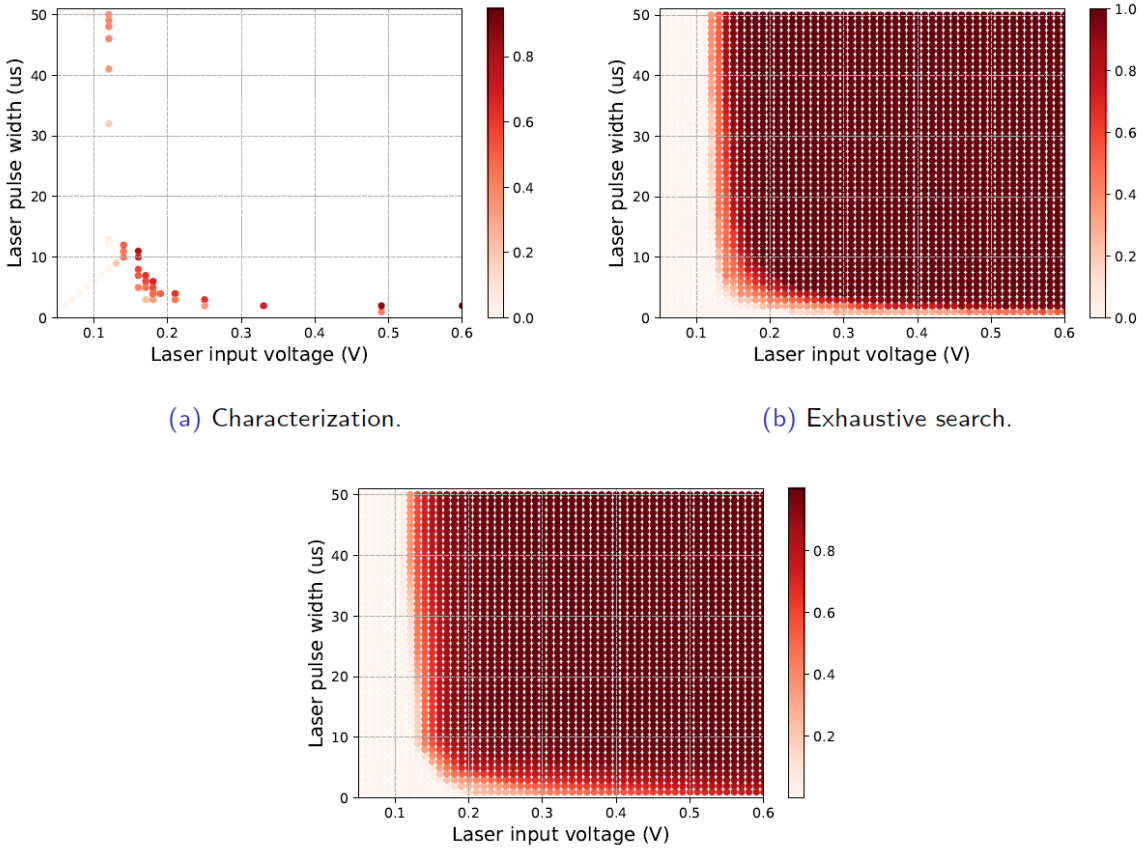


Figure: A depiction of search space for voltage glitching and two parameters.

Fault Injection



(a) Characterization.

(b) Exhaustive search.

Figure: Deep learning prediction.

Traditional Cryptanalysis

- Aims at recovering the secret key by using a knowledge of (P, C) pairs.
- Looking for patterns to distinguish encrypted data from random.
- Adversary's goal is to distinguish the output of a cipher from random data faster than brute force key search.
- Two common key-recovery attacks:
  1. differential cryptanalysis: exploits difference propagation
  2. linear cryptanalysis: exploits large $P - to - C$ correlations

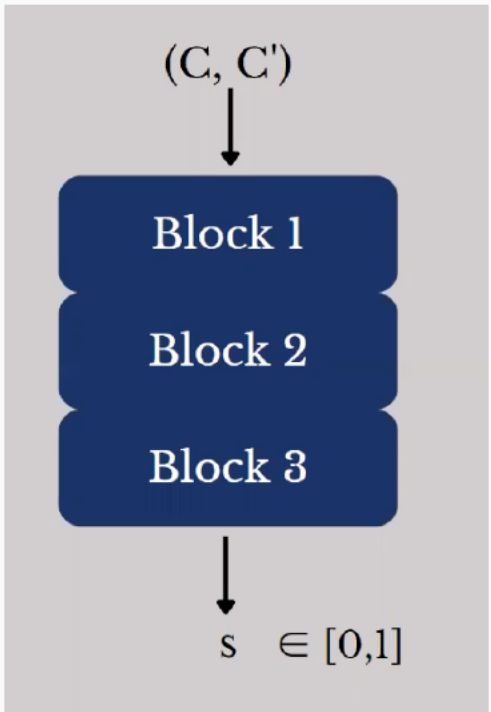
Differential Cryptanalysis

- Invented by Biham and Shamir in 1990 as a way to attack DES.
- Exploits a scenario where a particular ΔC occurs given a particular input difference ΔP with a "high" probability.
- It is a chosen plaintext attack, so the attacker will select pairs of inputs, P and $P^{'}$, to satisfy a particular ΔP.

Neural-aided Cryptanalysis

- Started by Gohr in 2019.
- Trained neural distinguishers of depth-10 and depth-1 for round-reduced versions of Speck32/64.
- The approach proved successful on 5-8 rounds (accuracy above 50%).
- Improved 11-round key recovery attack complexity on Speck32/64 (using Bayesian optimization).
- Up to now, used on more than 20 different cryptographic algorithms.

Neural-aided Cryptanalysis



A three-block **neural network**.

(C, C') is either the **encryption** of:
1. **(P, P')** where **P' = P ⊕ Δ**
2. **(P, P')** where there is **no fixed difference**

If **s ≥ 0.5** then (C, C') is a **real pair**
else (C, C') is a **random pair**.

Machine Learning Attack on LWE-based Cryptographic Schemes

- Lattice-based cryptosystems, based on a hard problem known as Learning With Errors (LWE), have emerged as strong contenders for PQC standardization.
- The idea is to train transformers to perform modular arithmetic and combine half-trained models with statistical cryptanalysis techniques.
- The attack can recover secrets for small-to-mid size LWE instances with sparse binary secrets.

Outline

Attacks on Machine Learning

	Revealing confidential information on the learning model or its users	Misclassifications not compromising normal system operation	Misclassifications compromising normal system operation
Attacker capability	**Confidentiality**	**Integrity**	**Availability**
Training data		Backdoor/targeted poisoning	Sponge poisoning
Test data	Model extraction/ stealing Model inversion Membership inference	Evasion attacks	Sponge attacks

Poisoning Attacks

- The goal of the attacker is to contaminate the machine model generated in the training phase so that predictions on new data will be modified in the testing phase.
- In targeted poisoning attacks, the attacker wants to misclassify specific examples.
- In non-targeted attacks, the attacker aims to degrade the model's performance (DoS attack).

Model Backdoors

- Backdoors are a particular type of poisoning attack, also named Trojans.
- Backdoor attacks aim to make a model misclassify some of its inputs to a preset-specific label while other classification results behave normally.
- This misclassification is activated when a specific pattern is added to the model input.
- This pattern is called the trigger and can be anything the targeted model understands.

Model Backdoors

- Goldwasser et al. show how to plant an undetectable backdoor into a classifier that, without an appropriate "backdoor key", cannot be detected by any computationally-bounded observer.
- First construction shows how to plant a backdoor in any classifier, leveraging the cryptographic notion of digital signatures.
- Assuming the existence of one-way functions, for every training procedure Train, there exists a model backdoor (Backdoor, Activate), which is non-replicable and black-box undetectable.
- "Planting Undetectable Backdoors in Machine Learning Models" (Goldwasser et al., 2024).

Model Stealing

- Model stealing attacks try to mimic or fully copy a target model.
- Leveraging oracle access to a model f, the attack tries to reconstruct it with oracle access $x : f(x) \simeq \hat{f}(x)$.
- The model $\hat{f}$ acquired by approximation relies on creating a model that, by iterative adjustments, performs similarly (same) to the original model.
- However, it is not architecturally the same.

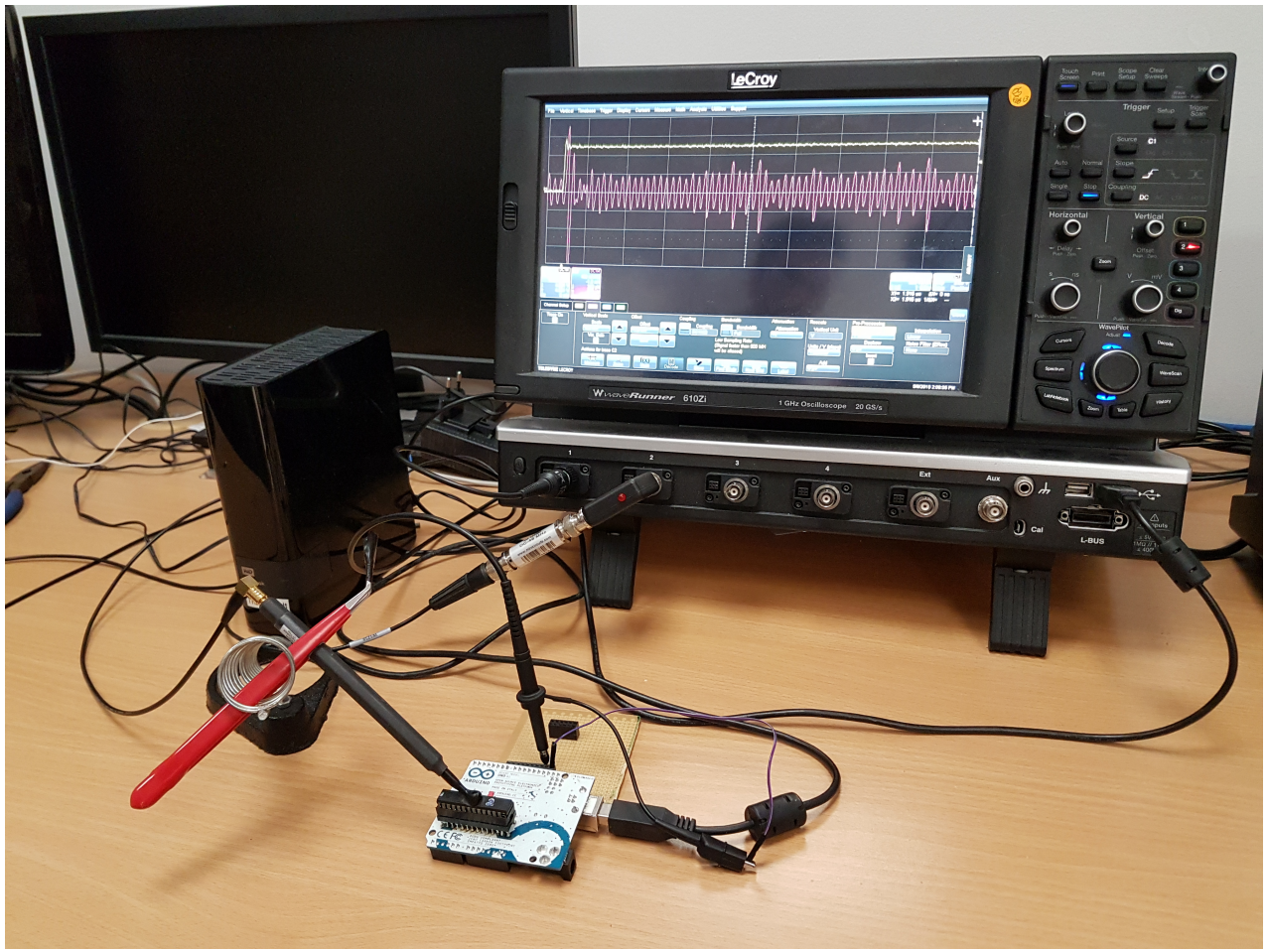
Cryptanalytic Extraction of Neural Network Models

- The machine learning problem of model extraction is a cryptanalytic problem in disguise.
- Given oracle access to a neural network, one can mount a differential attack that can efficiently steal the parameters of the remote model up to floating point precision.
- The attack relies on the fact that ReLU neural networks are piecewise linear functions, and thus queries at the critical points reveal information about the model parameters.

Cryptanalytic Extraction of Neural Network Models

- Performing a model extraction attack-learning the weights θ given oracle access to the function f_θ - is a similar problem to performing a chosen-plaintext attack on a nontraditional "encryption" algorithm.
- A differential attack that is effective at performing functionally-equivalent neural network model extraction attacks.
- The attack traces the neural network's evaluation on pairs of examples that differ in a few entries and uses this to recover the layers (analogous to the rounds of a block cipher) of a neural network one by one.
- All the weights and biases of black-box ReLU-based DNNs could be inferred using a polynomial number of queries and computational time.

Reverse Engineering of Neural Networks with SCA



(a) The complete measurement setup

Reverse Engineering the Number of Neurons and Layers

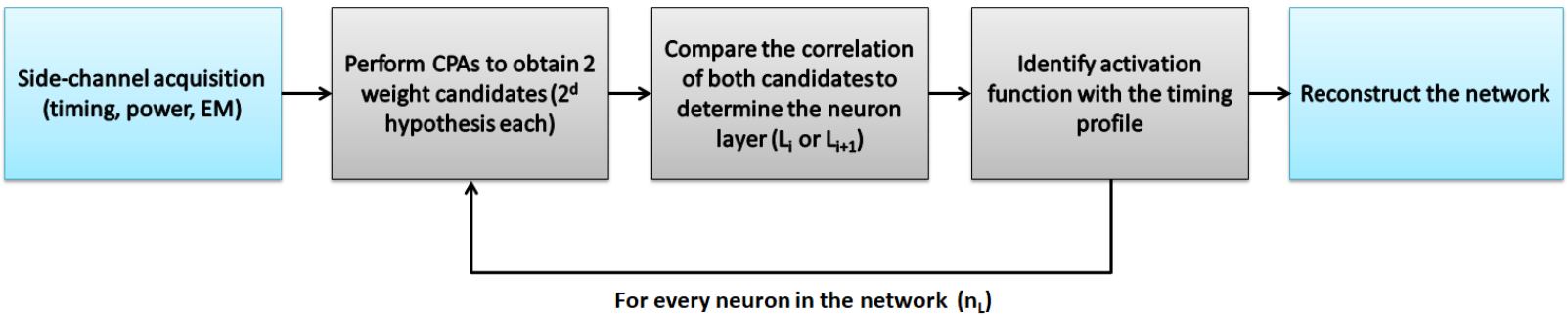


Figure: Methodology to reverse engineer the target neural network

ARM Cortex M-3 and MLP

- Tests with MNIST and DPAv4 datasets.
- DPAv4: the original accuracy equals 60.9%, and the accuracy of the reverse-engineered network is 60.87%.
- MNIST: the accuracy of the original network is equal to 98.16%, and the accuracy of the reverse-engineered network equals 98.15%, with an average weight error converging to 0.0025.

Outline

Conclusions

- AI has a prominent role in cryptography (and even more security).
- Current results are promising but we need more relevant problems.

- Cryptography also becomes more important for AI but we need more practical settings.
- In any way, it is good to see that AI is becoming more accepted in the crypto community.

Questions?

Thank you for your attention!

I am happy to answer your questions.

stjepan.picek@ru.nl