

# BoMaNet: Boolean Masking of an Entire Neural Network

Anuj Dubey  
aanujdu@ncsu.edu  
North Carolina State University  
Raleigh, North Carolina

Rosario Cammarota  
rosario.cammarota@intel.com  
Intel Labs  
San Diego, United States

Aydin Aysu  
aaysu@ncsu.edu  
North Carolina State University  
Raleigh, North Carolina

## ABSTRACT

Recent work on stealing machine learning (ML) models from inference engines with physical side-channel attacks warrant an urgent need for effective side-channel defenses. This work proposes the first *fully-masked* neural network inference engine design.

Masking uses secure multi-party computation to split the secrets into random shares and to decorrelate the statistical relation of secret-dependent computations to side-channels (e.g., the power draw). In this work, we construct secure hardware primitives to mask *all* the linear and non-linear operations in a neural network. We address the challenge of masking integer addition by converting each addition into a sequence of XOR and AND gates and by augmenting Trichina's secure Boolean masking style. We improve the traditional Trichina's AND gates by adding pipelining elements for better glitch-resistance and we architect the whole design to sustain a throughput of 1 masked addition per cycle.

We implement the proposed secure inference engine on a Xilinx Spartan-6 (XC6SLX75) FPGA. The results show that masking incurs an overhead of 3.5% in latency and 5.9× in area. Finally, we demonstrate the security of the masked design with 2M traces.

## KEYWORDS

Masking, neural networks, side-channel attacks, model stealing

## 1 INTRODUCTION

Physical side-channel attacks pose a major threat to the security of cryptographic devices. Attacks like the Differential Power Analysis (DPA) [27] can extract secret keys by exploiting the inherent correlation between the secret-key-dependent data being processed and the CMOS power consumption [9]. DPA has been shown to be effective against many cryptographic implementations in the last two decades [7, 16, 50]. Until recently, these attacks were confined to cryptographic schemes. But lately, the Machine Learning (ML) applications are shown to be vulnerable to physical side-channel attacks [13, 21, 76], where an adversary aims to reverse engineer the ML model. Indeed, these models are lucrative targets as they are costly to develop and hence become valuable IPs for the companies [65]. Knowledge about model parameters also eases fooling the model using adversarial learning, which is a serious problem if the model performs a critical task like fraud/spam detection [48].

Unfortunately, most of the existing work on the physical side-channel analysis of ML accelerators has focused only on attacks, not defenses. To date, there are three publications focusing specifically on the power/EM side-channel leakage of ML models. The first two discuss some countermeasures but limit the discussion of the mitigation to a qualitative level [21, 76]. The third one implements a hybrid of masking and hiding based countermeasures and this combination of empirical and provably secure mitigation is secure in practice to the best capabilities or today's attackers [13].

Masking uses secret-sharing to split the secrets into random shares and to decorrelate the statistical relation of secret-dependent computations to side-channels. Although similar work on cryptographic hardware has been fully masked [3], the earlier work on neural network hardware was *partially masked* to reduce costs [13]. Such solutions may work well for a regular IP where reasonable security at a low-cost is sufficient. However, *full masking* is a better alternative for IPs deployed in critical applications (like defense) requiring stronger defenses against side-channel attacks.

In this work, we propose the design of the *first fully-masked* neural network accelerator resistant against power-based side-channel attacks. We construct novel masked primitives for the various linear and non-linear operations in the network using gate-level *Boolean* masking and masked look-up tables (LUT). We analyze neural network-specific computations like weighted summations in fully-connected layers and build specialized masked adders and multiplexers to perform those operations in a secure way. We also design a novel hardware that finds the greatest integer out of a set of integers in a masked fashion, needed for the output layer.

We target an area-optimized Binarized Neural Network (BNN) in our work because of their preference for edge-based neural network inference [57, 71]. We optimize the hardware design to reduce the impact of masking on the performance. Specifically, we build an innovative adder-accumulator architecture that provides a throughput of one addition per cycle even with a multi-cycle masked adder with feedback. We reduce the effects of glitches [30] in the masking elements by adding registers at every stage, to synchronize the arrival of signals. We build the masked design in a modular fashion starting from small Trichina's AND gates blocks to the large 20-bit masked full adder. We have pipelined the full design to maintain a high throughput.

Finally, we implement both the baseline unmasked and the proposed first-order secure masked neural network design on an FPGA. We use the standard TVLA methodology to evaluate the first-order security of the design and demonstrate no leakage up to 2M traces [2]. Masking increases the latency by only 3.5% and the area by 5.9×. Our goal in this paper is to provide the first fully-masked design with certain optimizations and a practical evaluation of security. We also discuss potential further optimizations and extensions of masking for hardware design and security refinements.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ICCAD '20, November 2–5, 2020, Virtual Event, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8026-3/20/11...\$15.00

<https://doi.org/10.1145/3400302.3415649>

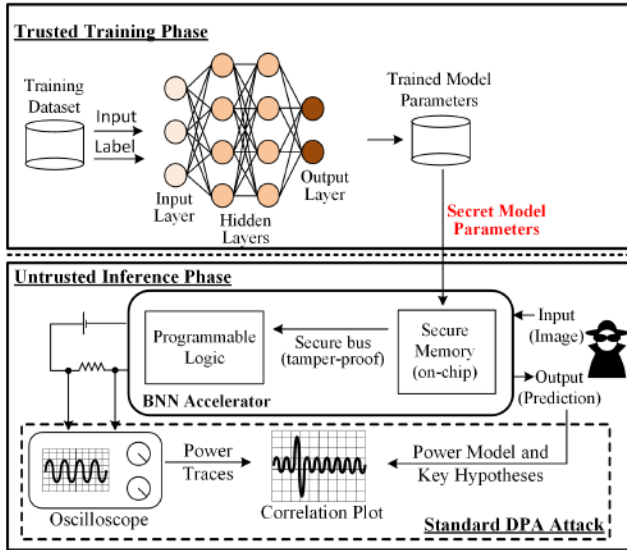


Figure 1: Standard DPA threat model applied to ML model stealing, where the trained neural network is deployed to an edge device running in an untrusted environment.

## 2 THREAT MODEL

We adopt the standard DPA threat model in which an adversary has direct physical access to the target device running inference [13, 21, 45], or can obtain power measurements remotely when the device executes neural network computations [78]. The adversary can control the inputs and observe the corresponding outputs from the device as in chosen-plaintext attacks.

Figure 1 shows our threat model where the training phase is trusted but the trained model is deployed to an inference engine operating in an untrusted environment. The adversary is after the trained model parameters (e.g., weights and biases)—input data privacy is out of scope [72].

We assume that the trained ML model is stored in a protected memory and the standard techniques are used to securely transfer it (i.e., bus snooping attacks are out of scope) [28]. The adversary, therefore, has gray-box access to the device, i.e., it knows all the design details up to the level of each individual logic gate but does not know the trained ML model. We restrict the secret variables to just the parameters and not the hyperparameters such as the number of neurons, following earlier work [13, 44, 69]. In fact, an adversary will still not be able to clone the model with just the hyperparameters if it does not possess the required compute power or training dataset. This is analogous to the scenario in cryptography where an adversary, even after knowing the implementation of a cipher, cannot break it without the correct key.

We target a hardware implementation of the neural network, not software. The design fully fits on the FPGA. Therefore, it does not involve any off-chip memory access and executes with constant-flow in constant time. These attributes make the design resilient to any type of digital (memory, timing, access-pattern, etc.) side-channel attack. However, the physical side-channels like power and EM emanations still exist; we address the power-based side-channel leakages in our work. Other implementation attacks on neural networks such as the fault attacks [4, 5] are out of scope.

## 3 BACKGROUND AND RELATED WORK

This section presents related work on the privacy of ML applications, the current state of side-channel defenses, preliminaries on BNNs, and our BNN hardware design.

### 3.1 ML Model Extraction

Recent developments in the field of ML point to several motivating scenarios that demand asset confidentiality. Firstly, training is a computationally-intensive process and hence requires the model provider to invest money on high-performance compute resources (eg. a GPU cluster). The model provider might also need to invest money to buy a labeled dataset for training or label an unstructured dataset. Therefore, knowledge about either the parameters or hyperparameters can provide an unfair business advantage to the user of the model, which is why the ML model should be private.

Theoretical model extraction analyzes the query-response pair obtained by repeatedly querying an unknown ML model to steal the parameters [23, 25, 55, 58]. This type of attack is similar to the class of theoretical cryptanalysis in the cryptography literature. Digital side-channels, by contrast, exploit the leakage of secret-data dependent *intermediate computations* like access-patterns or timing in the neural network computations to steal the parameters [11, 14, 34, 75], which can usually be mitigated by making the secret computations constant-flow and constant-time. Physical side-channels target the leak in the physical properties like CMOS power-draw or electromagnetic emanations that will still exist in a constant-flow/constant-time algorithm’s implementation [13, 21, 40, 72, 73]. Mitigating physical side-channels are thus harder than digital side-channels in hardware accelerator design and has been extensively studied in the cryptography community.

### 3.2 Side-Channel Defenses

The researchers have proposed numerous countermeasures against DPA. These countermeasures can be broadly classified as either *hiding-based* or *masking-based*. The former aims to make the power-consumption constant throughout the computation by using power-balancing techniques [53, 67, 77]. The latter splits the sensitive variable into multiple statistically independent shares to ensure that the power consumption is independent of the sensitive variable throughout the computation [3, 24, 35, 37, 56, 70].

The security provided by hiding-based schemes hinges upon the precision of the back-end design tools to create a near-perfect power-equalized circuit by balancing the load capacitances across the leakage prone paths. This is not a trivial task and prior literature shows how a well-balanced dual-rail based defense is still vulnerable to localized EM attacks [41]. By contrast, masking transforms the algorithm itself to work in a secure way by never evaluating the secret variables directly, keeping the security mostly independent of back-end design and making it a favorable choice over hiding.

### 3.3 Neural Network Classifiers

Neural network algorithms learn how to perform a certain task. In the learning phase, the user sends a set of inputs and expected outputs to the machine (a.k.a., training), which helps it to approximate (or learn) the function mapping the input-output pairs. The learned function can then be used by the machine to generate outputs for unknown inputs (a.k.a., inference).



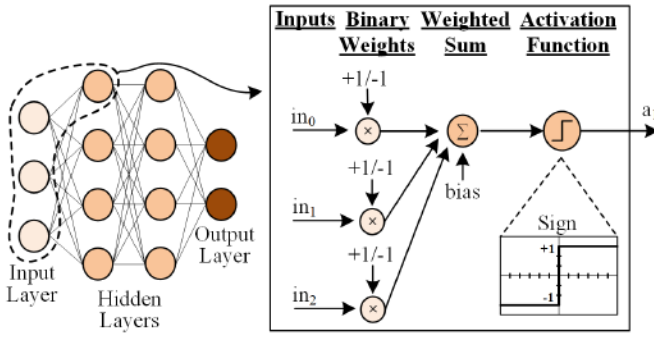


Figure 2: A typical Binarized Neural Network where the neuron performs weighted summations on binarized weights, and the activation function is a sign function.

A neural network consists of units called neurons (or nodes) and these neurons are usually grouped into layers. The neurons in each layer can be connected to the neurons in the previous and next layers. Each connection has a weight associated with it, which is computed during the training. The neurons work in a feed-forward fashion passing information from one layer to the next.

The weights and biases can be initialized to be random values or a carefully chosen set before training [29]. These weights and biases are the *critical parameters* that our countermeasure aims to protect. During training, a set of inputs along with the corresponding labels are fed to the network. The network computes the error between the actual outputs and the labels and tunes the weights and biases to reduce it, converging to a state where the accuracy is acceptable.

### 3.4 Binarized Neural Networks

The weights and biases of a neural network are typically floating-point numbers. However, high area, storage costs, and power demands of floating-point hardware do not fare well with the requirements of the resource-constrained edge devices. Binarized Neural Networks (BNNs) [8], with their low hardware cost and power needs fit very well in this use-case while providing a reasonable accuracy. BNNs restrict the weights and activations to binary values (+1 and -1), which can easily be represented in hardware by a single bit. This reduces the storage costs for the weights from floating-point values to binary values. The XNOR-POPCOUNT operation implemented using XNOR gates replaces the large floating-point multipliers resulting in a huge area and performance gain [57].

Figure 2 depicts the neuron computation in a fully-connected BNN. The neuron in the first hidden layer multiplies the input values with their respective binarized weights. The generated products are added to the bias, and the result is fed to the activation function, which is a sign function that binarizes the non-negative and negative inputs to +1 to -1, respectively. Hence, the activations in the subsequent layer are also binarized.

### 3.5 Our Baseline BNN Hardware Design

We consider a BNN having an input layer of 784 nodes, 3 hidden layers of 1010 nodes each, and an output layer of 10 nodes. The 784 input nodes denote the 784 pixel values in the 28×28 grayscale images of the Modified National Institute of Standards and Technology (MNIST) database and 10 output nodes represent the 10 output classes of the handwritten numerical digit.

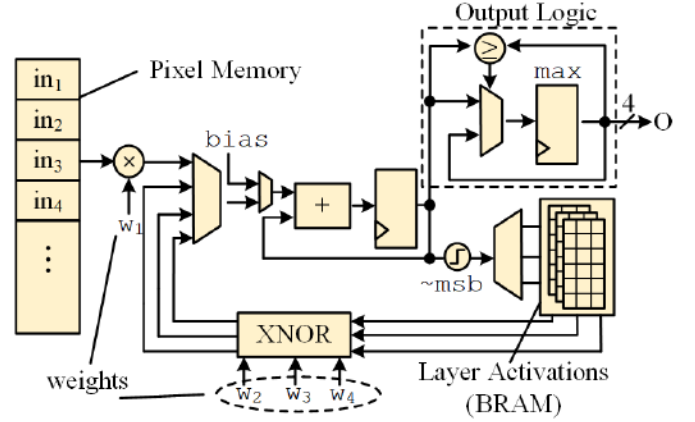


Figure 3: A sequentialized hardware design of the baseline BNN using a single adder.

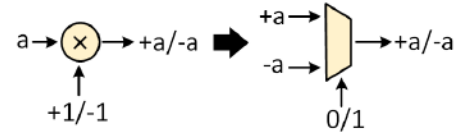


Figure 4: Multiplier expressed as a multiplexer in BNNs.

**3.5.1 Weighted Summations.** We choose to use a single adder in the design and sequentialize all the additions in the algorithm to reduce the area costs. Figure 3 shows our baseline BNN design. The computation starts from the input layer pixel values stored in the Pixel Memory. For each node of the first hidden layer, the hardware multiplies 784 input pixel values one by one and accumulates the sum of these products. The final summation is added with the bias reusing the adder with a multiplexed input and fed to the activation function. The hardware uses XNOR and POPCOUNT<sup>1</sup> operations to perform weighted summations in the hidden layers. The final layer summations are sent to the output logic.

In the input layer computations, the hardware multiplies an 8-bit unsigned input pixel value with its corresponding weight. The weight values are binarized to either 0 or 1 (representing a -1 or +1, respectively). Figure 4 shows the realization of this multiplication with a multiplexer that takes in the pixel value ( $a$ ) and its 2's complement ( $-a$ ) as the data inputs and weight ( $\pm 1$ ) as the select line. The 8-bit unsigned pixel value, when multiplied by  $\pm 1$ , needs to be sign-extended to 9-bits, resulting in a 9-bit multiplexer.

**3.5.2 Activation Function.** The activation function binarizes the non-negative and negative inputs to +1 and -1 respectively for each node of the hidden layer. In hardware, this is implemented using a simple NOT gate that takes the MSB of the summations as its input.

**3.5.3 Output Layer.** The summations in the output layer represent the confidence score of each output class for the provided image. Therefore, the final classification result is the class having the maximum confidence score. Figure 3 shows the hardware for computing the classification result. As the adder generates output layer summations, they are sent to the output logic block that performs a rolling update of the max register ( $max$ ) if the newly received sum

<sup>1</sup>The POPCOUNT operation also involves an additional step of subtracting the number of nodes (1010) from the final sum, which can be done as part of bias addition step.

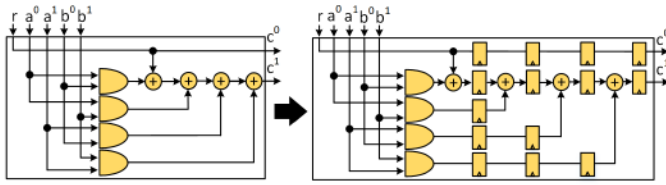


Figure 5: Trichina’s AND Gate implementation: glitch-prone (left) and glitch-resistant (right). Flip-flops synchronize arrival of signals at XOR gates’ inputs to mitigate glitches.

is greater than the previously computed max. In parallel, the hardware also stores the index of the current max node. The index stored after the final update is sent out as the final output of the neural network. The hardware takes 2.8M cycles to finish one inference.

#### 4 FULLY MASKING THE NEURAL NETWORK

This section discusses the hardware design and implementation of all components in the masked neural network. Prior work on masking of neural networks shows that arithmetic masking alone cannot mask integer addition due to a leakage in the sign-bit [13]. Hence, we apply gate-level *Boolean* masking to perform integer addition in a secure fashion. We express the entire computation of the neural network as a sequence of AND and XOR operations and apply gate-level masking on the resulting expression. XORs, being linear, do not require any additional masking, and AND gates are replaced with secure, Trichina style AND gates [70]. Furthermore, we design specialized circuits for BNN’s unique components like Masked Multiplexer and Masked Output Layer.

We first explain the notations in equations and figures. Any variable without a subscript or superscript represents an N-bit number. We use the subscript to refer to a single bit of the N-bit number. For example,  $a_7$  refers to the 8<sup>th</sup> bit of  $a$ . The superscript in masking refers to the different secret shares of a variable. To refer to a particular share of a particular bit of an N-bit number, we use both the subscript and the superscript. For example,  $a_4^1$  refers to the second Boolean share of the 5<sup>th</sup> bit of  $a$ . If a variable only has the superscript (say  $i$ ), we are referring to its full N-bit  $i^{th}$  Boolean share; N can also be equal to 1, in which case  $a$  is simply a bit.  $r$  (or  $r_i$ ) denotes a fresh, random bit.

##### 4.1 Why Trichina’s Masking Style?

Among the closely related masking styles [59], we chose to implement Trichina’s method due to its simplicity and implementation efficiency. Figure 5 (left) shows the basic structure and functionality of the Trichina’s gate, which implements a 2-bit, masked, AND operation of  $c = a \cdot b$ . Each input ( $a$  and  $b$ ) is split into two shares ( $a^0$  and  $a^1$  s.t.  $a = a^0 \oplus a^1$ , and  $b^0$  and  $b^1$  s.t.  $b = b^0 \oplus b^1$ ). These shares are sequentially processed with a chain of AND gates initiated with a fresh random bit ( $r$ ). A single AND operation thus uses 3 random bits. The technique ensures that output is the Boolean masked output of the original AND function, i.e.,  $c = c^0 \oplus c^1$ , while all the intermediate computations are randomized.

Unfortunately, the straightforward adoption of Trichina’s AND gate can lead to information leakage due to glitches [31]. For instance, in Figure 5 (left) if the products  $a_0 \cdot b_0$  and  $a_0 \cdot b_1$  reach the input of second XOR gate before random mask  $r$  reaches the input of first XOR gate, the output at the XOR gate will evaluate (glitch) to  $(a_0 \cdot b_0) \oplus (a_0 \cdot b_1) = a_0 \cdot (b_0 \oplus b_1)$  temporarily, which

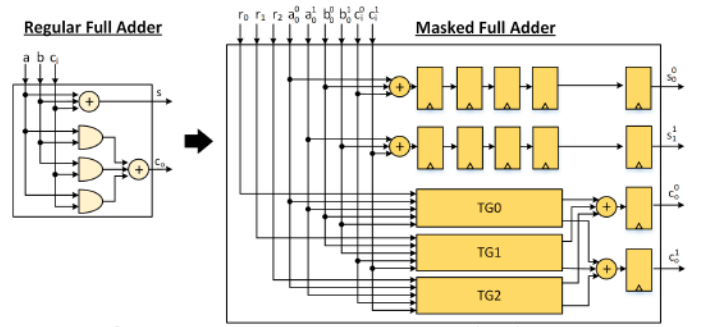


Figure 6: Regular operation of a Full Adder (left) and its gate-level masking using Trichina AND Gates (right).

leads to secret value  $b$  being unmasked. Therefore, we opted for an extension of the Trichina’s AND gate by adding flip-flops to synchronise the arrival of inputs at the XOR gates (see Figure 5 right). The only XOR gate not having a flip-flop at its input is the leftmost XOR gate in the path of  $c_1$ , which is not a problem because a glitching output at this gate does not combine two shares of the same variable. Similar techniques have been used in past [22]. Masking styles like the Threshold gates [20, 30, 49] may be considered for even stronger security guarantees, but they will add further area-performance-randomness overhead.

##### 4.2 Masked Adder

We adopt the ripple-carry style of implementation for the adder. It is formed using N 1-bit full adders where the carry-out from each adder is the carry-in for the next adder in the chain, starting from LSB. Therefore, ripple-carry configuration eases parameterization and modular design of the Boolean masked adders.

**4.2.1 Design of a Masked Full Adder.** A 1-bit full adder takes as input two operands and a carry-in and outputs the sum and the carry, which are a function of the two operands and the carry-in. If the input operand bits are denoted by  $a$  and  $b$  and carry-in bit by  $c$ , then the Boolean equation of the sum  $S$  and the carry  $C$  can be described as follows:

$$S = a \oplus b \oplus c \quad (1)$$

$$C = a \cdot b \oplus b \cdot c \oplus c \cdot a \quad (2)$$

Figure 6 shows the regular, 1-bit full adder (on the left), and the resulting masked adder with Trichina’s AND gates (on the right). In the rest of the subsection, we will discuss the derivation of the masked full adder equations.

First step is to split the secret variables ( $a$ ,  $b$  and  $c$ ) into Boolean shares. The hardware samples a fresh, random mask from a uniform distribution and performs XOR with the original variable. If we represent the random masks as  $a^0$ ,  $b^0$  and  $c^0$ , then the masked values  $a^1$ ,  $b^1$  and  $c^1$  can be generated as follows:

$$a^1 = a \oplus a^0, \quad b^1 = b \oplus b^0, \quad c^1 = c \oplus c^0 \quad (3)$$

A masking scheme always works on the two shares independently without combining them at any point in the operation because that will reconstruct the secret and create a side-channel leak.

The function of sum-generation is linear, making it easy to directly and independently compute the Boolean shares of  $S$ :

$$S = S^0 \oplus S^1$$

where,

$$S^0 = a^0 \oplus b^0 \oplus c^0, \quad S^1 = a^1 \oplus b^1 \oplus c^1$$



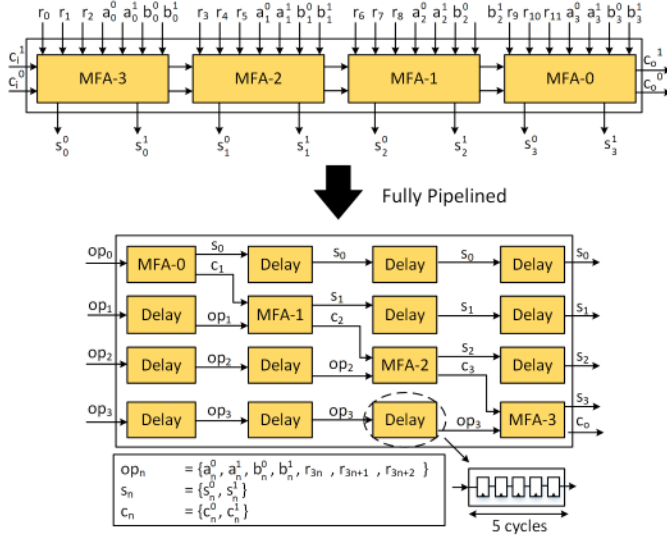


Figure 7: Modular design of a masked 4-bit adder using Masked Full Adders (top) and its pipelined version (bottom).

Unlike the sum-generation, carry-generation is a non-linear operation due to the presence of an AND operator. Hence, the hardware cannot directly and independently compute the Boolean shares  $C^0$  and  $C^1$  of  $C$ . We use the Trichina's construction explained in subsection 4.1 to mask carry-generation.

The hardware uses three Trichina's AND gates to mask the three AND operations in equation (2) using three random masks. This generates two Boolean shares from each Trichina AND operation. At this point, the expression is linear again, and therefore, the hardware can redistribute the terms, similar to the masking of sum operation. In the following equations, we use  $TG(x, y, r)$  to represent the product  $x \cdot y$  implemented via Trichina's AND Gate as illustrated in the following equation:

$$x \cdot y = TG(x, y, r) = m^0 \oplus m^1$$

where  $m^0$  and  $m^1$  are the two Boolean shares of the product. Replacing each AND operation in equation (2) with TG, we can write

$$TG(a, b, r_0) = d^0 \oplus d^1 \quad (4)$$

$$TG(b, c, r_1) = e^0 \oplus e^1 \quad (5)$$

$$TG(c, a, r_2) = f^0 \oplus f^1 \quad (6)$$

where  $d^0, d^1, e^0, e^1, f^0$ , and  $f^1$  are the output shares from each Trichina Gate. From equations (2), (4), (5), and (6) we get

$$carryout = TG(a, b, r_0) \oplus TG(b, c, r_1) \oplus TG(c, a, r_2)$$

Replacing the TGs from equation (4), (5), and (6) and rearranging the terms, we get

$$carryout = (d^0 \oplus e^0 \oplus f^0) \oplus (d^1 \oplus e^1 \oplus f^1)$$

which can also be written as a combination of two Boolean shares  $C^0$  and  $C^1$  where

$$C^0 = d^0 \oplus e^0 \oplus f^0, C^1 = d^1 \oplus e^1 \oplus f^1$$

Therefore, we create a masked full adder that takes in the Boolean shares of the two bits to be added along with a carry-in and gives out the Boolean shares of the sum and carry-out.

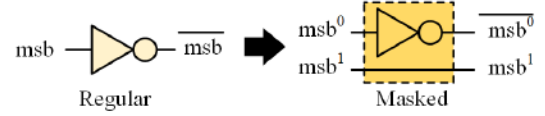


Figure 8: The unmasked activation function (left) is a NOT gate, and the masked version (right) receives two Boolean shares of MSB from masked adder and inverts one of them.



Figure 9: Masking a regular multiplexer using a masked LUT taking a fresh random mask  $r_i$ .

**4.2.2 The Modular Design of Pipelined N-bit Full Adder.** The masked full adders can be chained together to create an N-bit masked adder that can add two masked N-bit numbers. Figure 7 (top) shows how to construct a 4-bit masked adder as an example. We pipeline the N-bit adder to yield a throughput of one by adding registers between the full-adders corresponding to each bit (see Figure 7 (bottom)).

### 4.3 Masking of Activation Function

The baseline hardware implements the activation function as an inverter as discussed in 3.5.2. In the masked version, the MSB output from the adder is a pair of Boolean shares. To perform NOT operation in a masked way, the hardware simply inverts one of the Boolean shares as Figure 8 shows.

### 4.4 Masked Multiplexer

A 9-bit multiplexer is internally a set of parallel nine 1-bit multiplexers. We implement the masked 1-bit multiplexer using a 4-input 2-output masked look-up table (LUT). Figure 9 shows the masked LUT that takes in the original inputs ( $a, -a$ ) and an additional fresh random mask ( $r_i$ ) as inputs and outputs the random mask ( $r_o$ ) which is simply the bypassed  $r_i$  and the correct output XORed with the random mask. We assume that each LUT operation is atomic. Since the output functions are 4-input, 2-output, they can be mapped onto the same LUT of the target FPGA [74]. Lesser number of inputs also obviate the need for precautions like building a carefully balanced tree of smaller input LUTs [26]. Advanced masking constructions can be used to implement this function for a stronger security guarantee. As suggested in another work [26], masked look-ups can also be implemented using ROMs if the target is an ASIC, since ROMs are immutable and small in size. Thus, the (Boolean) masked output from the LUTs ensures that the secret intermediate-variable (multiplexed input pixel) always remains masked.

### 4.5 Masking the Output Layer

The hardware stores the 10 output layer summations in the form of Boolean shares. To determine the classification result, it needs to find the maximum value among the 10 masked output nodes. Specifically, it needs to *compare* two signed values expressed as Boolean shares. We transform the problem of *masked comparison* to *masked subtraction*.

Figure 10 shows the hardware design of the masked output layer. The hardware subtracts each output node value from the current maximum and swaps the current maximum (old max shares) with

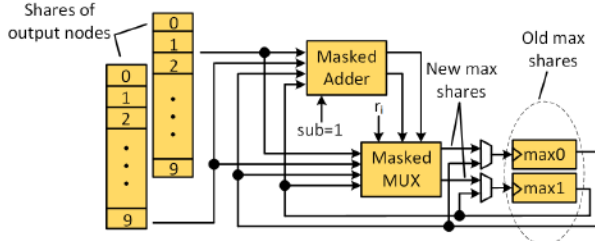


Figure 10: Masking of the Output Layer that uses a masked subtractor and a masked multiplexer to find the node with the maximum confidence score among the 10 output nodes.

the node value (new max shares) if the MSB is 1 using a masked multiplexer. An MSB of 1 signifies that the difference is negative and hence the new sum is greater than the latest max. Instead of building a new masked subtractor, we reuse the existing masked adder to also function as a subtractor through a *sub* flag, which is set while computing max. In parallel, the hardware uses one more masked multiplexer-based update-circuit to update the Boolean shares of the index corresponding to the current max node (not shown in the Figure). This is to prevent known-ciphertext attacks, ciphertext being the classification result in our case. Finally, the Masked Output Logic computes the classification result in the form of (Boolean) shares of the node's index having the maximum confidence score.

Subtraction is essentially adding a number with the 2's complement of another number. 2's complement is computed by taking bitwise 1's complement and adding 1 to it. A bitwise 1's complement is implemented as an XOR operation with 1 and the addition of 1 is implemented by setting the initial carry-in to 1. The additional XOR, being a linear operator, changes nothing with respect to the masking of the new adder-subtractor circuit.

#### 4.6 Scheduling of Operations

We optimize the scheduling in such a way that the hardware maintains a throughput of 1 addition per cycle. The latency of the masked 20-bit adder is 100 cycles. Therefore, the result from the adder will only be available after 101 cycles<sup>2</sup> from the time it samples the inputs. The hardware cannot feed the next input in the sequence until the previous sum is available because of the data dependency between the accumulated sum and the next accumulated sum. This incurs a stall for 101 cycles leading to a total of  $784 * 101 = 79184$  cycles for each node computation. That is a  $784\times$  performance drop over the unmasked implementation with a regular adder.

We solve the problem by finding useful work for the adder that is independent of the summation in-flight, during the stalls. We observe that computing the weighted summation of one node is completely independent of the next node's computation. The hardware utilizes this independence to improve the throughput by starting the next node computation while the result for the first node arrives. Similarly, all the nodes up till 101 can be computed upon concurrently using the same adder and achieve the exact same throughput as the baseline design. This comes at the expense of additional registers (see Figure 11<sup>3</sup>) for storing 101 summations<sup>4</sup> plus some control logic but a throughput gain of  $784\times$  (or  $1010\times$

<sup>2</sup>Need an additional cycle for the accumulator register as well.

<sup>3</sup>The register file also has a demultiplexing and multiplexing logic to update and consume the correct accumulated sum in sequence, which is not shown for simplicity.

<sup>4</sup>This is why we use 1010 neurons, which is a multiple of 101, in the hidden layers.

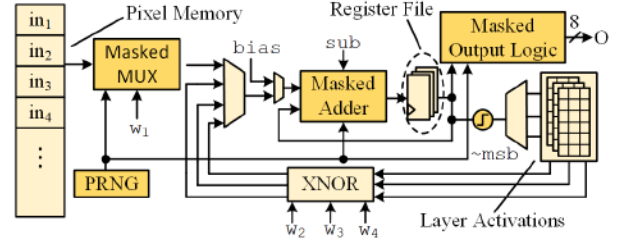


Figure 11: Hardware Design of the Fully Masked Neural Network. The components related to masking are shown in dark yellow. The register file helps in throughput optimization.

in hidden layers) is worthwhile. The optimization only works if the number of next-layer nodes is greater than, and a multiple of 101. This restricts optimizing the output layer (of 10 nodes) and contributes to the 3.5% increase in the latency of the masked design.

## 5 RESULTS

In this section, we describe the hardware setup used to implement the neural network and capture power measurements, the leakage assessment methodology that we follow to evaluate the security of the proposed design, and the hardware implementation results.

### 5.1 Hardware Setup

We use Verilog for the front-end design and Xilinx ISE 14.7 for the back-end. We use the DONT\_TOUCH synthesis attribute and disable the tool options like LUT combining, register reordering, etc. to prevent any optimization in the masked components.

Our side-channel evaluation platform is the SAKURA-G FPGA board [19]. It hosts Xilinx Spartan-6 (XC6SLX75-2CSG484C) as the main FPGA that executes the neural network inference. An on-board amplifier amplifies the voltage drop across a  $1\Omega$  shunt resistor on the power supply line. We use Picoscope 3206D [66] as the oscilloscope to capture the measurements from the board. The design is clocked at 24MHz and the oscilloscope sampling frequency is 125MHz. A higher sampling frequency leads to the challenges that we discuss in Section 6.2. However, to ensure a sound evaluation, we perform first and second-order t-tests on a smaller unit of the design at a much higher precision: the design frequency is 1.5MHz and the sampling frequency is 500MHz.

We use Riscure's Inspector SCA [61] software to communicate with the board and initiate a capture on the oscilloscope. By default, the Inspector software does not support SAKURA-G board communication. Hence, we develop our own custom software modules to automate the setup. The modules implement the communication protocol and perform the register reads/writes on the FPGA to start the inference and capture the result.

### 5.2 Leakage Evaluation

We perform the leakage assessment of the proposed design using the non-specific fixed vs random t-tests, which is a common and generic way of assessing the side-channel vulnerability in a given implementation [2]. A t-score lying within the threshold range of  $\pm 4.5$  implies that the power traces do not leak any information about the data being processed, with up to 99.99% confidence. The measurement and evaluation is quite involved and we refer the reader to Section 6.2 for further details. We demonstrate the security



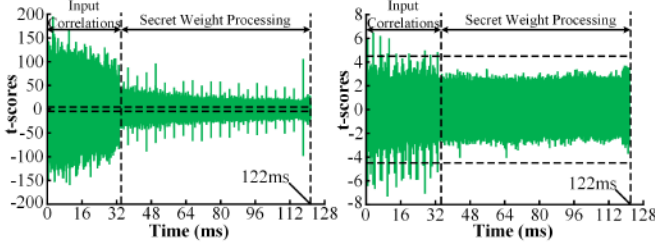


Figure 12: TVLA results of the unmasked (left) and masked (right) implementation. The results clearly show that the unmasked design is insecure, whereas the masked design is secure with 99.99% confidence (t-scores always below  $\pm 4.5$ ).

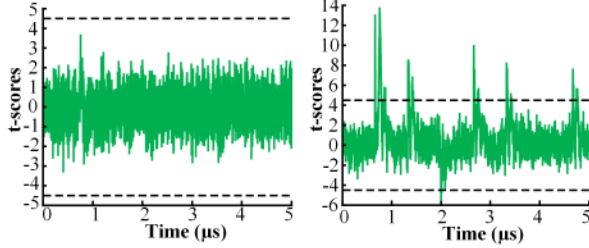


Figure 13: First-order (left) and second-order (right) t-tests on Trichina's AND gate at a low design frequency of 1.5MHz and sampling frequency of 500MHz.

up to 2M traces, which is much greater than the first-order security of the currently best-known defense that leaks at 40k traces [13].

Pseudo Random Number Generators (PRNG) produce the fresh, random masks required for masking. We choose TRIVIUM [10] as the PRNG, which is a hardware implementation friendly PRNG specified as an International Standard under ISO/IEC 29192-3, but any cryptographically-secure PRNG can be employed. TRIVIUM generates  $2^{64}$  bits of output from an 80-bit key; hence, the PRNG has to be re-seeded before the output space is exhausted.

**5.2.1 First-order tests.** We first perform the first-order t-test on the design with PRNGs disabled, which is equivalent to an unmasked (baseline/unprotected) design. Figure 12 (left) shows the result for this experiment where we clearly observe significant leakages since the t-scores are greater than the threshold of  $\pm 4.5$  for the entire execution. Then, we perform the same test, but with PRNGs switched on this time, which is equivalent to a masked design. Figure 12 (right) shows the results for this case, where we observe that the t-scores never cross the threshold of  $\pm 4.5$  except the initial phase.

The initial phase leakages are due to the input correlations during input layer computations. The hardware loads the input pixel after every 101 cycles and feeds it to the masked multiplexer. The secret variable is the *weight*, which is never exposed because the masked multiplexer randomises the output using a fresh, random mask.

**5.2.2 High Precision First and Second-order tests.** We performed univariate second-order t-test on the fully masked design [64], but 1M traces were not sufficient to reveal the leakages. Due to the extremely lengthy measurement and evaluation times it was infeasible to continue the test for more traces. Therefore, we perform first and second-order evaluation on the isolated synchronized Trichina's AND gate, which is one of the main building blocks of the design. We reduce the design frequency to 1.5MHz to increase the accuracy of the measurement and prevent any clock cycle aliasing. The

Table 1: Area (LUT/FF/BRAM) and Latency (in cycles) Comparison of the Unmasked and Masked Implementations.

Metric	Unmasked	Masked	Change
Area	1833/1125/163	9833/7624/163	$5.3\times / 6.8\times / 1\times$
Latency	$2.85 \times 10^6$	$2.94 \times 10^6$	$1.04\times$

Table 2: Block-level Area Distribution of the Unmasked and Masked Implementations (LUT/FF/BRAM).

Design Blocks	Unmasked	Masked	Fraction(%)
Adder	10/0/0	954/1050/0	12/16/-
PRNGs	0/0/0	1125/1314/0	14/20/-
Output Layer	7/16/0	32/22/0	0.3/0.09/-
Throughput Optimization	0/20/0	5337/4040/0	66/62/-
ROMs	411/1009/159	672/1009/159	4/-/-
RWMs	0/0/4	0/0/4	-
Misc	486/108/0	1233/2605/0	9/38/-

"-" denotes no change in the area of the unmasked and masked design.

SNR for a single gate was not sufficient to see leakage even at 10M traces, hence we amplify the SNR by instantiating 32 independent instances of the Trichina's AND gate in the design, driven by the same inputs. We present the results for this experiment in Figure 13 that shows no leakage in the first-order t-test but significant leakages in the second-order t-tests for 500k traces. Thus, the successful second-order t-test validates the correctness of our measurement setup and the first-order masking implementation.

### 5.3 Masking Overheads

Table 1 shows that the impact of masking on the latency is  $1.04\times$ , and on the number of LUTs and FFs is  $5.3\times$  and  $6.7\times$  respectively. We also summarize the area contribution from each design component in Table 2. The fourth column indicates what fraction of the total increase in area (i.e., 8000 LUTs and 6499 FFs) does each component contribute. Most of the area increase is due to the throughput optimization logic—the register file accumulator logic described in subsection 4.6. The masked adder contributes 12% and 16% to the overall increase in the LUTs and FFs respectively. The increase due to the output layer logic is minimal. ROMs refer to the read-only memories storing the weights and bias values where the increase is minimal<sup>5</sup>. RWMs refer to the read-write memories storing the layer activations, which also do not show any increase as the masked version stores two bits (the Boolean shares) instead of one for the activations accommodated in the same BRAM tile.

We compare the area-delay product (ADP) of our proposed design, BoMaNet, to MaskedNet [13], where area is defined as the sum of the number of LUTs and FFs, and delay is the latency in number of cycles. The ADP of our design is  $5 \times 10^{10}$  whereas the ADP of MaskedNET is  $6.4 \times 10^8$ , which is approximately  $80\times$  lower. This is expected since MaskedNet was designed for cost-effectiveness using partial masking, but on BoMaNet every operation is masked at the gate-level to improve side-channel security. Similar overheads were observed in previous works on Boolean masking of AES [36].

<sup>5</sup>The slight increase in the number of LUTs is because one of the memories is implemented using LUTs that might redistribute even for the same memory size.

## 6 DISCUSSIONS

### 6.1 Proof-of-Concept vs. Optimizations

Our solutions utilize simple yet effective techniques to mask an inference engine. But certainly, there is scope for improvement both in terms of the hardware design and the security countermeasures. In this section, we discuss possible optimizations/extensions of our work and alternate approaches taken in the field of privacy for ML.

**6.1.1 Design Optimizations.** The ripple-carry adder (RCA) used in this work can be replaced with advanced adder architectures like carry-lookahead [12], carry-skip [47], or Kogge-Stone [46] (KG). These architectures commonly possess an additional logic block to pre-compute the *generate* and *propagate* bits. Therefore, additional randomness will be needed to mask the non-linear *generate* expression. All these adders have more combinational logic than the ripple-carry adder, which may make it harder to avoid glitches. To that end, prior work on TI-based secure versions of RCA and KG can be extended [63]. Another potential optimization is the use of other masking styles like DoM [38] or manual techniques [15] to reduce the area and randomness overheads.

**6.1.2 Limitations.** We reduce glitch-related vulnerabilities using registers at each stage, which is a low-cost, practical solution. Other works have proposed stronger countermeasures, at the cost of higher performance and area overheads [31, 38]. The quest for stronger and more efficient countermeasures is never-ending; masking of AES is still being explored, even 20 years after the initial work [24], due to the advent of more efficient or secure masking schemes [32] and more potent attacks [52, 70].

Our solution is first-order secure but there is scope for construction of higher-order masked versions. However, higher-order security is a delicate task; Moos *et al.* recently showed that a straightforward extension of masking schemes to higher-order suffers from both local and compositional flaws [33] and masking extensions were proposed in another recent work [18]. This is the first work on fully-masked neural networks and we foresee follow ups as we have experienced in the cryptographic research of AES masking, even after 20 years of intensive study.

### 6.2 Measurement Challenges

We faced some unique challenges that are not generally seen with the symmetric-key cryptographic evaluations. Inference becomes a lengthy operation, especially for an area-optimized design—the inference latency of our design is roughly 3 million cycles. For a design frequency of 24MHz, the execution time translates to 122ms per inference. If the oscilloscope samples at 125MHz the number of sample points to be captured per power trace is equal to 15 million. This significantly slows down the capturing of power traces. In our case, capturing 2 million power traces took one week, which means capturing 100 million traces as AES evaluation [32] will take roughly a year. Performing TVLA on such large traces (28TB, in our case) also takes a significant amount of time: it took 3 days to get one t-score plot during our evaluations on a high-end PC<sup>6</sup>. One possibility to avoid this problem is looking at a small subset of representative traces of the computation [6], but, we instead conduct a comprehensive evaluation of our design.

<sup>6</sup>Intel Core i9-9900K, 64GB RAM.

### 6.3 Theoretical vs Side-Channel Attacks

Theoretical model extraction by training a proxy model on a synthetically generated dataset using the predictions from the unknown victim model is an active area of research [23, 25]. These attacks mostly assume a black-box access to the model and successfully extract the model parameters after a certain number of queries. This number ranges typically in the order of  $2^{20}$  [25]. By contrast, physical side-channel attacks only require a few thousand queries to successfully steal all the parameters [13]. This is partly due to fact that physical side-channel attacks can extract information about intermediate computations even in a black-box setting. Physical side-channel attacks also do not require the generation of the synthetic dataset, unlike most theoretical attacks.

### 6.4 Orthogonal ML Defences

There has been some work on defending the ML models against stealing of inputs and parameters using other techniques like Homomorphic Encryption (HE) and Secure Multi-Party Computation (SMPC) [17, 51, 60], Watermarking [1, 62], and Trusted Execution Engines (TEE) [39, 43, 68]. The survey by Isakov *et al.* and the NIST draft is a good reference for a more exhaustive list [42, 54]. Purely HE-based or HE+SMPC solutions initially incurred a large overhead (30000× in CryptoNet [17] and 500× in SecureML [51]), but follow-up work made them more efficient (13× in XONN[60]). We propose masking, which is an extension of SMPC on hardware and we believe that it is a promising direction for ML side-channel defenses as it has been on cryptographic applications. Watermarking techniques are punitive methods that cannot prevent physical side-channel attacks. TEEs are subject to ever-evolving microarchitectural attacks and typically are not available in edge/IoT nodes.

## 7 CONCLUSIONS AND FUTURE OUTLOOK

Physical side-channel analysis of neural networks is a new, promising direction in hardware security where the attacks are rapidly evolving compared to defenses. We proposed the first fully-masked neural network, demonstrated the security with up to 2M traces, and quantified the overheads of a potential countermeasure. We addressed the key challenge of masking integer addition [13] through Boolean masking. We furthermore presented ideas on how to mask the unique linear and non-linear computations of a fully-connected neural network that do not exist in cryptographic applications.

The large variety in neural network architectures in terms of the quantization-level and the types of layer operations (e.g., Convolution, Maxpool, Softmax), and activation functions (e.g., ReLU, Sigmoid, Tanh) presents a large design space for neural network side-channel defenses. This paper focused on BNNs as they are a good starting point. Our ideas serve as a benchmark to analyze the vulnerabilities that exist in neural network computations and to construct more robust and efficient countermeasures.

## 8 ACKNOWLEDGEMENTS

This project is supported in part by NSF under Grants No. 1943245 and SRC GRC Task 2908.001. We also thank Dr. Vikram Suresh and Dr. Sohrab Aftabjahani for their valuable feedback on the design.



## REFERENCES

- [1] Yossi Adi et al. 2018. Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring. In *USENIX Security '18*.
- [2] George Becker et al. 2013. Test Vector Leakage Assessment (TVLA) methodology in practice. In *International Cryptographic Module Conference*, Vol. 1001.
- [3] Johannes Blömer et al. 2005. Provably Secure Masking of AES. In *Selected Areas in Cryptography*.
- [4] Jakub Breier et al. 2018. Practical Fault Attack on Deep Neural Networks. In *2018 ACM SIGSAC Conference on Computer and Communications Security*.
- [5] Jakub Breier et al. 2020. SNIFF: Reverse Engineering of Neural Networks with Fault Attacks. *arXiv preprint arXiv:2002.11021* (2020).
- [6] Mathieu Carbone et al. 2019. Deep Learning to Evaluate Secure RSA Implementations. *TCHES* 2019 (2019).
- [7] Cong Chen et al. 2015. Differential Power Analysis of a McEliece Cryptosystem. In *International Conference on Applied Cryptography and Network Security*. Springer.
- [8] Matthieu Courbariaux et al. 2016. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. (2016).
- [9] Debayan Das et al. 2019. STELLAR: A Generic EM Side-Channel Attack Protection through Ground-Up Root-cause Analysis. In *HOST '19*.
- [10] Christophe De Cannière et al. 2008. Trivium. In *New Stream Cipher Designs*.
- [11] Gaofeng Dong et al. 2019. Floating-Point Multiplication Timing Attack on Deep Neural Network. In *IEEE International Conference on Smart Internet of Things*.
- [12] Robert W Doran. 1988. Variants of an improved carry look-ahead adder. *IEEE Trans. Comput.* 37, 9 (1988), 1110–1113.
- [13] Anuj Dubey et al. 2019. MaskedNet: A Pathway for Secure Inference against Power Side-Channel Attacks. *arXiv preprint arXiv:1910.13063* (2019).
- [14] Vasisht Duddu et al. 2018. Stealing Neural Networks via Timing Side Channels. *arXiv preprint arXiv:1812.11720* (2018).
- [15] Amir Moradi et al. 2012. Glitch-Free Implementation of Masking in Modern FPGAs. In *HOST '12*.
- [16] Aesun Park et al. 2018. Side-Channel Attacks on Post-Quantum Signature Schemes based on Multivariate Quadratic Equations. *TCHES* 2018, 3 (2018).
- [17] Gilad-Bachrach et al. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *ICML*. 201–210.
- [18] Gaëtan Cassiers et al. 2020. Hardware Private Circuits: From Trivial Composition to Full Verification. ePrint, Report 2020/185. (2020).
- [19] H. Guntur et al. 2014. Side-channel Attack User Reference Architecture board SAKURA-G. In *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*.
- [20] Kris Tiri et al. 2007. Changing the Odds Against Masked Logic. In *SAC 2006: 13th Annual International Workshop on Selected Areas in Cryptography*, Vol. 4356.
- [21] Lejla Batina et al. 2019. CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel. In *USENIX Security '19*.
- [22] Monjur Alam et al. 2009. Effect of glitches against masked AES S-box implementation and countermeasure. *IET Information Security* (2009).
- [23] Matthew Jagielski et al. 2019. High Accuracy and High Fidelity Extraction of Neural Networks. *arXiv:cs.LG/1909.01838*
- [24] Mehdi-Laurent Akkar et al. 2001. An Implementation of DES and AES, Secure against Some Attacks. In *CHES 2001*, Vol. 2162. Springer, Heidelberg, Germany.
- [25] Nicholas Carlini et al. 2020. Cryptanalytic Extraction of Neural Network Models. *arXiv:cs.LG/2003.04884*
- [26] Oscar Reparaz et al. 2015. A Masked Ring-LWE Implementation. In *CHES 2015*.
- [27] Paul C. Kocher et al. 1999. Differential Power Analysis. In *Advances in Cryptology – CRYPTO'99*. Springer, Heidelberg, Germany.
- [28] Rosario Cammarota et al. 2018. Machine Learning IP Protection. In *ICCAD '18*.
- [29] Sinno Jialin Pan et al. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* (2010).
- [30] Stefan Mangard et al. 2005. Successfully Attacking Masked AES Hardware Implementations. In *CHES 2005*.
- [31] Svetla Nikova et al. 2006. Threshold Implementations Against Side-Channel Attacks and Glitches. In *ICICS '06*.
- [32] Thomas De Cnudde et al. 2016. Masking AES with d+1 Shares in Hardware. *IACR ePrint*, 2016/631.
- [33] Thorben Moos et al. 2019. Glitch-Resistant Masking Revisited. *IACR TCHES* 2019, 2 (2019).
- [34] King Hu et al. 2019. Neural Network Model Extraction Attacks in Edge Devices by Hearing Architectural Hints. *arXiv:cs.CR/1903.03916*
- [35] Yuval Ishai et al. 2003. Private Circuits: Securing Hardware against Probing Attacks. In *Advances in Cryptology – CRYPTO 2003*, Vol. 2729.
- [36] Yuan Yao et al. 2018. Fault-Assisted Side-Channel Analysis of Masked Implementations. In *HOST '18*.
- [37] Jovan D Golić et al. 2003. Multiplicative Masking and Power Analysis of AES. In *CHES 2002*, Vol. 2523. Springer, Heidelberg, Germany.
- [38] Hannes Groß et al. 2016. Domain-Oriented Masking: Compact Masked Hardware Implementations with Arbitrary Protection Order. *IACR ePrint* (2016).
- [39] Lucjan Hanzlik et al. 2018. MLCapsule: Guarded Offline Deployment of Machine Learning as a Service. *arXiv preprint arXiv:1808.00590* (2018).
- [40] Weizhe Hua et al. 2018. Reverse Engineering Convolutional Neural Networks through Side-Channel Information Leaks. In *DAC '18*.
- [41] Vincent Immler et al. 2017. Your Rails Cannot Hide From Localized EM: How Dual-Rail Logic Fails on FPGAs. In *CHES 2017*.
- [42] Mihailo Isakov et al. 2019. Survey of Attacks and Defenses on Edge-Deployed Neural Networks. In *HPEC '19*.
- [43] Mihailo Isakov et al. 2018. Preventing Neural Network Model Exfiltration in Machine Learning Hardware Accelerators. In *AsianHOST '18*.
- [44] Mika Juuti et al. 2019. PRADA: Protecting Against DNN Model Stealing Attacks. In *EuroS&P '19*.
- [45] Paul Kocher et al. 2011. Introduction to Differential Power Analysis. *Journal of Cryptographic Engineering* 1 (2011).
- [46] Peter M Kogge et al. 1973. A Parallel Algorithm for the Efficient Solution of a General Class of Recurrence Equations. *IEEE Trans. Comput.* 100, 8 (1973).
- [47] M Lehman et al. 1961. Skip Techniques for High-Speed Carry-Propagation in Binary Arithmetic Units. *IRE Transactions on Electronic Computers* 4 (1961).
- [48] Daniel Lowd et al. 2005. Adversarial Learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*.
- [49] Stefan Mangard et al. 2005. Side-Channel Leakage of Masked CMOS Gates. In *Topics in Cryptology – CT-RSA 2005*.
- [50] Stefan Mangard et al. 2008. *Power analysis attacks: Revealing the secrets of smart cards*. Vol. 31. Springer Science & Business Media.
- [51] Payman Mohassel et al. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*.
- [52] Thorben Moos et al. 2017. Static Power Side-Channel Analysis of a Threshold Implementation Prototype Chip. In *DATE '17*.
- [53] Maxime Nassar et al. 2010. BCDL: A High Speed Balanced DPL for FPGA with Global Precharge and No Early Evaluation. In *DATE '10*.
- [54] NIST. 2019. *A Taxonomy and Terminology of Adversarial Machine Learning*. <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>
- [55] Seong Joon Oh et al. 2019. Towards Reverse-Engineering Black-Box Neural Networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*.
- [56] Elisabeth Oswald et al. 2005. A Side-Channel Analysis Resistant Description of the AES S-Box. In *Fast Software Encryption*.
- [57] Mohammad Rastegari et al. 2016. XNOR-Net: Imagenet Classification using Binary Convolutional Neural Networks. In *ECCV '16*.
- [58] Robert Nikolai Reith et al. 2019. Efficiently Stealing Your Machine Learning Models. In *18th ACM Workshop on Privacy in the Electronic Society*.
- [59] Oscar Reparaz et al. 2016. Additively Homomorphic Ring-LWE Masking. In *International Workshop on Post-Quantum Cryptography*.
- [60] M Sadegh Riazi et al. 2019. XONN: XNOR-based Oblivious Deep Neural Network Inference. In *USENIX Security '19*.
- [61] Riscure. 2019. *Riscure Inspector*. Retrieved May 7, 2020 from [https://www.riscure.com/uploads/2017/08/inspector\\_brochure.pdf](https://www.riscure.com/uploads/2017/08/inspector_brochure.pdf)
- [62] Bitá Darvish Rouhani et al. 2018. Deepsigns: A Generic Watermarking Framework for IP Protection of Deep Learning Models. *arXiv preprint arXiv:1804.00750* (2018).
- [63] Tobias Schneider et al. 2015. Arithmetic Addition over Boolean Masking. In *Applied Cryptography and Network Security*.
- [64] Tobias Schneider et al. 2016. Leakage Assessment Methodology. *Journal of Cryptographic Engineering* 6, 2 (2016), 85–99.
- [65] Emma Strubell et al. 2019. Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).
- [66] Pico Technology. 2020. . <https://www.picotech.com/oscilloscope/3000/picoscope-3000-oscilloscope-specifications>
- [67] Kris Tiri et al. 2004. A Logic Level Design Methodology for a Secure DPA Resistant ASIC or FPGA implementation. In *DATE '04*, Vol. 1.
- [68] Florian Tramer et al. 2018. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. *arXiv preprint arXiv:1806.03287* (2018).
- [69] Florian Tramèr et al. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security '16*.
- [70] Elena Trichina et al. 2004. Small Size, Low Power, Side Channel-Immune AES Coprocessor: Design and Synthesis Results. In *AES '04*.
- [71] Yaman Umuroglu et al. 2017. FINN: A Framework for Fast, Scalable Binarized Neural Network Inference. In *FPGA '17*.
- [72] Lingxiao Wei et al. 2018. I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators. In *ACSAC '18*.
- [73] Yun Xiang et al. 2020. Open DNN Box by Power Side-Channel Attack. *IEEE Transactions on Circuits and Systems II: Express Briefs* (2020).
- [74] Xilinx. 2010. *Spartan-6 FPGA Configurable Logic Block User Guide*. [https://www.xilinx.com/support/documentation/user\\_guides/ug384.pdf](https://www.xilinx.com/support/documentation/user_guides/ug384.pdf)
- [75] Mengjia Yan et al. 2020. Cache Telepathy: Leveraging Shared Resource Attacks to Learn DNN Architectures. In *USENIX Security '20*.
- [76] Honggang Yu et al. 2020. DeepEM: Deep Neural Networks Model Recovery through EM Side-Channel Information Leakage. (2020).
- [77] Pengyuan Yu et al. 2007. Secure FPGA Circuits using Controlled Placement and Routing. In *(CODES+ISSS) '07*.
- [78] Mark Zhao et al. 2018. FPGA-based Remote Power Side-Channel Attacks. In *2018 IEEE Symposium on Security and Privacy (SP)*.