IRISA

PhD position at IRISA/Inria, Rennes, Taran team

Guess What I'm Learning: Side-Channel Analysis of Edge AI Training Accelerators

Summary. The importance of data privacy in Deep Learning (DL) systems is without doubt. However, recent studies have shown how to use side-channels to guess key parameters or recover inputs from DL model inference running on accelerators. For example, input images of a DL model were estimated from collected power traces without knowing the detailed model parameters [1]. However, this thesis goes further than current practice by studying if private information can be retrieved during training phase, if it is possible to disrupt the training quality, and how to secure training accelerators. The objective is to study training-time, side-channel analysis, hardware attacks and required countermeasures, focusing on fault injections in edge AI accelerators.

Context

As Deep Learning is computationally intensive and power hungry, the use of dedicated and customized hardware accelerators is imposing. This is the case for FPGAs, increasingly adopted to build highly customized and flexible DL accelerators [2–4], including the recent trend on approximate Deep Neural Network (DNN) implementations [5].

The convergence of edge computing and AI brings *Edge Intelligence* [6], which moves the bulk of intelligent data processing from the core of the network to the edge, closer to where data is produced and resides. This therefore reduces latency and increases privacy [7].

However, the connectivity and accessibility of these edge devices enable both local and remote attacks, unveiling an enormous attack surface with large potential impacts on security, safety and privacy. In the context of DL hardware security, recent works are reporting increasing attacks to DNN implementations [8, 9]. These include Side-Channel Analysis (SCA) attacks [9, 10], either using power consumption [1, 11–17] or Electromagnetic (EM) emanations [18–20], and Fault Injection (FI) attacks [21–29]. In the former case, the objective of the attack is to compromise *confidentiality*, enabling the recovery of secret DL assets (like models and private data inputs) that jeopardize privacy and enable counterfeiting by model reverse engineering. In the latter, the objective is to compromise both *integrity*, altering the expected performance through misclassifications and controlled behaviours, and *availability*, rendering the system useless through denied access or reduced quality or performance [30]. Physical SCA and FI attacks to AI-enabled edge devices are particularly worrying given their higher accessibility and exposure to attackers [31].

Distributed training at the edge can be traced back to 2016 [32], where a decentralized Stochastic Gradient Descent (SGD) method is proposed to solve a large linear regression problem. More recently it has evolved to the concept of collaborative or federative learning, which is based on the same general principle but is more efficient [33]. Other techniques for edge training are to train or retrain models on single edge devices, taking advantage of modern training features such as transfer learning, incremental learning, and continuous learning. In all these approaches, local data are processed on each edge device, which prevents the devices from revealing private data to the cloud. However, the server should neither trust edge devices completely, since these can be attacked and forced into abnormal behaviors, which can poison training data. This would thus result in inadequate model updates, and hence in a low-quality trained model. For example, in a backdoor-attacked face recognition-based authentication system, attackers could mislead systems to identify them as a person who can access a building through impersonation [34],[6].

The described scenario is helping rise general concerns on AI trust, which calls for a major research effort to protect critical infrastructures and sensitive data that rely on AI-based processing. As a consequence, protecting DNN implementations is a key concern to keep their models and internal data private and secure from attacks, as this has a large potential for major impacts on privacy, safety and secret corporate IP. To help unlock the full potential of AI and enable efficient and secure deployments, **our objective is to build secure DL hardware accelerators for edge and cloud systems, hence resistant to both local and remote hardware attacks**.

Objectives of the Thesis

The main goals of this thesis are (1) to investigate the implementation vulnerabilities against SCA and FI attacks of custom, reduced-precision hardware implementations of DNN accelerators built in FPGAs and (2) to develop adequate countermeasures to build secure accelerators.

On the FI case, the objective is to investigate how these attacks can impact the integrity and availability of the system (accuracy, training/inference time, energy consumption). We will especially focus on electromagnetic FI using the facilities at the Laboratoire Haute Sécurité (LHS) at Inria Rennes. On the SCA case our objective is to understand how the attacks can impact the confidentiality of the system by revealing key secret information like training/inference inputs and by enabling reverse engineering of DL models and architectures. For local attacks, we focus on capturing power/EM side-channel leakage traces.

As mentioned, we will in particular focus on hardware security of DL accelerators at **training time**, especially in (semi-)supervised, cooperative edge scenarios, through a holistic approach that combines training methodologies, algorithms and design of custom accelerators in FPGA.

Training-time attacks to DNNs have not focused on hardware vulnerabilities, but on datasets to compromise the training, like software adversarial attacks that contaminate the training dataset to increase the misclassification probability at inference time [35]. As already mentioned, security-enhanced edge training includes research on how to secure the communication protocols to avoid data to be corrupted in a federated-learning setting [36–38]. The hardware is assumed to be secured and fault-free. However, when this assumption fails, data and/or model computation can be corrupted, hence harming the global model training result.

This work will take place in the Taran team from IRISA/Inria, in collaboration with CentraleSupélec (Rubén Salvador, IETR) and Inria LHS (Ronan Lashermes).

To get more information and to apply please contact:

- Angeliki Kritikakou: angeliki.kritikakou@irisa.fr
- Olivier Sentieys: olivier.sentieys@irisa.fr

References

[1] L. Wei et al. "I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators". *ACSAC*. ACM, 2018, pp. 393–406. DOI: 10.1145/3274694.3274696.

[2] K. Abdelouahab et al. *Accelerating CNN Inference on FPGAs: A Survey*. May 26, 2018. arXiv: 1806.01683 [cs].

[3] A. G. Blaiech et al. "A Survey and Taxonomy of FPGA-based Deep Learning Accelerators". *J. Syst. Archit.* 98 Sept. 1, 2019, pp. 331–345. DOI: 10.1016/j.sysarc.2019.01.007.

[4] K. Guo et al. "[DL] A Survey of FPGA-based Neural Network Inference Accelerators". *ACM TRETS* 12.1 Mar. 28, 2019, 2:1–2:26. DOI: 10.1145/3289185.

[5] E. Wang et al. "Deep Neural Network Approximation for Custom Hardware: Where We've Been, Where We're Going". *ACM Comput. Surv.* 52.2 May 30, 2019, 40:1–40:39. DOI: 10.1145/3309551.

[6] D. Xu et al. "Edge Intelligence: Empowering Intelligence to the Edge of Network". *Proc. IEEE* 109.11 2021, pp. 1778–1837. DOI: 10.1109/JPROC.2021.3119950.

[7] S. Deng et al. "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence". *IEEE Internet Things J.* 7.8 2020, pp. 7457–7469. DOI: 10.1109/JIOT.2020.2984887.

[8] Q. Xu et al. "Security of Neural Networks from Hardware Perspective: A Survey and Beyond". *ASPDAC*. ACM, 2021, pp. 449–454. DOI: 10.1145/3394885.3431639.

[9] S. Mittal et al. "A Survey on Hardware Security of DNN Models and Accelerators". *J. Syst. Archit.* 117 2021, p. 102163. DOI: 10.1016/j.sysarc.2021.102163.

[10] M. Méndez Real et al. "Physical Side-Channel Attacks on Embedded Neural Networks: A Survey". *Appl. Sci.* 11 15, 2021, p. 6790. DOI: 10.3390/app11156790.

[11] A. Dubey et al. "MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection". *IEEE HOST*. 2020, pp. 197–208. DOI: 10.1109/HOST45689.2020.9300276.

[12] S. Moini et al. "Power Side-Channel Attacks on BNN Accelerators in Remote FPGAs". *IEEE J. Emerg. Sel. Top. Circuits Syst.* 11.2 2021, pp. 357–370. DOI: 10.1109/JETCAS.2021.3074608.

[13] Y. Zhang et al. "Stealing Neural Network Structure Through Remote FPGA Side-Channel Analysis". *IEEE Trans. Inf. Forensics Secur.* 16 2021, pp. 4377–4388. DOI: 10.1109/TIFS.2021.3106169.

[14] S. Tian et al. "Remote Power Attacks on the Versatile Tensor Accelerator in Multi-Tenant FPGAs". *IEEE FCCM*. 2021, pp. 242–246. DOI: 10.1109/FCCM51124.2021.00037.

[15] T. Li et al. "Model Extraction and Adversarial Attacks on Neural Networks Using Switching Power Information". *ICANN*. Springer, 2021, pp. 91–101. DOI: 10.1007/978-3-030-86362-3_8.

[16] K. Yoshida et al. "Model Reverse-Engineering Attack Using Correlation Power Analysis against Systolic Array Based Neural Network Accelerator". *ISCAS*. 2020, pp. 1–5. DOI: 10.1109/ISCAS45731.2020.9180580.

[17] K. Yoshida et al. "Model Reverse-Engineering Attack against Systolic-Array-Based DNN Accelerator Using Correlation Power Analysis". *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* E104-A.1 2021, pp. 152–161. URL: https://search.ieice.org/bin/summary.php?id=e104-a_1_152.

[18] H. Yu et al. "DeepEM: Deep Neural Networks Model Recovery through EM Side-Channel Information Leakage". *IEEE HOST*. 2020, pp. 209–218. DOI: 10.1109/HOST45689.2020.9300274.

[19] K. Yoshida et al. "Model-Extraction Attack Against FPGA-DNN Accelerator Utilizing Correlation Electromagnetic Analysis". *FCCM*. 2019, pp. 318–318. DOI: 10.1109/FCCM.2019.00059.

[20] V. Yli-Mäyry et al. "Extraction of Binarized Neural Network Architecture and Secret Parameters Using Side-Channel Information". *ISCAS*. May 2021, pp. 1–5. DOI: 10.1109/ISCAS51556.2021.9401626.

[21] S. Tajik et al. "Artificial Neural Networks and Fault Injection Attacks". *arXiv* 2020. arXiv: 2008.07072 [cs].

[22] J. Breier et al. "SNIFF: Reverse Engineering of Neural Networks With Fault Attacks". *IEEE Trans Reliab* 2021, pp. 1–13. DOI: 10.1109/TR.2021.3105697.

[23] J. Breier et al. "Practical Fault Attack on Deep Neural Networks". *ACM CCS*. 2018, pp. 2204–2206. DOI: 10.1145/3243734.3278519.

[24] S. Hong et al. "Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks". *USENIX Security Symp*. 2019, pp. 497–514. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/hong.

[25] M. M. Alam et al. "RAM-Jam: Remote Temperature and Voltage Fault Attack on FPGAs Using Memory Collisions". *FDTC*. 2019, pp. 48–55. DOI: 10.1109/FDTC.2019.00015.

[26] X. Hou et al. "Physical Security of Deep Learning on Edge Devices: Comprehensive Evaluation of Fault Injection Attack Vectors". *Microelectron. Reliab.* 120 2021, p. 114116. DOI: 10.1016/j.microrel.2021.114116.

[27] A. Boutros et al. "Neighbors From Hell: Voltage Attacks Against Deep Learning Accelerators on Multi-Tenant FPGAs". *ICFPT*. 2020, pp. 103–111. DOI: 10.1109/ICFPT51103.2020.00023.

[28] Y. Luo et al. "DeepStrike: Remotely-Guided Fault Injection Attacks on DNN Accelerator in Cloud-FPGA". *arXiv* 2021. arXiv: 2105.09453 [cs].

[29] A. S. Rakin et al. "Deep-Dup: An Adversarial Weight Duplication Attack Framework to Crush Deep Neural Network in Multi-Tenant FPGA". *arXiv* 2020. arXiv: 2011.03006 [cs].

[30] P.-A. Moellic et al. *Security of Software Embedded Neural Network Models: State of the Art and Threat Modelling*. ANR PICTURE Project D11. 2021. URL: https://picture-anr.cea.fr.

[31] M. Isakov et al. "Survey of Attacks and Defenses on Edge-Deployed Neural Networks". *IEEE HPEC*. 2019, pp. 1–8. DOI: 10.1109/HPEC.2019.8916519.

[32] G. Kamath et al. "Pushing Analytics to the Edge". *IEEE GLOBECOM*. 2016, pp. 1–6. DOI: 10.1109/GLOCOM.2016.7842181.

[33] B. McMahan et al. "Federated learning: Collaborative machine learning without centralized training data". *Google AI Blog*. Vol. 3. 2017.

[34] Z. Zhou et al. "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing". *Proc. IEEE* 107.8 2019, pp. 1738–1762. DOI: 10.1109/JPROC.2019.2918951.

[35] N. Pitropakis et al. "A Taxonomy and Survey of Attacks against Machine Learning". *Comput. Sci. Rev* 34 2019, p. 100199. DOI: 10.1016/j.cosrev.2019.100199.

[36] C. Ma et al. "On Safeguarding Privacy and Security in the Framework of Federated Learning". *IEEE Network* 34.4 2020, pp. 242–248. DOI: 10.1109/MNET.001.1900506.

[37] K. Bonawitz et al. "Practical Secure Aggregation for Privacy-Preserving Machine Learning". *ACM CCS*. 2017, pp. 1175–1191. DOI: 10.1145/3133956.3133982.

[38] P. Blanchard et al. "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent". *NIPS*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.