Bus Electrocardiogram: Vulnerability of SoC-FPGA Internal AXI Buses to Electromagnetic Side-Channel Analysis

May Myat Thu, Maria Mendez Real, Maxime Pelcat, Philippe Besnier

Bus Electrocardiogram: Vulnerability of SoC-FPGA Internal AXI Buses to Electromagnetic Side-Channel Analysis

May Myat Thu[1], **Maria Méndez Real**[1], **Maxime Pelcat**[1], **Philippe Besnier**[1]
[1]Univ Rennes, INSA Rennes, Nantes Université, CNRS, IETR-UMR 6164, F-35000 Rennes, France
May-Myat.Thu@insa-rennes.fr

***Abstract*—This article demonstrates a confidential data vulnerability in integrated circuits, especially in System-on-Chip (SoC)-field programmable gate array (FPGA) circuits. We demonstrate that the electromagnetic (EM) emanations of internal Advanced eXtensible Interface (AXI) standard data buses are exploitable. Electromagnetic pulses recorded by a near-field probe in the vicinity of such data buses can be mapped to the underlying data flux, just as an electrocardiogram (ECG) would map the heart activity and blood circulation. This vulnerability is demonstrated by spying on the EM emanations caused by internal buses while MNIST handwritten digit images are sent through them. The transmitted MNIST data are found to be framed by preamble and post-amble signals enabling the detection of the transmitted data themselves. Data (images, in this paper) are then reconstituted by using a simple algorithm.**

***Index Terms*—Electromagnetic side-channel attacks, near-field, hardware security, on-chip communication buses, magnetic probe.**

I. Introduction

In recent years, a trend has developed towards automated acceleration of artificial intelligence (AI) algorithms on specialized hardware such as SoC-FPGAs. Nowadays, Convolutional neural networks (CNNs) are one of the most deployed AI technologies. They have shown state-of-the-art performance in several application domains such as computer vision, autonomous driving and medical imaging. Due to their extensive use and rising popularity, many CNNs are now used for processing a variety of sensitive and confidential data, and in particular, confidential images. However, as for any computation, during the inference process of CNNs on hardware accelerators, unintended signals are produced from hardware, which are called side-channels. These side-channels include variations in energy consumption, EM emissions, temperature, etc. Emanations in the near-field or the far-field can be observed and exploited to compromise secret information of the system, particularly the structure of the CNN or the data processed by the system. However, regarding the data themselves, they have to be transmitted within the chip through dedicated buses. To the best of our knowledge, this article is the first to highlight the vulnerability of internal communication buses on SoC-FPGAs, regardless of the process running inside the system. This paper demonstrates the recovery of SoC-FPGAs internal images, sent as inputs to the CNN implemented on the FPGA while exploiting near-field EM emanations of AXI bus.

The paper is organised as follows. The background and related work are briefly described in section II. The following sections are devoted to methodology (section III), signal of interest (section IV) and processing of trace signals (section V). After that, section VI presents the results achieved before the conclusion in section VII.

II. Background and Related Work

EM side-channel analysis is a technique for spying on computational activities and private data from unintentional electromagnetic emissions [1]. In computer security, sensitive assets should be protected from malicious access. We may distinguish different types of assets:

(a) confidential, non-encrypted data [2] [3] [4],
(b) confidential, encrypted data with secret key [5] and
(c) confidential details about implemented algorithms and models [6] [7] [8] [9].

Several research works based on EM side-channels have proven successfully regarding these different aspects.

A. Non-encrypted data

In 1985, EM eavesdropping of computer display was first demonstrated to the public by Van Eck using cathode-ray tubes (CRT) [2]. Authors of [3] proved that this eavesdropping methodology is not only limited to CRT but can also be applied to modern flat panel displays. Recently, authors of [4] created a novel EM-based approach that spies on a mobile device screen, without a direct line of sight, exposing any private data that may be displayed on there. It is safe to assume that this non-encrypted private data from the screens and cables can be compromised by EM side-channel analysis. In the frame of TEMPEST, non-encrypted sensitive data are called red data, and protection against EM eavesdropping requires zoning protection.

B. Encrypted Data with Secret Key

Sensitive assets can sometimes be protected by encryption. However, the secret key of an encryption algorithm can be deduced from unwanted EM traces, strongly jeopardizing system confidentiality and integrity. In [5], authors make an investigation of EM side-channels produced by the execution of Advanced Encryption Standard (AES) by three different implementations 1) by an ARM integer core, 2) by a proprietary

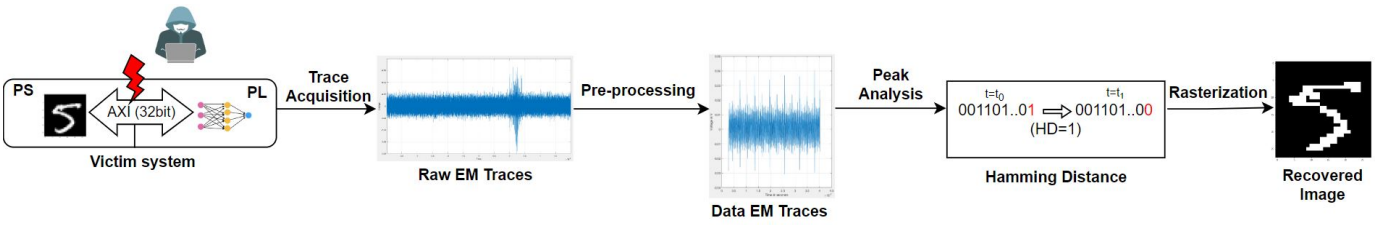

Fig. 1: Flow-chart of Bus ECG data retrieval from near-field probing.

co-processor and 3) by an ARM NEON Single-Instruction-Multiple-Data core. The key is recovered by analyzing the correlation of Hamming Weight to EM emanations. Once the key is obtained, it enables the decryption of sensitive data. In the frame of TEMPEST, encrypted data, protected by a secret key, are called black data and would not require the same zoning restrictions.

C. Details about implemented algorithms and models

1) Identification of Executed Operations: In addition to the asset leakage from screens and cryptographic cores, EM emanations can also expose other confidential information, such as the operations of the system running inside the hardware platform. Authors of [6] succeed in the identification of three different cryptographic algorithms executed on the two different hardware processor types (micro-controller and micro-processor) running on Internet of Things (IoT) devices. They are able to distinguish the executed algorithm between 3DES, AES-128 or AES-256, by processing EM signals. Then, with the help of a simple neural network based classifier, authors deduce the corresponding algorithm executed.

2) Recovery of neural network parameters: In the case of AI application systems, especially neural networks (NNs), EM side-channels can be associated with the retrieval of different non-encrypted information such as NN parameters. CSI-NN [7] fully reverse engineers the CNN parameters that should have been kept secret. Authors recover the activation function, pre-trained weights, number of hidden layers and neurons in each layer from EM side-channel analysis. DeepEM [8] demonstrates a new theft attack that uses EM emanations to retrieve information about network architecture and the designs inside SoC-based binarized neural network (BNN) accelerators. Similarly, [9] demonstrates how an attacker can do a full recovery of the secret BNN weights used in the network by capturing EM traces emitted from a target FPGA platform.

D. Current Work: Bus ECG

Unlike previous works mentioned above, our work develops a new method to retrieve non-encrypted input images sent to a BNN by taking advantage of the vulnerability of SoC-FPGA data buses to EM analysis. In our case scenario, the non-encrypted BNN input data, transferred through the internal bus, is a private asset that should be kept secret. This would correspond to red data in the frame of TEMPEST. However, contrary to the normal TEMPEST scenario, we focus on near-field attacks, where the embedded system platform executing the classification operations is physically accessible in the near-field to a potential attacker. This is the case for many civil applications such as camera surveillance in train stations and airports. Such systems are vulnerable as they are not covered by TEMPEST protection rules.

III. Methodology

In the case of a typical scenario such as video surveillance, smart cameras tend to embed more and more processing, implementing near-sensor computing, such as neural network inference, so as to reduce communication requirements between systems. For these systems to remain low-power, hardware accelerators such as FPGAs are commonly employed. Real-time images recorded by the sensor are sent to the FPGA where neural network inference is done. At the time images are transferred through internal buses, this creates an opportunity for capturing data using a near-field probe.

These electromagnetic emanations, converted to electric signal traces, are collected via an oscilloscope connected through a coaxial cable and a low-noise amplifier (LNA) to the near-field probe. It requires a physical access to the targeted system but remains a non-destructive approach.

The traces are then analyzed with only a partial knowledge of the victim system. In our experimental case, this includes the width of the bus and the general characteristics of the transmitted input image. However, in general, data buses are widely defined. The default width for standard data buses is 32-bit, whereas newer systems that have to deal with a large amounts of data now have 64-bit buses [10]. This reduces the possibilities of attacker's guess on the bus width. As for the characteristics of input image, most popular open datasets, MNIST [11], SVHN [12], etc., disclose their basic features to the public. Contrary to the majority of the works in the state-of-the-art, one may note that the proposed side-channel vulnerability in this work needs neither interacting with the architecture of CNN, nor triggering its inference. This significantly lowers the attacker constraints. Furthermore, no a-priori knowledge of inference inputs and outputs, weights, layers and neurons of the CNN is required.

The flow-chart of Bus ECG data retrieval is represented schematically in Fig. 1. The acquisition of signals is triggered by the bus data transmission protocol itself while the probe is placed in a zone of interest. The waveform of the trace allows

the acquisition system to trigger and find out positively and negatively polarized transient signals. This waveform is then correlated to the Hamming Distance (HD) between successive binary sequences on the bus. Such information is sufficient to, at least approximately, reconstitute the transmitted image on the bus. The principle will be implemented in detail in the context of an application to a specific SoC-FPGA target.

IV. Signal of Interest

A. Test Bench

The above principle is experimented on a Pynq Z1 target board which embeds a Zynq-7020 SoC-FPGA. Pynq Z1 is one of the few boards supported by the FINN, which is a popular framework by Xilinx for the generation and implementation of pre-trained CNNs on FPGA targets. A SoC-FPGA contains an ARM processor, also called Processing System (PS), and an FPGA called Programmable Logic (PL). Among the default FINN accelerators designed by Xilinx, we use BNN with LeNet-5 architecture which is trained on the MNIST hand-written character database. As mentioned in the section III, the attacker can already predict that the expected image is in gray scale formatted with 8 bits per pixel. BNNs are one of the sub-categories of NNs with their weight values quantized to either -1 or +1, hence, the name *binarized*. Pynq Z1 is a device used specifically for IoT applications and owns the advantage of not having a build-in heat-sink, which facilitates the movements of a near field probe on the surface of the chip. The AXI ports between the PS and the PL are configured with a 32-bit data width and have access directly from the PL to the dynamic random access memory (DRAM) memory controller of the PS. The test bench hardware setup is illustrated in Fig. 2.

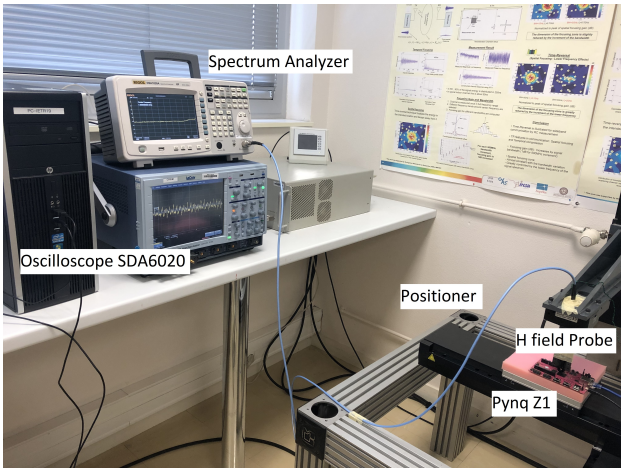

Fig. 2: Hardware setup of the experiment

B. Identification of Signals of Interest

First, it is useful to locate the areas of the chip where the leakage from AXI data transfer occurs. To do this, the surface of the SoC-FPGA is scanned by a customized magnetic field probe which is engraved on a dielectric substrate, as in Fig. 3.

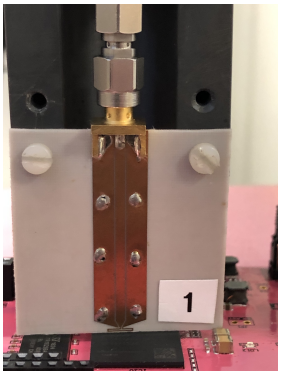

Fig. 3: Customized near field H-probe

According to the documentation of the Xilinx Zynq-7000 [13] the AXI buses ensuring the communication between PS and PL is clocked at the frequency of 525 MHz. A spectral analysis of the captured signals allows to formulate the hypothesis of localization of the bus. Fig. 4 presents a typical signature of the trace signals collected in the zone of interest. This chosen time window illustrates the behavioral difference between the absence and in the presence of images transmitted periodically. In the absence of transmission, the recorded signals are close to the noise floor.

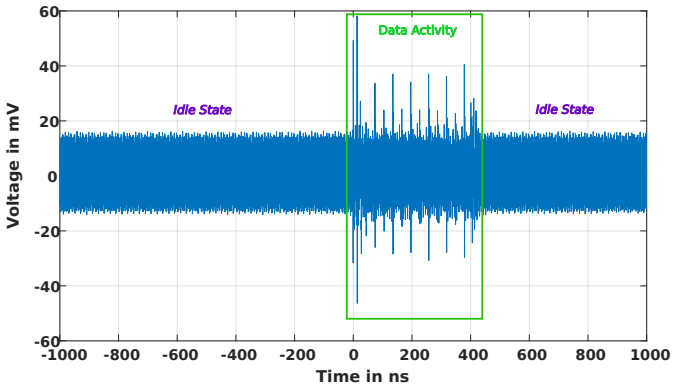

Fig. 4: Signals observed at the input of the oscilloscope connected to the near-field magnetic probe (through a cable and a LNA) in the zone of interest when the target is idle vs. active.

C. Cartography

The EM cartography is produced with a spatial resolution of 0.5 mm and the magnetic probe placed in vertical position Fig. 3. Taking Fig. 5 as reference, the field components H_y and H_x are measured for both horizontal or vertical orientation of the magnetic loop, respectively. The amplitude reported in the coloured map represents the average of the positive peaks of voltage difference in the trace signal when a checker print image (to induce maximum Hamming Distance (HD) between two bit sequences) is sent for inference. The measured signal is indeed composed of a regular train of short pulses. Fig. 6 shows the cartography with respect to H_y for the zone

indicated by the red box in Fig. 5. Field according to H_x was globally much weaker. This phenomenon is due to the fact that the highest levels of potential difference are located on a horizontal axis which, according to the manufacturer's datasheet, corresponds to the location of the AXI bus.

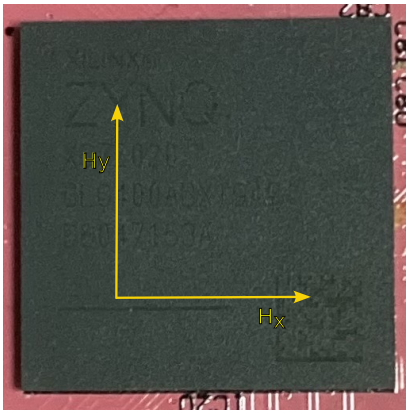

Fig. 5: Picture of the studied SoC. Yellow arrows refer to horizontal and vertical axis where field components H_y and H_x are measured.

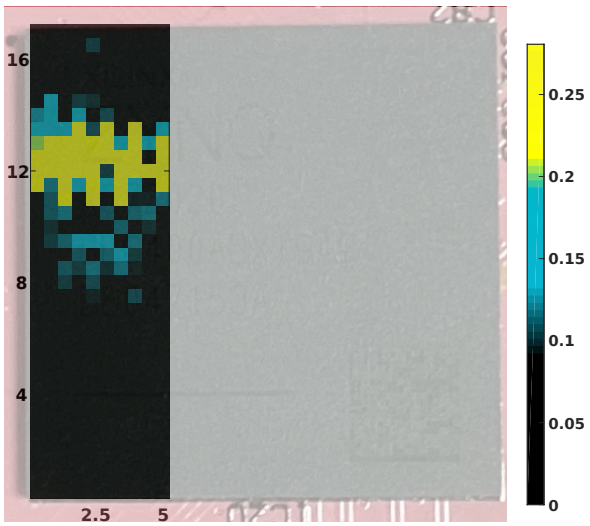

Fig. 6: Cartography of the voltage difference at the input of the oscilloscope using an horizontally oriented magnetic field probe in the region of interest. The probe is sensitive to the Hy component of the field, with y the vertical axis in Fig. 5.

V. Processing of Trace Signals

A. Acquisition and Processing of Traces

In these experiments, 200 traces per input image are captured. The measured signal comprises a *Pre-amble* and a *Post-amble* which occur before and after data transmission. The data traces capture is synchronized with the help of these two data-independent signals, as illustrated in Fig. 7. The *pre-amble* therefore allows a systematic triggering of the acquisition from the oscilloscope without an external synchronization signal as recourse. Therefore, it is safe to assume that these two signals define the start and end point of the data transfer from PS to PL before the inference operation inside PL. After the removal of *pre-amble* and *post-amble* from the collected traces, the remaining data dependant signals are averaged 200 times to remove Gaussian noise floor.

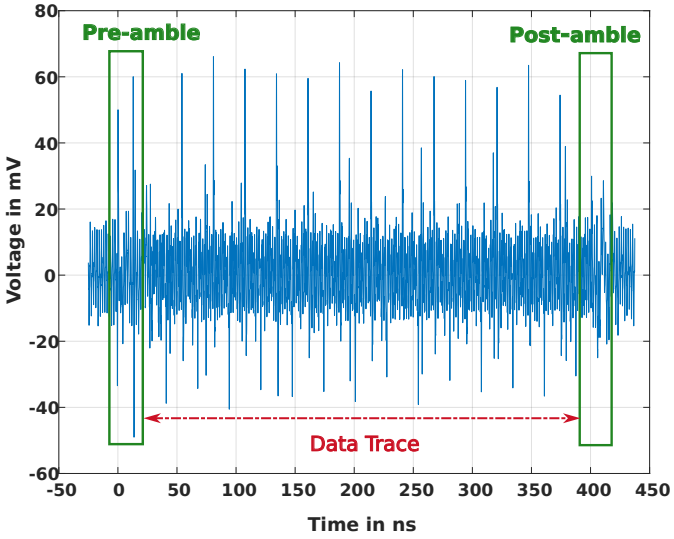

Fig. 7: Signal acquired by the oscilloscope at the output of the near field probe during the transfer of an image on AXI bus.

B. Interpretation of Traces and Partial Recovery of Data

The data width of AXI bus defines the value of *X*, which is the number of input pixels sent simultaneously on a single clock cycle. The HD of two groups of *X* pixels sent between two consecutive clock cycles represents the number of bit positions in which a transition appears (from $0 \rightarrow 1$ or from $1 \rightarrow 0$). HD therefore provides important information about the pixels transferred on the bus.

The peaks in the recorded trace represent the HD between two successive groups of X pixels. The higher the amplitude of the peak, the higher the HD value. Additionally, while placing the magnetic probe in a specific polarity, if the pixel bit values transmitted in cycle, *t+1* are greater than the bit values of the previous cycle *t*, it can be assumed that a positive peak will occur in the acquired trace signal. This positive peak represents the upward transition among the *X* bits sent. On the contrary, if the bit values of the pixels sent in the cycle *t+1* are lower than the values of cycle *t*, the result is a negative transient signal and it is interpreted as a downward transition. A positive bit transition (respectively negative) on the AXI bus corresponds to the the charge (respectively discharge) to the output gate capacitance and the parasitic capacitance of the bus lines. It has also been verified that by rotating the probe by 180°, the polarisation of the peaks are reversed which seems to confirm the detection of a magnetic flux through the loop. Note that signed HD is sensitive to the transition of bits between two successive group of pixels sent per clock cycle and not to the

values of the pixels directly. Thus, some information is lost in the interception process. Despite this loss, a great deal of information is disclosed and can be exploited.
The detection of positive and negative peaks requires the selection of two appropriate thresholds. After observation of a large number of traces, these thresholds have been set empirically at 40% of the maximum value of the positive peaks and 25% of the maximum amplitude of the negative peaks. These choices ensure a good discrimination between data signal and noise and decrease the risk of false detection. These thresholds are to be defined once before the attack, and do not depend on the application on the targeted PL.

C. Image Reconstruction Algorithm

We propose here a method to reconstruct the binarized format of the original image. Due to the fact that the transitions on the least significant bits cannot be distinguished from the transitions on the most significant bits, the binarization becomes the best modality as it allows the extraction of decent amount of information with a limited complexity. The reconstruction of the image is done in sequence, pixel by pixel until the image matrix is obtained entirely. This matrix is first initialized to a certain pixel value, for instance, the background color (*e.g.*, black). The dimension of the matrix is assumed to be known beforehand as the standard image size of MNIST dataset is public knowledge. Let X be the number of pixels transmitted simultaneously (X = 4 for a 32 bit bus). The processed trace signal is taken and its amplitude (voltage) values at each time instant are scanned. Once it has reached the positive threshold, the detection of a positive peak is triggered. As a result, the value of the pixel in that specific position, as well as the next $X-1$ pixels in the scanning order of the image frame are assigned to the white color. This white pixel value is kept for all the following pixels until the detection of a new negative threshold is triggered. In the same way, when a negative threshold is reached, the corresponding pixels group in the matrix and its following $X-1$ pixels are set to the black value until the next positive peak is detected. This procedure is repeated until the end of the EM trace. Therefore, all pixels of the image are assigned to either black or white color.

VI. Results

The experimental results presented in this article are obtained from 8-bit pixel images of 28×28 in size. Since the AXI ports are configured to the data width of 32 bits, a group of 4 pixels is sent through the bus per clock cycle. The detection of peaks in the trace signal therefore represents the HD of two consecutive groups of 4 pixels sent between two clock cycles. To start the reconstitution of input images, a 2D image matrix of 28×28 (same size than the original image to be recovered) is first initialized with all values assigned by an assumed background color, in this case, black.

The tests have been performed using a set of images from the MNIST database containing handwritten numbers. This database serves as a training base for our BNN, implemented on FPGA. The examples of results obtained with Bus ECG are presented in Fig. 8. The first row of numbers corresponds to the original input images transmitted to the FPGA by the AXI bus. The second line of numbers corresponds to the recovery of the images from the traces acquired using the algorithm mentioned in Section V-C. Each image is transmitted a certain number of times in order to increase the signal-to-noise ratio by averaging. This is a common practice to evaluate the information carried by a side-channel signal [5]. In our case, each image is transmitted 200 times, which is realistic in a video camera surveillance as it will correspond to around 8 seconds for a video camera with standard frame rate of 24 frames per second (FPS). In these conditions, the restitution of the pixel information carried by the images is almost perfect.

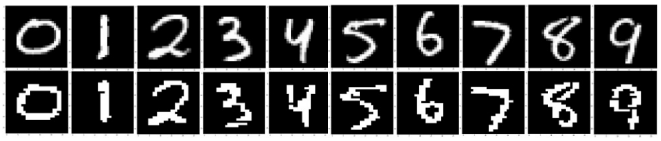

Fig. 8: First line: original MNIST images fed to the BNN. Second line: recovered images from Bus ECG with 200 averaged traces

VII. Conclusion

This work has shown that the internal AXI bus of a SoC FPGA can become the target of a malicious exploitation using an electromagnetic side-channel attack. The EM emanations of the chip can be extracted and exploited using a near field magnetic probe located at the surface of the component. A methodology has been used to demonstrate that internal secret images can be reconstructed from trace signals. Since the exploited vulnerability exploits an AXI communication bus, this attack is likely not to be limited to SoC-FPGAs and may be applicable to other hardware platforms as long as they contain similar internal buses for data transmission. Even if the network exchanged data are encrypted, once the data is inside the chip, it usually circulates as non-encrypted data on internal buses. One may also note that the magnetic field sensor can be made very small, making it possible to hide a surface sensor inside a package. Countermeasures to prevent this *Bus ECG* side-channel analysis are still to be deployed, and may impact the cost and energy consumption of the system. Future work will be devoted to the implementation of the attack on different types of data and different SoC platforms in order to prove that the threat exists in a large set of systems and applications.

References

[1] Mark Scanlon Asanka P. Sayakkara, Nhien-An Le-Khac. A survey of electromagnetic side-channel attacks and discussion on their case-progressing potential for digital forensics. *CoRR*, 2019.

[2] Wim van Eck. Electromagnetic radiation from video display units: An eavesdropping risk? *Computers and Security*, 1985.

[3] Markus Kuhn. Electromagnetic eavesdropping risks of flat-panel displays. *LNCS*, 2004.

[4] Zhuoran Liu, Niels Samwel, Leo Weissbart, Zhengyu Zhao, Dirk Lauret, Lejla Batina, and Martha A. Larson. Screen gleaning: A screen reading TEMPEST attack on mobile devices exploiting an electromagnetic side channel. *CoRR*, 2020.

[5] J. Longo, E. De Mulder, D. Page, and M. Tunstall. Soc it to em: Electromagnetic side-channel attacks on a complex system-on-chip. In *Conference on Cryptographic Hardware and Embedded Systems (TCHES)*, 2015.

[6] Mark Scanlon Asanka P. Sayakkara, Nhien-An Le-Khac. Leveraging electromagnetic side-channel analysis for the investigation of iot devices. *CoRR*, 2019.

[7] Lejla Batina, Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. Csi nn: Reverse engineering of neural network architectures through electromagnetic side channel. In *Proceedings of the 28th USENIX Conference on Security Symposium*, 2019.

[8] Honggang Yu, Haocheng Ma, Kaichen Yang, Yiqiang Zhao, and Yier Jin. Deepem: Deep neural networks model recovery through em side-channel information leakage. In *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2020.

[9] Ville Yli-Mäyry, Akira Ito, Naofumi Homma, Shivam Bhasin, and Dirmanto Jap. Extraction of binarized neural network architecture and secret parameters using side-channel information. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[10] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2017.

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012.

[12] Ian Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks, 2013.

[13] Xilinx. Zynq-7000 soc data sheet: Overview, 2018. https://docs.xilinx.com/v/u/en-US/ds190-Zynq-7000-Overview.