Report from Dagstuhl Seminar 22412

Intelligent Security: Is "AI for Cybersecurity" a Blessing or a Curse

Nele Mentens*[1], Stjepan Picek*[2], and Ahmad-Reza Sadeghi*[3]

1 **Leiden University, NL.** n.mentens@liacs.leidenuniv.nl
2 **Radboud University Nijmegen, NL.** stjepan.picek@cs.ru.nl
3 **TU Darmstadt, DE.** ahmad.sadeghi@trust.tu-darmstadt.de

Abstract

This report documents the outcomes of Dagstuhl Seminar 22412 "Intelligent Security: Is "AI for Cybersecurity" a Blessing or a Curse". The seminar brought together 25 attendees from 10 countries (Canada, Croatia, Czech Republic, France, Germany, Netherlands, Singapore, Sweden, Switzerland, and the USA). There were 17 male and 8 female participants. Three participants were from the industry, and the rest were from academia.

The gathered researchers are actively working in the domains of artificial intelligence and cybersecurity, emphasizing hardware security, fuzzing, physical security, and network security. The seminar aims to foster sharing experiences and best practices between various cybersecurity applications and understand how and when certain approaches are transferable. The first two days were devoted to 20-minute self-introductions by participants to achieve these goals. At the end of the second day, we made a list of topics that were decided to be the focus of the seminar and that will be discussed in the groups in the next few days. On the third and fourth days, the work was conducted in four discussion groups where at the end of each day, all participants gathered to report the results from the discussion groups and to align the goals. On the last day, we again worked in one group to summarize the findings and foster networking among participants. A hike was organized in the afternoon of the third day. The seminar was a success. The participants actively participated in the working groups and the discussions and went home with new ideas and collaborators. This report gathers the abstracts of the presented talks and the conclusions from the discussion groups, which we consider relevant contributions toward better interdisciplinary research on artificial intelligence and cybersecurity.

Seminar October 9–14, 2022 – http://www.dagstuhl.de/22412

2012 ACM Subject Classification Security and privacy → Cryptography; Security and privacy → Intrusion/anomaly detection and malware mitigation; Security and privacy → Security in hardware; Security and privacy → Systems security; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Computer systems organization → Real-time systems

Keywords and phrases Cybersecurity, Artificial Intelligence, Hardware Security, Machine Learning, Explainability

Digital Object Identifier 10.4230/DagRep.12.10.106

* Editor / Organizer

Intelligent Security: Is "AI for Cybersecurity" a Blessing or a Curse, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 106–128
Editors: Nele Mentens, Stjepan Picek, and Ahmad-Reza Sadeghi

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Stjepan Picek (Radboud University, NL)

In recent years, artificial intelligence (AI) has become an emerging technology to assess security and privacy. Moreover, we can see that AI does not represent "only" one of the options for tackling security problems but instead a state-of-the-art approach. Besides providing better performance, AI also brings automated solutions that can be faster and easier to deploy but are also resilient to human errors. We can only expect that future AI developments will pose even more unique security challenges that must be addressed across algorithms, architectures, and hardware implementations. While there are many success stories when using AI for security, there are also multiple challenges. AI is commonly used in the black-box setting, making the interpretability or explainability of the results difficult. Furthermore, research on AI and cybersecurity commonly look at the various sub-problems in isolation, mostly relying on best practices in the domain. As a result, we often see how techniques are "reinvented", but also that strong approaches from one application domain are introduced to another only after a long time.

The Dagstuhl Seminar 22412 on Intelligent Security: Is "AI for Cybersecurity" a Blessing or a Curse brought together experts from diverse domains of cybersecurity and artificial intelligence with the goal of facilitating the discussion at different abstraction levels to uncover the links between scaling and the resulting security, with a special emphasis on the hardware perspective. The seminar started with two days of contributed talks by participants. At the end of the second day, every participant suggested topics to be discussed in more detail. From the initial pool of nine topics, we decided to concentrate on four topics on the third and fourth day of the seminar: 1) the explainability of AI for cybersecurity, 2) AI and implementation attacks, 3) AI and fuzzing, and 4) the security of machine learning. The first group approached the problem of the explainability of AI for cybersecurity. The discussion mainly revolved around scenarios where deep learning is used as the attack method, but explainability is necessary to understand why the attack worked and, more importantly, how to propose new defense mechanisms that will be resilient against such AI-based attacks. During the discussion, we considered two perspectives: a) understanding the features and b) understanding deep neural networks.

The second group focused on how AI can improve the performance of implementation attacks. More precisely, we discussed the side-channel analysis and fault injection. Most of the discussion aimed at usages of deep learning for side-channel analysis and evolutionary algorithms for fault injection. However, we also discussed how the lessons learned from one domain could be used in another one. The third group worked on the topic of security fuzzing. We discussed how techniques like evolutionary algorithms are used for evolving diverse mutations and mutation scheduling. At the same time, machine learning is (for now) somewhat less used, but there are many potential scenarios to explore. For instance, instead of using evolutionary algorithms, it should be possible to use reinforcement learning to find mutation scheduling. The fourth group discussed the topic of the security of machine learning. More precisely, it focused on backdoor attacks and federated learning settings. While both attack and defense perspectives were discussed, the discussion group emphasized the need for stronger defenses. Each group followed a cross-disciplinary setting where the participants exchanged groups based on their interests. We had one group switch per day to allow sufficient time to discuss a topic. At the end of each day, all participants joined a

meeting to discuss the findings and tweak the topics for the discussion groups. On the last day of the seminar, all participants worked together on fine-tuning the findings and discussing possible collaborations. The reports of the working groups, gathered in the following sections, constitute the main results from the seminar. We consider them the necessary next step toward understanding the interplay between artificial intelligence and cybersecurity, as well as the interplay among diverse cybersecurity domains using AI. Moreover, we expect that the seminar (and this report) will help better understand the main open problems and how to use techniques from different domains to tackle cybersecurity problems. This will encourage innovative research and help to start joint research projects addressing the issues.

2 Table of Contents

3 Overview of Talks

3.1 Can AI clone the microarchitecture of a microcontroller?

Ileana Buhan (Radboud University Nijmegen, NL)

Early attempts to create automated tooling and the recently increased efforts toward this purpose prove the appeal of leakage simulators. A leakage simulator translates a sequence of assembly instructions into a power trace. The challenge for the wide-scale adoption lies in the manual effort required to create a leakage simulator. ABBY is the first post-silicon leakage simulator, where we used deep learning to automate the profiling of the target.

3.2 Deep Learning Application for Side-Channel Analysis and Fault Injection

Lukasz Chmielewski (Radboud Universiteit Nijmegen, NL & Masaryk University – Brno, CZ)

Joint work of Guilherme Perin, Lejla Batina, Stjepan Piecek, Madura Shelton, Niels Samwel, Markus Wagner, Leo Weissbart, Yuval Yarom

Main reference Guilherme Perin, Lukasz Chmielewski, Lejla Batina, Stjepan Picek: "Keep it Unsupervised: Horizontal Attacks Meet Deep Learning", IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2021(1), pp. 343–372, 2021.

URL https://doi.org/10.46586/tches.v2021.i1.343-372

Main reference Lukasz Chmielewski, Leo Weissbart: "On Reverse Engineering Neural Network Implementation on GPU", in Proc. of the Applied Cryptography and Network Security Workshops – ACNS 2021 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, and SiMLA, Kamakura, Japan, June 21-24, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12809, pp. 96–113, Springer, 2021.

URL https://doi.org/10.1007/978-3-030-81645-2_7

Main reference Madura A. Shelton, Lukasz Chmielewski, Niels Samwel, Markus Wagner, Lejla Batina, Yuval Yarom: "Rosita++: Automatic Higher-Order Leakage Elimination from Cryptographic Code", in Proc. of the CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 – 19, 2021, pp. 685–699, ACM, 2021.

URL https://doi.org/10.1145/3460120.3485380

This presentation covers selected topics in Deep Learning (DL) applications to physical attacks, including Side-Channel Analysis (SCA) and Fault Injection (FI). The following topics are covered: horizontal attack against Elliptic Curve Cryptography (ECC) and RSA, XYZ hotspot selection (SCA & FI), attacks against hardware DL accelerators, and DL-based power simulators.

3.3 Backdoor Detection in Federated Learning via Deep Layer Predictions

Alexandra Dmitrienko (Universität Würzburg, DE)

Joint work of Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi
Main reference Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi: "Close the Gate: Detecting Backdoored Models in Federated Learning based on Client-Side Deep Layer Output Analysis", CoRR, Vol. abs/2210.07714, 2022.
URL https://doi.org/10.48550/arXiv.2210.07714

This talk discusses the challenges of backdoor detection in federated learning (FL) related to adaptive attackers and non-independent and identically distributed (non-IID) data. It then presents an approach to identify backdoored local contribution of FL clients by analyzing local client predictions of deep learning layers and comparing those to predictions made by a global model. The approach can handle an extended non-IID scenarios compare to the related work and is resilient to adaptive adversaries.

3.4 Breaking cryptographic algorithms using power and EM side-channels

Elena Dubrova (KTH Royal Institute of Technology – Kista, SE)

Side-channel attacks are one of the most efficient physical attacks against implementations of cryptographic algorithms at present. They exploit the correlation between physical measurements (power consumption, electromagnetic emissions, timing) taken at different points during the algorithm's execution and the secret key. In this talk, I will give an introduction to power and EM-based side-channel attacks and present some of our recent results.

3.5 Blockchain tools for privacy-preserving machine learning

Oğuzhan Ersoy (TU Delft, NL)

In recent years, blockchain technology get the attention of both industry and academia. Thanks to the interest, there are several cryptographic tools developed for decentralized systems that can be used in other domains including machine learning. Among these tools, VDF, VRF, and adaptor signatures are mentioned in this talk. Firstly, VDFs (Verifiable Delay Functions) allow a prover to show a verifier that a certain amount of time running a function was spent. In a machine learning setting, VDFs can be used to limit the number of queries on a machine learning model. Specifically, by requesting parties to provide VDF proofs when they query the model, we can restrict the number of queries sent to the system. Compared to proof-of-work-based techniques [1], VDF-based query limitations would also guarantee that the adversary cannot parallelize the VDF challenge. Secondly, VRFs (Verifiable Random

Functions) are used to generate random numbers that can be verifiable by all parties involved. In collaborative machine learning, this can be used, for example, cryptographic sortition and leader selection [2]. In this selection, an adversary would not be able to predict the leader in advance. Finally, adaptor signatures allow parties to embed a condition into the signature. It has been used to improve the fungibility of transactions in the blockchain domain. However, it is yet an open question how to utilize adaptor signatures in the machine learning domain.

References

1 Adam Dziedzic; Muhammad Ahmad Kaleem; Yu Shen Lu; Nicolas Papernot, *Increasing the Cost of Model Extraction with Calibrated Proof of Work*, International Conference on Learning Representations (ICLR), 2022.
2 Rui Wang; Oğuzhan Ersoy; Hangyu Zhu; Yaochu Jin; Kaitai Liang, *FEVERLESS: Fast and Secure Vertical Federated Learning based on XGBoost for Decentralized Labels*, IEEE Transactions on Big Data, 1–15, 2022.

3.6 Mitigating Backdoor Attacks in Federated Learning (FL) using Frequency Analysis of the Local Model updates

Hossein Fereidooni (TU Darmstadt, DE) and Ahmad-Reza Sadeghi (TU Darmstadt, DE)

Joint work of Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Ahmad-Reza Sadeghi

Federated learning (FL) is a distributed machine learning technique enabling participating clients to collaboratively learn a shared global model without sharing their potentially private data. Despite its benefits (i.e., communication efficiency and reduced requirements for hardware), federated learning has been shown to be vulnerable to adversarial threats such as backdoor attacks where the adversary stealthily manipulates the global model so that adversary-selected inputs result in adversary-selected outputs. Although there are multiple defense mechanisms proposed by previous works, the backdoor attacks with sophisticated hiding techniques still pose a threat to FL. Existing defense solutions cannot fully mitigate backdoor attacks and have a number of deficiencies such as unrealistic assumptions for data distributions and attack strategies. The core idea of this talk is that backdoored model might be related to frequency analyses of neural networks. We are going to we investigate a relationship between backdoor and frequency components of the model parameters (i.e., weights) that can be used for model filtering during the aggregation process in FL to implement backdoor attack defense. More specifically, we set up the FL process and implement state-of-the-art backdoor attacks (i.e., Semantic attack, Stealthy Model Poisoning, etc.) and then transform tensor weights (i.e., local model updates) to the frequency domain and apply frequency analysis (i.e., Discrete Cosine Transform – DCT) to find a relationship between backdoor patterns and frequency components of the weights.

References

1 E. Bagdasaryan, Andreas Veit, Yiqing Hua, D. Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In AISTATS, 2020.
2 A. Bhagoji, Supriyo Chakraborty, Prateek Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. In ICML, 2019.
3 Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. ArXiv, abs/1808.04866, 2018.

4 Tianyu Gu, Brendan Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. ArXiv, abs/1708.06733, 2017.
5 Hongyi Wang, Kartik K. Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. ArXiv, abs/2007.05084, 2020.
6 Chen Wu, Xiangwen Yang, Sencun Zhu, and P. Mitra. Mitigating backdoor attacks in federated learning. ArXiv, abs/2011.01767, 2020.
7 Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. ArXiv, abs/1807.01251, 2019.
8 Yi Zeng, Won Park, Z. M. Mao, and R. Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. ArXiv, abs/2104.03413, 2021.

3.7 Neural Networks: predators and prey

Fatemeh Ganji (Worcester Polytechnic Institute, US)

Joint work of Fatemeh Ganji, Domenic Forte, Rabin Acharya, Mohammad Hashemi, Steffi Roy
Main reference Rabin Yu Acharya, Fatemeh Ganji, Domenic Forte: "Information Theory-based Evolution of Neural Networks for Side-channel Analysis", IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2023(1), pp. 401–437, 2023.
URL https://doi.org/10.46586/tches.v2023.i1.401-437
Main reference Mohammad Hashemi, Steffi Roy, Domenic Forte, Fatemeh Ganji: "HWGN^2: Side-Channel Protected NNs Through Secure and Private Function Evaluation", in Proc. of the Security, Privacy, and Applied Cryptography Engineering – 12th International Conference, SPACE 2022, Jaipur, India, December 9-12, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13783, pp. 225–248, Springer, 2022.
URL https://doi.org/10.1007/978-3-031-22829-2_13

This talk covers two main topics relevant to how neural networks (NNs) have become a powerful tool to assess the security of cryptographic primitives and how NNs themselves have been targeted to extract their assets. The first part of the talk is devoted to NN-enabled side-channel analysis (SCA), in particular, profiled SCA that leverages leakage from cryptographic implementations to extract the secret key. It is known that when combined with advanced methods in NNs, profiled SCA can successfully attack even crypto-cores with protection devised to impair the effectiveness of SCA. Similar to other machine learning tasks, a range of questions have remained unanswered about NN-enabled SCA, namely: how to choose an NN with an adequate configuration, how to tune the NN's hyperparameters, when to stop the training, etc. This talk introduces "InfoNEAT," which tackles these issues in a natural way. InfoNEAT relies on the concept of neural structure search (NAS), enhanced by information-theoretic metrics to guide the evolution, halt it with novel stopping criteria, and improve time-complexity and memory footprint. Besides the considerable advantages regarding the automated configuration of NNs, InfoNEAT demonstrates significant improvements over other approaches for effective key recovery in terms of the number of epochs and the number of attack traces compared to both MLPs and CNNs, as well as a reduction in the number of trainable parameters compared to MLPs. Furthermore, through experiments, it is demonstrated that InfoNEAT's models are robust against noise and desynchronization in traces.

In the second part of the talk, SCA against NNs has been taken into account. In fact, recent work has highlighted the risks of intellectual property (IP) piracy of deep learning (DL) models from the side-channel leakage of DL hardware accelerators. In response, fundamental cryptographic approaches, specifically built upon the notion of multi-party computation,

could potentially improve the robustness against side-channel leakage. To examine this and weigh the costs and benefits, we introduce hardware garbled NN (HWGN^2), a DL hardware accelerator implemented on FPGA. HWGN^2 also provides NN designers with the flexibility to protect their IP in real-time applications, where hardware resources are heavily constrained, through a hardware-communication cost trade-off. Concretely, we apply garbled circuits, implemented using a MIPS architecture that achieves up to 62.5× fewer logical and 66× less memory utilization than the state-of-the-art approaches at the price of communication overhead. Further, the side-channel resiliency of HWGN^2 is demonstrated by employing the test vector leakage assessment (TVLA) test against both power and electromagnetic side channels.

3.8 AI for Cybersecurity: a taste of things to come... or papers of future past?

Domagoj Jakobovic (University of Zagreb, HR)

Designing a secure system requires a lot of expertise in the security domain. In that process, some of the tasks can be automated with the help of Artificial Intelligence (AI). The use of AI methods does not aim to replace the human designer; rather, they can help in the design optimization process, where standardized algorithms can be readily applied to increase the efficiency. As long as a complex system design task can be decomposed into simpler elements, AI methods can substantially facilitate the optimization of individual components. Furthermore, most methods can be used to optimize an arbitrary (set of) design criteria.

However, although there are problems that can be efficiently solved with AI techniques, it is not always obvious *which* AI technique or optimization algorithm should be applied. In practice, a bit of knowledge in both domains is needed to select the appropriate method and to efficiently apply it to the problem at hand. Even then, for many AI methods there are no formal guarantees of efficiency, which is especially evident for obscure machine learning models such as deep neural networks.

Ideally, the AI component should provide *explainability*, so the decision making process can be justified at each step. We may even employ less efficient but explainable models to evaluate obscure models which bring performance. There are use cases in which a part of a black-box model may be replaced with an equivalent white-box component offering the same level of performance. Additionally, different optimization algorithms may be used to prune "fat" models, either to provide insight into their functionality or to reduce application complexity. In this regard, neuroevolution methods may be used to design and optimize the structure and hyperparameters of deep neural models.

The application of the above techniques can be found in model building efforts in various domains; the usual goals are knowledge representation, model parameter optimization, feature extraction and selection, etc. Some of the efficient examples of this paradigm are already evident in cryptology and security where different AI techniques, most notably evolutionary algorithms, have been applied. Here, the focus was mainly on the design of different cryptography primitives, such as Boolean functions, S-boxes and pseudo-random number generators. Successful applications also include fault injection, intrusion detection, hyper-parameter optimization etc. Recently, evolutionary algorithm methods have also been applied to fuzzing, where they obtained competitive performance in a target-based comparison with commonly used solutions.

3.9 Hardware Security and Deep Learning

Dirmanto Jap (Nanyang TU – Singapore, SG)

Joint work of Yoo-Seung Won, Xiaolu Hou, Dirmanto Jap, Jakub Breier, Shivam Bhasin, Soham Chatterjee, Arindam Basu Leijla Batina, Stjepan Picek

Main reference Yoo-Seung Won, Xiaolu Hou, Dirmanto Jap, Jakub Breier, Shivam Bhasin: "Back to the Basics: Seamless Integration of Side-Channel Pre-Processing in Deep Neural Networks", IEEE Trans. Inf. Forensics Secur., Vol. 16, pp. 3215–3227, 2021.

URL https://doi.org/10.1109/TIFS.2021.3076928

Main reference Yoo-Seung Won, Soham Chatterjee, Dirmanto Jap, Arindam Basu, Shivam Bhasin: "DeepFreeze: Cold Boot Attacks and High Fidelity Model Recovery on Commercial EdgeML Device", in Proc. of the IEEE/ACM International Conference On Computer Aided Design, ICCAD 2021, Munich, Germany, November 1-4, 2021, pp. 1–9, IEEE, 2021.

URL https://doi.org/10.1109/ICCAD51958.2021.9643512

Main reference Lejla Batina, Shivam Bhasin, Dirmanto Jap, Stjepan Picek: "CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel", in Proc. of the 28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019, pp. 515–532, USENIX Association, 2019.

URL https://www.usenix.org/conference/usenixsecurity19/presentation/batina

In this presentation, we provided the discussion on two main direction on the area of hardware security and deep learning (DL). First, we discussed about the use of feature extraction or pre-processing techniques, which could help improving the performance of DL based side-channel attacks (SCA). In most of the research works done, the main goal is towards the direction of designing an efficient network that can provide the best attacks against each side-channel trace dataset. On the other hand, little work has been done to investigate the possibility of strengthening DL architecture with the capability of integrating existing side-channel pre-processing or filtering techniques, which have been thoroughly investigated over the past decades. As such, one of the aim is to minimize the necessity for architecture adjustments while enabling seamlessly integration of pre-processing method for attack. In our work, we propose to incorporate feature extraction and classification in a single framework by using a multi-branch model. The experimental results indicated that the model can perform better than the benchmark model even though it is not specifically tailored for the dataset. These show that it is an inherent property of MCNN which allows it to learn more feature representations and result in better attacks. As for the potential future direction, we discussed the possibility of using other DL based approach as a way to further automate the feature pre-processing method.

Next, we discussed about the vulnerability of DL implementation on physical device against side-channel and fault attacks. Due to the rapid growth of DL application, more and more efforts are being allocated to build and train critical DL models. These DL models have then become valuable Intellectual Properties (IPs) that cost companies lots of time and resources, which inadvertently attract malicious parties to steal them. We presented the work on model extraction and reverse engineering of the neural networks model through electromagnetic (EM) side-channel leakage. We also presented alternative work for reverse engineering of neural network models through cold boot attacks. The work is then conducted targeting edge AI hardware accelerators, Intel Neural Compute Stick 2 (NCS2). It is based on the observation that the model architecture and parameters have to be loaded to Intel NCS2 before the inference, and thus, by performing cold boot attack on host device, it is possible to recover the information, albeit with correction required. As for potential future direction, we proposed to investigate different target devices or more complex architectures. We also discussed on possible countermeasures for the implementation as well as the security evaluation of these countermeasures.

3.10 AI for fault injection

Marina Krcek (TU Delft, NL)

Joint work of Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek
Main reference Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek: "The More You Know: Improving Laser Fault Injection with Prior Knowledge", in Proc. of the Workshop on Fault Detection and Tolerance in Cryptography, FDTC 2022, Virtual Event / Italy, September 16, 2022, pp. 18–29, IEEE, 2022.
URL https://doi.org/10.1109/FDTC57191.2022.00012

Fault injection types such as laser FI, electromagnetic FI, or voltage glitching have different parameters to define. Nevertheless, the parameter search space becomes large for all types because of many parameters and possibilities. Since the search space is large, commonly used methods like grid and random search lead to suboptimal performance/results. We use AI techniques discussed in this talk to improve the efficiency of the search. Specifically, genetic and memetic algorithms from evolutionary computation were shown to find more parameter combinations that lead to erroneous outputs compared to random search [1]. Additionally, hyperparameter tuning methods like successive halving and reinforcement learning from the machine learning domain were also shown to be quite successful [2, 3]. On the other hand, machine learning can be helpful for transferability issues in fault injection. As discussed during the talk, we can use prior knowledge from tested devices and parameter combinations generalized with decision trees to find more vulnerabilities on a new target or bench in the same amount of tested parameters.

References

1 Maldini, Antun; Samwel, Niels; Picek, Stjepan; Batina, Lejla, Genetic algorithm-based electromagnetic fault injection. In: 2018 Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC). IEEE, 2018. p. 35-42.
2 Werner, Vincent; Maingault, Laurent; Potet, Marie-Laure, Fast calibration of fault injection equipment with hyperparameter optimization techniques. In: Smart Card Research and Advanced Applications: 20th International Conference, CARDIS 2021, Lübeck, Germany, November 11–12, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022. p. 121-138.
3 Moradi, Mehrdad; Oakes, Bentley James; Saraoglu, Mustafa; Morozov, Andrey; Janschek, Klaus; Denil, Joachim, Exploring fault parameter space using reinforcement learning-based fault injection. In: 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2020. p. 102-109.

3.11 Assessing the Trustworthiness of AI Sytems

Jesus Luna Garcia (Robert Bosch GmbH – Stuttgart, DE)

Joint work of Parmar, Manojkumar Somabhai; Serna, Jetzabel
Main reference European Commission, On Artificial Intelligence – A European approach to excellence and trust. 2020.
URL https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Despite the topic of AI (cyber)security has received lots of academic and industrial attention in recent years, these communities have started to realize the need for a holistic approach related to this topic. We do not mean only from a system perspective, where the different

implementation layers (e.g., cloud) also contribute to the security (and even to the attack surface) of the AI-application, but also from equally important features like privacy, transparency/explainability, bias, and safety (to name just a few). Extrapolating relevant security research to this much needed holistic approach is critical for the uptake of trusted AI system. This talk discusses some relevant industrial and regulation-related aspects on the field of AI trustworthiness, along with few identified challenges which are being tackled from an EU perspective. One of the main points relates to the need of developing a framework for supporting the assessment of AI systems for cybersecurity certification purposes. The referred framework should be able to leverage realistic levels of automation which can pave the road for continuous (automated) certification. It is expected that such a framework might provide support for accelerating the uptake of relevant standards and regulations like the EU AI Act.

3.12 Use cases of side-channel data analysis

Damien Marion (IRISA – Rennes, FR)

Joint work of Damien Marion, Duy-Phuc Pham, Annelie Heuser

Abstract. In this talk, we went through different use cases of side- channel analysis for different security purposes. The first use case was the analysis of micro-architectural leakage, in order to address the gap between leakage and unknown micro-architectures. The second use case was the usage of electromagnetic leakage to classify and/or detect malware and rootkits[1, 2]. Then the talk quickly tackled some problems of securing PQ-cryptography from side-channel point of view. From a more general point of view, side-channel analysis could be viewed as a subpart of data analysis for security. How to extract or quantify sensitive information present in huge amounts of noise data, this where IA (or machine learning) can solve existing issues.

References

1 Duy-Phuc Pham, Damien Marion, and Annelie Heuser, ULTRA: Ultimate Rootkit Detection over the Air. 25th International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2022, Limassol, Cyprus, October 26-28, 2022 (2022)
2 Duy-Phuc Pham, Damien Marion, Matthieu Mastio, and Annelie Heuser, Obfuscation Revealed: Leveraging Electromagnetic Signals for Obfuscated Malware Classification. ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 – 10, 2021 (2021)

3.13 New Directions in AI-Based Cryptography

Luca Mariot (Radboud University Nijmegen, NL)

Main reference Luca Mariot, Domagoj Jakobovic, Thomas Bäck, Julio Hernandez-Castro: "Artificial Intelligence for the Design of Symmetric Cryptographic Primitives", pp. 3–24, 2022.
URL https://doi.org/10.1007/978-3-030-98795-4_1

In this talk, we give a general overview of AI methods and computational models to design cryptographic primitives. These include the use of bio-inspired optimization techniques (particularly evolutionary algorithms) to construct symmetric primitives with good cryptographic properties, like Boolean functions and S-boxes. The approach leverages also on the

use of AI computational models like Cellular Automata (CA) as an efficient representation technique for such primitives. In the second part of the talk, new directions of research are illustrated based on the experience gained with regard to the above AI methods and models. In particular, we focus on the use of evolutionary algorithms to design algebraic constructions of symmetric primitives, to evolve differential distinguishers for small symmetric ciphers, and to explore the space of adversarial examples in machine learning models. Particular emphasis is given to the inherent interpretability and explainability of the solutions provided by evolutionary algorithms, specifically in the case of Genetic Programming (GP).

3.14 High-throughput network intrusion detection based on deep learning

Nele Mentens (Leiden University, NL)

Joint work of Nele Mentens, Laurens Le Jeune

The evolution of our digital society relies on networks that can handle an increasing amount of data, exchanged by an increasing number of connected devices at an increasing communication speed. With the growth of the online world, criminal activities also extend onto the Internet. Network Intrusion Detection Systems (NIDSs) detect malicious activities by analyzing network data. While neural network-based solutions can effectively detect various attacks in an offline setting, it is not straightforward to deploy them in high-bandwidth online systems. This talk elucidates why Field-Programmable Gate Arrays (FPGAs) are the preferred platforms for online network intrusion detection, and which challenges need to be overcome to develop FPGA-based NIDSs for Terabit Ethernet networks.

3.15 Fuzz testing with machine learning

Irina Nicolae (Robert Bosch GmbH, Bosch Center for AI – Stuttgart, DE)

Joint work of Maria-Irina Nicolae, Max Eisele, Andreas Zeller

Fuzzing – testing software and hardware with randomly generated inputs – has gained significant traction due to its success in exposing program vulnerabilities automatically. Machine learning has increasingly been applied to different parts of the fuzzing loop, with the goal of improving fuzzing efficiency. In this talk, we examine *neural program smoothing* for fuzzing, a family of methods that approximate the tested program with a neural network for novel test case generation. We uncover fundamental and practical limitations of neural program smoothing, which prevent it from reaching its advertised performance and limit its practical interest.

3.16 Explainability of deep learning-based side-channel analysis

Stjepan Picek (Radboud University, NL)

Joint work of Guilherme Perin, Lichao Wu, Stjepan Picek
Main reference Guilherme Perin, Lichao Wu, Stjepan Picek: "I Know What Your Layers Did: Layer-wise Explainability of Deep Learning Side-channel Analysis", IACR Cryptol. ePrint Arch., p. 1087, 2022.
URL https://eprint.iacr.org/2022/1087

Deep learning-based side-channel analysis is an extremely powerful option as it can work without feature engineering and defeats various hiding and masking countermeasures. Still, from the evaluator's perspective, even after a successful evaluation (attack), a crucial detail is missing: how did the neural network break the target? Thus, the explainability of deep learning-based side-channel analysis becomes an important issue. Unfortunately, up to now, there are only sporadic attempts to understand how neural network defeats countermeasures and none that gives the complete answer. Some early explored techniques include SVCCA [1] and ablation [2]. While good first steps, these techniques do not provide enough information to understand how countermeasures are circumvented. This talk concentrated on a recent approach to explaining the deep learning-based side-channel attack: layer-wise explainability and its comparative advantages over previous approaches.

References

1 Daan van der Valk, Stjepan Picek, Shivam Bhasin: Kilroy Was Here: The First Step Towards Explainability of Neural Networks in Profiled Side-Channel Analysis. COSADE 2020: 175-199.
2 Lichao Wu, Yoo-Seung Won, Dirmanto Jap, Guilherme Perin, Shivam Bhasin, Stjepan Picek: Explain Some Noise: Ablation Analysis for Deep Learning-based Physical Side-channel Analysis. IACR Cryptol. ePrint Arch. 2021: 717 (2021).

3.17 Engineering Models versus Scientific Models

Patrick Schaumont (Worcester Polytechnic Institute, US)

Cybersecurity implementations, in hardware or software, are created from engineering models, not from scientific models. Scientific models reflect the laws of nature in formulae, while engineering models aim at the opposite: we use the laws of nature to mimic an abstraction.

The observations of secure implementations in the real world are noisy distortions from the ideal, noiseless engineering models. However, we *know* that the ground truth corresponds to the engineering model, which is noiseless and undistorted.

This has an important consequence on machine learning applications. We can use simulation (of engineering models) to create a ground truth to improve inference on measured, distorted implementation. For example, using simulated data, we can build attacks on real-world systems that outperform real-world measurements [1].

References

1 Dillibabu Shanmugam, Patrick Schaumont, "Improving Side-channel Leakage Assessment using Pre-silicon Leakage Models," 14th International Workshop on Constructive Side-channel Analysis and Secure Design (COSADE 2023), Munch, Germany, April 2023.

3.18 Remote Electrical-Level Attacks on Cloud FPGAs: The Role of AI

Mirjana Stojilović (EPFL – Lausanne, CH)

Main reference Ognjen Glamocanin, Louis Coulon, Francesco Regazzoni, Mirjana Stojilovic: "Are Cloud FPGAs Really Vulnerable to Power Analysis Attacks?", in Proc. of the 2020 Design, Automation & Test in Europe Conference & Exhibition, DATE 2020, Grenoble, France, March 9-13, 2020, pp. 1007–1010, IEEE, 2020.
URL https://doi.org/10.23919/DATE48585.2020.9116481

Field-programmable gate arrays (FPGAs) have made their way into the cloud, allowing users to gain remote access to the state-of-the-art reconfigurable fabric and implement their custom accelerators. As FPGAs are large enough to accommodate multiple independent designs, the multi-tenant user scenario may soon be prevalent in cloud computing environments. However, shared FPGAs are vulnerable to remote power-side channel and fault-injection attacks [1, 3, 4]. Machine learning (ML) further broadens the attack space: (1) ML accelerators may be the targets of remote attacks, (2) ML techniques can be used to infer the type of workloads or the computations the FPGA is running [2], and (3) ML can help detecting malicious circuits in FPGA bitstreams. This talk has two parts: In the first, the techniques enabling remote electrical-level attacks on cloud FPGAs are explained. In the second, the opportunities for using ML for detecting and locating malicious activity, or for guiding the cloud hypervisors in managing the FPGA users in a security-aware manner are discussed.

References

1 Ognjen Glamočanin; Louis Coulon; Francesco Regazzoni; Mirjana Stojilović, *Are Cloud FPGAs Really Vulnerable to Power Analysis Attacks?*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1007–10, 2020.
2 Ognjen Glamočanin; Hajira Bazaz; Mathias Payer; Mirjana Stojilović, *Temperature Impact on Remote Power Side-Channel Attacks on Shared FPGAs*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1–6, 2023.
3 Dina G. Mahmoud; Samah Hussein; Vincent Lenders; Mirjana Stojilović, *FPGA-to-CPU Undervolting Attacks*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 999–1004, 2022.
4 Dina G. Mahmoud; Mirjana Stojilović, *Timing Violation Induced Faults in Multi-Tenant FPGAs*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1745–50, 2019.

3.19 AI-Assisted System-level Tamper Detection

Shahin Tajik (Worcester Polytechnic Institute, US)

Joint work of Shahin Tajik, Tahoura Mosavirik, Patrick Schaumont

To mount physical attacks adversaries might need to place probes in the proximity of the integrated circuits (ICs) package, create physical connections between their probes/wires and the system's PCB, or physically tamper with the PCB's components, chip's package, or substitute the entire PCB to prepare the device for the attack. In this talk, inspired by methods known from the field of power integrity analysis, we show how the impedance

characterization of the system's power distribution network (PDN) using an on-chip circuit-based network analyzer can detect various categories of tamper events. By analyzing the frequency response of the system different classes of tamper events from board to chip level are revealed. Using the Wasserstein Distance as a metric, we demonstrate that we can confidently detect tamper events. We demonstrate that even environment-level tampering activities, e.g., proximity of contactless EM probes to the IC package or slightly polished IC package, can be detected using on-chip impedance sensing.

3.20 Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis

Trevor Yap (Nanyang TU – Singapore, SG)

Joint work of Trevor Yap, Andrien Benamira, Shivam Bhasin, Thomas Peyrin
Main reference Trevor Yap, Adrien Benamira, Shivam Bhasin, Thomas Peyrin: "Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis", 2022.
URL https://eprint.iacr.org/2022/1247

Deep neural networks (DNN) have become a significant threat to the security of cryptographic implementations with regards to side-channel analysis (SCA), as they automatically combine the leakages without any preprocessing needed, leading to a more efficient attack. However, these DNNs for SCA remain mostly black-box algorithms that are very difficult to interpret. Benamira *et al.* recently proposed an interpretable neural network called Truth Table Deep Convolutional Neural Network (TT-DCNN), which is both expressive and easier to interpret. In particular, a TT-DCNN has a transparent inner structure that can entirely be transformed into SAT equations after training. This talk gives a brief outline of why we need explainability, and on what TT-DCNN is. The talk also presented a way to analyse the SAT equations of TT-DCNN and show some results. Furthermore, we give a possible direction to analyse this paper.

4 Working Groups

4.1 Explainability of AI in Cybersecurity

Stjepan Picek (Radboud University, NL)
Nele Mentens (Leiden University, NL)

The explainability of AI in cybersecurity represents an important problem since often, it is not sufficient to only have a successful solution. Still, we also must explain why that solution works. For instance, in side-channel analysis, from the perspective of a security evaluator, it is important to know how secure a target is. But, if the target gets broken, a necessary step is to report back to the implementation designers and explain what went wrong (e.g., how a countermeasure got broken). Unfortunately, while deep learning can break various targets, the explainability part is still very much unexplored and vague [6, 4, 7]. For instance, in deep learning-based side-channel analysis, the state-of-the-art approaches can easily break

implementations protected with various countermeasures (masking, hiding, or a combination of masking and hiding). At the same time, understanding why the attack works is based on intuition or general terms from the machine learning domain, e.g., desynchronization is defeated due to the spatial invariance of convolutional neural networks.

Furthermore, deep learning has recently been shown to be a very powerful option in mounting cryptanalysis attacks where the neural networks serve as distinguishers. More precisely, the differential-neural distinguishers are based on distinguishing ciphertext-pairs that belong to a fixed plaintext difference from random ones. While the approach works well, and for several ciphers, the researchers managed to find attacks that are at least competitive with classical approaches. Unfortunately, even after the successful attack, the question remains why the attack works and how to fix the cipher to make it more secure. Works addressing such issues are sparse and far from conclusive [2, 1, 3, 5].

The discussion centered on two questions we consider at the core of explainability. Finally, the discussion from this group was also connected with other discussion groups since explainability is of relevance whenever applying AI in cybersecurity.

- Why?
  - To improve the model: more efficient implementation, more powerful in solving the intended task (e.g., getting the key, increasing the performance metrics, lowering the number of false alarms), more efficient test cases for fuzzing.
  - To improve the security of the implementation against attacks (e.g., SCA, crypto): understand the vulnerabilities of the implementation under attack, fix the implementation based on the position of the leakage, and fix the countermeasure based on the discovered vulnerabilities.
  - To improve trust in the model: important in intrusion detection systems, lower the number of false positives and false negatives, enable application in online systems.
  - To contribute to the security of AI: discover which parts are weak against backdoors, etc.
- How?
  - Understand the features:
    - Feature visualization: activation maximization, code inversion.
    - Feature attributions: LIME, occlusion, delivery maps, Shapley values.
    - Rule extraction: DeepRed, scalability challenges (data, model).
  - Understand the neural network:
    - ablation.
    - SVCCA.
    - layer-wise explainability for side-channel analysis.

References

1 Aron Gohr, Gregor Leander, Patrick Neumann: An Assessment of Differential-Neural Distinguishers. IACR Cryptol. ePrint Arch. 2022: 1521 (2022).

2 Aron Gohr: Improving Attacks on Round-Reduced Speck32/64 Using Deep Learning. CRYPTO (2) 2019: 150-179.

3 Adrien Benamira, David Gérault, Thomas Peyrin, Quan Quan Tan: A Deeper Look at Machine Learning-Based Cryptanalysis. EUROCRYPT (1) 2021: 805-835.

4 Trevor Yap, Adrien Benamira, Shivam Bhasin, Thomas Peyrin: Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2023(2), 24–53. https://doi.org/10.46586/tches.v2023.i2.24-53.

5 Nicoleta-Norica Bacuieti, Lejla Batina, Stjepan Picek: Deep Neural Networks Aiding Cryptanalysis: A Case Study of the Speck Distinguisher. ACNS 2022: 809-829.
6 Lichao Wu, Yoo-Seung Won, Dirmanto Jap, Guilherme Perin, Shivam Bhasin, Stjepan Picek: Explain Some Noise: Ablation Analysis for Deep Learning-based Physical Side-channel Analysis. IACR Cryptol. ePrint Arch. 2021: 717 (2021).
7 Guilherme Perin, Lichao Wu, Stjepan Picek: I Know What Your Layers Did: Layer-wise Explainability of Deep Learning Side-channel Analysis. IACR Cryptol. ePrint Arch. 2022: 1087 (2022).

4.2 AI for Implementation Attacks

Stjepan Picek (Radboud University, NL)
Nele Mentens (Leiden University, NL)

Implementation attacks aim at the weaknesses of the implementation and not the algorithm. The most common options for implementation attacks are side-channel attacks and fault injection attacks. In both domains, AI is used extensively. In side-channel attacks, it is common to use machine learning in the profiling attack scenario. There, the attacker has a copy of the device to be attacked under control and uses it to build a model of a device. Later, the model is used to attack the target and obtain secret information. Machine learning attacks in such a setup have been used for over a decade and show excellent attack performance. More recent results with deep learning provide even better attack performance against targets protected with countermeasures and with no need to conduct feature engineering [3]. Still, multiple open issues need to be resolved. For instance, the attacks assume that the attacker has access to a copy of a device to be attacked, which is often not a realistic assumption. As such, one of the big challenges to be solved is how to mount non-profiling deep learning-based attacks [2]. Next, leakage assessment is important as it provides the first information on whether the target has secure implementation or if there is some leakage. The results with deep learning are promising but sparse [4]. Mounting an attack once the device is produced is a common setup but results in large expenses for manufacturers once security vulnerabilities are detected. As such, it is important to understand whether we can use various simulation-based approaches and techniques to construct synthetic measurements to assess the security of devices even before they are produced [5, 8]. Finally, as previously discussed, the explainability perspective is important for side-channel attacks. While most of the AI-based approaches for side-channel analysis use machine (deep) learning, there are also some efforts in feature engineering or hyperparameter tuning [1, 7]. More open challenges discussed during the workshop can be found in [6].

On the other hand, in fault injection, AI is mostly used to allow fast characterization of the target (cartography). In that context, various evolutionary and local search algorithms are used [11, 9]. More recently, deep learning is also used to predict if a point on a target will result in a faulty response [10]. We identified the research gaps in making the approaches more stable and maintaining the balance between exploring various regions of the target and fast convergence to a region with many faulty responses. Finally, research rarely explores how to use the located faults in mounting the attacks (which could help understand if all located faults are equally important).

Finally, implementation attacks can be used to attack machine learning, connecting this topic with the security of AI [12, 13].

References

1 Karlo Knezevic, Juraj Fulir, Domagoj Jakobovic, Stjepan Picek, Marko Đurasevic: NeuroSCA: Evolving Activation Functions for Side-Channel Analysis. IEEE Access 11: 284-299.
2 Lichao Wu, Guilherme Perin, Stjepan Picek: Hiding in Plain Sight: Non-profiling Deep Learning-based Side-channel Analysis with Plaintext/Ciphertext. IACR Cryptol. ePrint Arch. 2023: 209 (2023).
3 Guilherme Perin, Lichao Wu, Stjepan Picek: Exploring Feature Selection Scenarios for Deep Learning-based Side-channel Analysis. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2022(4): 828-861 (2022).
4 Thorben Moos, Felix Wegener, Amir Moradi: DL-LA: Deep Learning Leakage Assessment A modern roadmap for SCA evaluations. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2021(3): 552-598 (2021).
5 Omid Bazangani, Alexandre Iooss, Ileana Buhan, Lejla Batina: ABBY: Automating the creation of fine-grained leakage models. IACR Cryptol. ePrint Arch. 2021: 1569 (2021).
6 Stjepan Picek, Guilherme Perin, Luca Mariot, Lichao Wu, Lejla Batina: SoK: Deep Learning-based Physical Side-channel Analysis. ACM Computing SurveysVolume 55Issue 11Article No.: 227pp 1–35.
7 Unai Rioja, Lejla Batina, Jose Luis Flores, Igor Armendariz: Auto-tune POIs: Estimation of distribution algorithms for efficient side-channel analysis. Comput. Networks 198: 108405 (2021)
8 Naila Mukhtar, Lejla Batina, Stjepan Picek, Yinan Kong: Fake It Till You Make It: Data Augmentation Using Generative Adversarial Networks for All the Crypto You Need on Small Devices. CT-RSA 2022: 297-321
9 Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek: The More You Know: Improving Laser Fault Injection with Prior Knowledge. FDTC 2022: 18-29.
10 Lichao Wu, Gerard Ribera, Noemie Beringuier-Boher, Stjepan Picek: A Fast Characterization Method for Semi-invasive Fault Injection Attacks. CT-RSA 2020: 146-170.
11 Rafael Boix Carpi, Stjepan Picek, Lejla Batina, Federico Menarini, Domagoj Jakobovic, Marin Golub: Glitch It If You Can: Parameter Search Strategies for Successful Fault Injection. CARDIS 2013: 236-252.
12 Lejla Batina, Shivam Bhasin, Dirmanto Jap, Stjepan Picek: CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel. USENIX Security Symposium 2019: 515-532.
13 Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin, Yang Liu: Practical Fault Attack on Deep Neural Networks. CCS 2018: 2204-2206.

4.3 Security Fuzzing

Stjepan Picek (Radboud University, NL)

Vulnerabilities caused by programming errors are a major threat to today's programs. For instance, memory corruption vulnerabilities can lead to uncontrolled behavior in the program, which attackers can often abuse. A modern strategy to uncover such programming errors is automated software testing using fuzz testing (fuzzing). Fuzzing automatically generates inputs from testcases and feeds them to the program under test while monitoring it. If a programming error has been reached, the fuzzer notices that the program hangs or crashes. Mutational fuzzing requires a set of program inputs (seeds) that can be obtained from

testcases or real inputs. The process of mutation can be influenced by 1) the location in the input that gets mutated and 2) the mutation that is applied, with the selection done randomly or guided by a heuristic. A common option is to use evolutionary algorithms for such goals [1]. While the approach works well, there are issues. Due to a wide number of available evolutionary algorithms, selecting what algorithm to use and how to customize it for the task is not trivial. Moreover, since evolutionary algorithms are guided through an objective function, appropriate evaluations should be done. Machine learning is also used in fuzzing for various tasks like seed file generation, testcase generation, or mutation operator selection [2]. It is important to understand whether evolutionary algorithms or machine learning produce better results for tasks that can be achieved by both (e.g., mutation operator selection) and in what scenarios to select a specific AI technique. For instance, finding the states in stateful fuzzing is not easy, and machine learning could be used for this task.

References

1 Patrick Jauernig, Domagoj Jakobovic, Stjepan Picek, Emmanuel Stapf, Ahmad-Reza Sadeghi: DARWIN: Survival of the Fittest Fuzzing Mutators. CoRR abs/2210.11783 (2022).
2 Wang Y, Jia P, Liu L, Huang C, Liu Z (2020) A systematic review of fuzzing based on machine learning techniques. PLoS ONE 15(8): e0237749. https://doi.org/10.1371/journal.pone.0237749.

4.4 Security of Machine Learning

Stjepan Picek (Radboud University, NL)

Machine (deep) learning found its place in various real-world applications, where many applications have security requirements. Unfortunately, as these systems become more pervasive, understanding how they fail becomes more challenging. There are several failure modes in machine learning, but one category received significant attention in the last few years: backdoor attacks. 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 property is included in the model input. This property is called the trigger and can be anything the targeted model understands. Deep learning is evaluated in either a centralized or distributed setting. While the centralized one is simpler, it poses privacy concerns due to the need to have the training data available (and, for instance, shared in the case of online training). Then, a common option is to use federated learning as a distributed learning paradigm that works on isolated data. In federated learning, clients can collaboratively train a shared global model under the orchestration of a central server while keeping the data decentralized. Multiple backdoor attacks and defenses exist on machine learning systems (centralized and distributed) and for diverse data types: computer vision (e.g., images, video), sound, text, and graph data. While many observations can be transferred from one setup to another, unique characteristics also require detailed experimentalism [1, 2]. We need more systematic evaluations of diverse attack factors in different domains and with larger (more realistic) datasets and neural network models. Finally, more effort must be given to designing powerful, transferable, and efficient defenses [4, 3].

References

1 Gorka Abad, Oguzhan Ersoy, Stjepan Picek, Aitor Urbieta: Sneaky Spikes: Uncovering Stealthy Backdoor Attacks in Spiking Neural Networks with Neuromorphic Data. CoRR abs/2302.06279 (2023)

2 Jing Xu, Rui Wang, Stefanos Koffas, Kaitai Liang, Stjepan Picek: More is Better (Mostly): On the Backdoor Attacks in Federated Graph Neural Networks. ACSAC 2022: 684-698.

3 Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, Thomas Schneider: FLAME: Taming Backdoors in Federated Learning. USENIX Security Symposium 2022: 1415-1432.

4 Kavita Kumari, Phillip Rieger, Hossein Fereidooni, Murtuza Jadliwala, Ahmad-Reza Sadeghi: BayBFed: Bayesian Backdoor Defense for Federated Learning. CoRR abs/2301.09508 (2023).

Participants

- Ileana Buhan
Radboud University
Nijmegen, NL
- Lukasz Chmielewski
Radboud University Nijmegen,
NL & Masaryk University –
Brno, CZ
- Alexandra Dmitrienko
Universität Würzburg, DE
- Elena Dubrova
KTH Royal Institute of
Technology – Kista, SE
- Oguzhan Ersoy
TU Delft, NL
- Hossein Fereidooni
TU Darmstadt, DE
- Fatemeh Ganji
Worcester Polytechnic
Institute, US
- Houman Homayoun
University of California –
Davis, US
- Domagoj Jakobovic
University of Zagreb, HR
- Dirmanto Jap
Nanyang TU – Singapore, SG
- Florian Kerschbaum
University of Waterloo, CA
- Marina Krcek
TU Delft, NL
- Jesus Luna Garcia
Robert Bosch GmbH –
Stuttgart, DE
- Damien Marion
IRISA – Rennes, FR
- Luca Mariot
Radboud University
Nijmegen, NL
- Nele Mentens
Leiden University, NL
- Irina Nicolae
Bosch Center for AI –
Renningen, DE
- Stjepan Picek
Radboud University
Nijmegen, NL
- Jeyavijayan Rajendran
Texas A&M University –
College Station, US
- Ahmad-Reza Sadeghi
TU Darmstadt, DE
- Patrick Schaumont
Worcester Polytechnic
Institute, US
- Matthias Schunter
INTEL ICRI-SC –
Darmstadt, DE
- Mirjana Stojilović
EPFL – Lausanne, CH
- Shahin Tajik
Worcester Polytechnic
Institute, US
- Trevor Yap
Nanyang TU – Singapore, SG