An Improved Machine-Learning Model for the Identification and Classification of Memory-Based PUF Responses

Nico Mexis[1], Nikolaos Athanasios Anagnostopoulos[1], Tolga Arul[1,2], Elif Bilge Kavun[1], and Stefan Katzenbeisser[1]

[1]Faculty of Computer Science and Mathematics, University of Passau, Germany
[2]Department of Computer Science, Technical University of Darmstadt, Germany

35th Crypto Day, 25./26. May 2023

In recent years, the Internet of Things (IoT) has found wide adoption. The IoT is based on the direct interconnection and communication among computer systems, without the need for human intervention. It is comprised of devices that possess different capabilities, ranging from high-end servers and infrastructure devices to extremely resource-constrained devices, such as low-end sensors, data aggregators, and single-board computers and microprocessors. Thus, there is an inherent need for lightweight security solutions in the IoT.

In this context, Physical Unclonable Functions (PUFs) have been proposed as adequate security anchors, especially in the framework of the IoT. PUFs are physical objects, such as hardware memories, that intrinsically exhibit characteristics acquired by minor imperfections of their manufacturing process, which usually do not affect their normal operation, but are highly distinguishable and, thus, rather "unique" per PUF instance.[1] In general, the measurement of these characteristics leads to a rather stable output, which most often is in a binary form for computer-hardware-based PUFs, and is referred to as the *PUF response*.

However, since PUF responses do exhibit noise to some extent, they can usually not be directly used as security tokens, e.g., cryptographic keys. For this reason, fuzzy extractors are most often deployed to remove the relevant noise. Fuzzy extractors use an error-correction code and helper data that has to be generated in an enrollment phase, in order to correct the PUF responses by considering their noise as "errors". But since error-correction codes rely on complex procedures for encoding and decoding data, they can often not be directly deployed in low-end devices and a more lightweight solution is needed.

For this reason, Suragani, Nazarenko, Anagnostopoulos, Mexis & Kavun (2022) have instead suggested using Machine Learning (ML) based on a Convolutional Neural Network (CNN) to identify noisy (and even corrupted) PUF responses. Their architecture consists of five convolution layers and three fully connected dense layers, and provides a promising accuracy of up to 90% even with 30% corruption. In this work, we propose and examine an improved machine-learning model which is illustrated in Figure 1. This model consists only of two convolutional layers and a dense layer, which results in a much lower training time, while still providing comparable accuracy. The trained ML model is able to correctly categorise all the DRAM PUF responses from our training and validation datasets. Evaluation on a test dataset of a further 30 responses from each of the PUFs also yields a test accuracy of 100%, although these responses also exhibit noise. The trained model is also able to provide an accuracy of around 86% on up to 70%-corrupted PUF responses. However, a major drawback of using a conventional CNN is that it is not able to detect random PUF responses, which are then classified as one of the trained classes with very high confidence. This is also evident in our trained model.

crypto day matters 35 (2023)
Beiträge vom 35. Kryptotag, 25.-26. Juni 2023, Regensburg.
doi:10.18420/cdm-2023-35-03

*This work has been partially funded by DFG Projects 440182124 and 439892735, and by Project BA0100016 of the Programm INTERREG VI-A Bayern–Österreich 2021–2027, co-funded by the European Union. The authors would also like to thank Panagiotis Kessanidis for his work within the framework of the "Security Engineering Lab" course of the University of Passau during the academic semester of Winter 2022–2023.

[1] The term "Physical Unique Function" can more accurately describe a PUF.

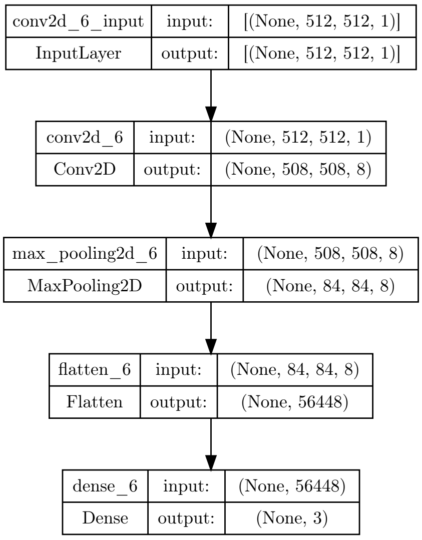

Figure 1: Our proposed CNN architecture, created in TensorFlow Keras.

References

Reshmi Suragani, Emiliia Nazarenko, Nikolaos Athanasios Anagnostopoulos, Nico Mexis & Elif Bilge Kavun (2022). Identification and Classification of Corrupted PUF Responses via Machine Learning. In *2022 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 137–140. IEEE. URL `https://doi.org/10.1109/host54066.2022.9839919`.