Investigation of Deep Learning Attacks on Nonlinear Silicon Photonic PUFs

Iskandar Atakhodjaev, Bryan T. Bosworth, Brian C. Grubel, Michael R. Kossey, Jesús Villalba, A. Brinton Cooper, Najim Dehak, Amy C. Foster, and Mark A. Foster

Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
mark.foster@jhu.edu

Abstract: We demonstrate that nonlinear silicon photonic Physical Unclonable Functions (PUFs) are resistant to adversarial deep learning attacks. We find that this resistance is rooted in the optical nonlinearity of the silicon photonic PUF token.

OCIS codes: (100.4998) Pattern recognition, optical security and encryption (190.4390) Nonlinear optics, integrated optics

Introduction

Physical Unclonable Functions (PUFs) are a modern day realization of a millennia old concept, the physical key. However, unlike conventional physical keys it is not feasible to copy a PUF. In modern authentication and cryptography, PUFs present a promising alternative to storing digital keys in physical memory as the security of these schemes relies on the presumption that these digital keys are not known to adversaries. Specifically, modern security schemes suffer from a number vulnerabilities such as the proper implementation and physical attacks (e.g. invasive, semi-invasive, and side-channel attacks), as well as software attacks and viruses. Such attacks often compromise the system's integrity by gaining access to the stored digital key [1]. As an alternative to this conventional approach, PUFs provide device specific, non-deterministic transfer functions that can be used for authentication and secret key storage without the drawbacks mentioned above. Instead of storing the digital key in physical memory, PUFs derive a secret key from a physical process that is sensitive to the random and unpredictable idiosyncrasies of a physical device [2]. As a result, it is impossible, even for the manufacturer, to clone or reproduce PUFs and their behavior. PUFs have been realized using a variety of technologies including electronic circuits and complex optical materials. Notably, optical PUFs based on volumetric scattering have been shown to provide orders of magnitude higher information content than their electronic counterparts [3, 4], however these optical scattering PUFs suffer from poor repeatability and problematic compatibility with electronic integration. Recently our group developed a novel photonic PUF based on ultrafast nonlinear optical interactions in chaotic silicon photonic micro-cavities [5,6]. These interactions produce a highly complex and unpredictable, yet deterministic, ultrafast response that can serve as a source of secret key material. Our nonlinear silicon photonic PUFs provide significant improvements over optical scattering PUFs in terms of repeatability, key generation rates, and ease of integration with CMOS electronics and telecommunications hardware, while also providing comparably large information capacity [5, 6].

In this paper, we investigate a set of eavesdropping attacks against this novel PUF. Specifically, we extract a subset of key material from a nonlinear silicon photonic PUF and use this material construct modeling attacks with the help of deep learning methods. The ultimate goal of this work is to understand whether an attacker armed with some subset of knowledge about the PUF's behavior can, within a limited time frame, design a mathematical model that is capable of correctly emulating the entire PUF behavior. According to previous studies, such machine learning (ML) attacks have been highly successful at emulating both electronic and optical PUFs [7]. For example, studies have shown that integrated optical scattering PUFs are vulnerable to ML attacks when a linear optical medium is employed [7]. Intriguingly, here we find that nonlinear silicon photonic PUFs are highly resistant to ML attacks and that this resistance is directly rooted in the nonlinear optical behavior of the devices.

Experiment and results

In the machine learning community, deep learning is widely acknowledged as the state-of-the-art method and outperforms other solutions in numerous fields such as computer vision, image and speech recognition, classification and machine translation [8]. A major advantage of the deep learning framework is that it is readily adapted to new problems and for this reason we employ deep learning to investigate ML attacks on our nonlinear silicon photonic PUF.

A PUF token is typically interrogated with the sequence of input signals called challenges and the output signals, called responses, are recorded to generate digital key material. A set of input-ouput signal pairs, which are known as challenge-response pairs (CRPs) are stored as a CRP database also known as a challenge-response library (CRL). The whole process of populating a CRL is referred as an enrollment phase.

To build the CRL, we employed the same experimental setup as demonstrated in [4, 5]. To generate the challenge signals, ultrafast 300-fs laser pulses with 90-MHz repetition rate are dispersively stretched with a spool of dispersion compensation fiber and are then encoded by a 128-bits random binary sequence generated by a pulse pattern generator operating at a 11.52-Gbit/s rate. These challenge pulses are then amplified in and erbium doped fiber amplifier, compressed in single-mode optical fiber, and coupled into the silicon photonic micro-cavity PUF (e.g. an example device is shown Fig. 1a). The sequence of response pulses are amplified and then measured by filtering with a spectral mask and recording the transmitted

pulse energy with an ADC. Finally, a post-processing algorithm derives a digital sequence from the analog power samples. At the whole enrollment phase, we collect 960,000 CRPs to form the CRL for testing deep learning attacks. We repeated this process at three different input optical power levels to study the impact of optical nonlinearity on the results of adversary attack. Additionally, we repeated CRL generation at each power level to determine the repeatability of the CRL.

In the attack scenario, we assume that the attacker (Eve) has stolen some part of the CRL and Eve's goal is to find an algorithm that can emulate the mapping of input challenges to output responses using the stolen part of database. To investigate this susceptibility, we construct a Deep Neural Network (DNN) and train it to learn the mapping function between challenges and responses, this is known as an end to end deep learning architecture. During the learning phase, we train the DNN with 70% of the total collected data and use the remaining 30% to test the performance of DNN at emulating the devices behavior. In Fig.1b, three learning curves showing DNN prediction error versus the size of the training set are presented for different input optical power level of the challenge laser pulses. We find that in all cases, the performance of the DNN plateaus after approximately 50,000 training samples. Furthermore, we find that the lower power behavior is better learned than higher power behavior indicating that the PUF's optical nonlinearity is providing resistance to the ML attack.

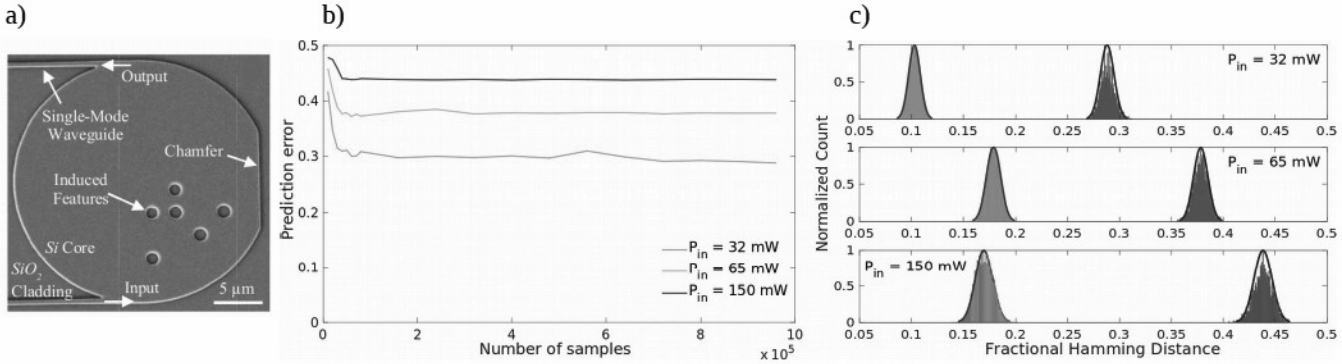

Figure 1. Deep Learning attack results. a) Scanning electronic microscope image of chaotic silicon micro-cavity b) Convergence of NN generalization errors with respect to amount of the dataset at average power levels 32mW, 65mW and 150mW b) Normalized FHD distributions and histograms calculated against CRL of legitimate PUF token at different power levels in the setup: "like" distribution (green) represent the FHD values between repetitions and the response sequence from CRL of legitimate PUF, ML "clone" distribution (blue) represent the FHD values between ML predicted response sequences and the response from CRL of legitimate PUF.

The performance of PUFs is measured using Fractional Hamming Distance (FHD) between response bit sequences and the CRL. FHD is used as a metric to quantify repeatability and differentiability of individual PUF tokens [2]. The histogram of FHDs between responses to repeated challenges to a PUF and the mean response of the same device is referred to as "like" distribution, whereas the histogram of FHDs between each of those repetitions and CRL of ML generated responses is referred to as ML "clone" distribution. Such histograms are shown in Fig. 1c at each of the three power levels. We calculated the mean and standard deviations for each histogram that account for uniqueness and repeatability of the token respectively. In ideal scenario, the mean of FHD distribution that accounts for comparison of one PUF device to different PUF should be centered around the value of 0.5. As Fig.1c shows, we observe good separability of genuine PUF device from the ML clone PUF even at low power level indicating resistance to the ML attacks at all power levels. Furthermore, at higher optical power levels, the ML "clone" distribution is further separated from the "like" distributions, consistent with the observation that optical nonlinearity in the system enhances its resistance to ML attacks.

References

[1] U. Rührmair, C. Hilgers, S. Urban, A. Weiershäuser, E. Dinter, B. Forster and C. Jirauschek. "Optical PUFs Reloaded." (2013).
[2] R. Maes, I. Verbauwhede, Physically unclonable functions: a study on the state of the art and future research directions, 2010
[3] R. Pappu, B. Recht, J. Taylor, N. Gershenfeld: Physical One-Way Functions, Science, vol. 297, 2002.
[4] R. Horstmeyer, B. Judkewitz, I. M. Vellekoop, S. Assawaworrarit & C. Yang, "Physical key-protected one-time pad", Scientific Reports 3, 3543 (2013)
[5] B. C. Grubel, B. T. Bosworth, M. R. Kossey, H. Sun, A. B. Cooper, M. A. Foster, A. C. Foster, "Silicon photonic physical unclonable function," Opt. Express 25, 12710-12721 (2017)
[6] B. C. Grubel, B. T. Bosworth, M. R. Kossey, A. B. Cooper, M. A. Foster, A. C. Foster, Information-Dense Nonlinear Photonic Physical Uncloneable Function, **arXiv:1711.02222**
[7] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, J. Schmidhuber: Modeling Attacks on Physical Unclonable Functions. ACM CCS, 2010.
[8] Y. LeCun., Y. Bengio & G. Hinton. Deep learning. Nature 521, 436–444 (2015)