

# On the Resilience of Deep Learning for Reduced-voltage FPGAs

Kamyar Givaki<sup>1</sup>, Behzad Salami<sup>2</sup>, Reza Hojabr<sup>1</sup>, S. M. Reza Tayarani<sup>1</sup>, Ahmad Khonsari<sup>1,3</sup>, Dara Rahmati<sup>3</sup>, Saeid Gorgin<sup>4</sup>, Adrian Cristal<sup>2,5,6</sup>, Osman S. Unsal<sup>2</sup>.

<sup>1</sup> School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

<sup>2</sup> Barcelona Supercomputing Center, Barcelona, Spain

<sup>3</sup> School of Computer Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

<sup>4</sup> Iranian Research Organization for Science and Technology (IROST), Tehran, Iran

<sup>5</sup> Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>6</sup> Artificial Intelligence Research Institute (IIIA), Centro Superior de Investigaciones Científicas (CSIC), Barcelona, Spain  
givakik@ut.ac.ir, behzad.salami@bsc.es, r.hojabr@ut.ac.ir, m.taiarani@ut.ac.ir, a\_khonsari@ut.ac.ir, dara.rahmati@ipm.ir, gorgin@irost.ir, adrian.cristal@bsc.es, osman.unsal@bsc.es

**Abstract**—Deep Neural Networks (DNNs) are inherently computation-intensive and also power-hungry. Hardware accelerators such as Field Programmable Gate Arrays (FPGAs) are a promising solution that can satisfy these requirements for both embedded and High-Performance Computing (HPC) systems. In FPGAs, as well as CPUs and GPUs, aggressive voltage scaling below the nominal level is an effective technique for power dissipation minimization. Unfortunately, bit-flip faults start to appear as the voltage is scaled down closer to the transistor threshold due to timing issues, thus creating a resilience issue.

This paper experimentally evaluates the resilience of the training phase of DNNs in the presence of voltage underscaling related faults of FPGAs, especially in on-chip memories. Toward this goal, we have experimentally evaluated the resilience of LeNet-5 and also a specially designed network for CIFAR-10 dataset with different activation functions of Rectified Linear Unit (Relu) and Hyperbolic Tangent (Tanh). We have found that modern FPGAs are robust enough in extremely low-voltage levels and that low-voltage related faults can be automatically masked within the training iterations, so there is no need for costly software- or hardware-oriented fault mitigation techniques like ECC. Approximately 10% more training iterations are needed to fill the gap in the accuracy. This observation is the result of the relatively low rate of undervolting faults, i.e., <0.1%, measured on real FPGA fabrics. We have also increased the fault rate significantly for the LeNet-5 network by randomly generated fault injection campaigns and observed that the training accuracy starts to degrade. When the fault rate increases, the network with Tanh activation function outperforms the one with Relu in terms of accuracy, e.g., when the fault rate is 30% the accuracy difference is 4.92%.

**Index Terms**—DNN, Hardware accelerator, FPGA, Training, Resilience, Voltage underscaling.

## I. INTRODUCTION

Hardware accelerators are designed to perform required computations in a specific application efficiently [1]–[6]. Deep Neural Networks (DNNs) need a huge amount of computations, which categorizes them as power- and energy-hungry applications. Using hardware accelerators is a promising solution to answer this requirement. Recently, many accelerators based on Graphics Processing Units (GPUs) [7], Field Programmable Gate Arrays (FPGAs) [8]–[11], and Application Specific Integrated Circuits (ASICs) [12], [13] have been proposed for

various DNNs. Among them, FPGAs have unique features, which makes them increasingly popular, thanks to their massively parallel architecture, reconfiguration capability, data-flow execution model, and the recent advances on the High-Level Synthesis (HLS) tools. However, the power consumption of FPGAs is still a key concern, especially comparing against equivalent ASICs, and FPGAs can be at least an order of magnitude less energy-efficient than ASIC designs [14]. To mitigate this gap, aggressive voltage underscaling is an effective solution [15], [16], by considering the quadratic saving in dynamic power and exponential saving in static power [17]; aggressive voltage underscaling can be described as decreasing the supply voltage of either whole or some components of a circuit below the nominal voltage which is set by the vendor. However, some reliability issues might appear as a result of the circuit delay increase. These timing issues may cause some faults; therefore, the circuit can produce the wrong results. Generally mitigating these effects is done by hardware design changes [18], [19] or by using the built-in Error Correction Code (ECC) of FPGAs [20]. These types of efforts are also carried out under research projects like LEGaTO [21], [22]. DNN applications are inherently tolerant to some faults. This is a unique property that distinguishes DNNs as good applications to apply aggressive voltage underscaling. The reason is that no additional hardware or technique would be needed to be applied to mitigate the effects of aggressive voltage underscaling in DNN applications.

Typically, a working DNN has two phases of operation. The first phase, *training*, is the process of tuning the parameters of a specific network. Training is an iterative task in which sample inputs are iteratively injected into the network. Then, the predicted results of the network are evaluated by the desired results using a specific function known as *loss function*. After that, the loss is propagated backward for parameter tuning. This process continues until the loss of the network decreases under a threshold value. In most cases, the process of training a network is performed only once. Fig. 1 shows a high-level abstraction of one iteration in the training process in which a sample enters the network, then the loss of the

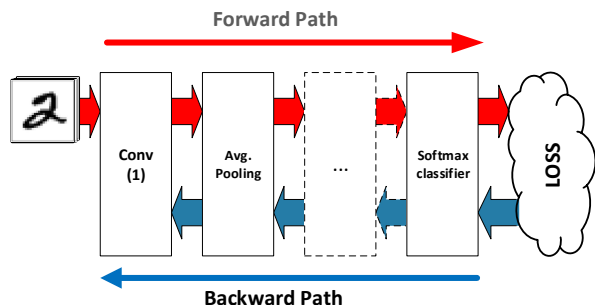


Fig. 1. High-level abstraction of an iteration in the training process of LeNet-5 network.

network is computed and the loss propagates to the first layer through the backward path. In the second phase (known as *inference*), a sample is injected into the network, and the network generates an output based on the parameters that are learned in the training process. In this paper, our focus is on the training phase of neural networks as it is more energy-hungry than the inference phase, and reducing the power consumption by voltage undervolting can be directly translated to lower energy consumption. The effect of the faults is mainly investigated in the inference phase of DNNs in hardware, software, embedded, and HPC platforms [23]. There are some recent efforts on the training phase too [24], [25], but these faults are related to manufacturing defects or soft errors. However, to the best of our knowledge, no work addresses undervolting related faults of COTS hardware in the training phase of DNNs.

In this paper, we contribute by examining the resilience of DNN training by using the on-chip SRAM-based memory fault maps of real FPGA fabrics [15]. Note that SRAMs play an important role in the structure of DNN accelerators [26]. They also have a significant contribution in range of 30%-70% on the total power consumption of such DNN systems [27], [28]. Thus, the focus of this paper is on the on-chip memories. Our experiments confirm the idea that the faults related to the aggressive voltage undervolting are masked in the training process due to the inherent fault resiliency of DNNs. The fault rate of undervolting FPGAs is less than 0.1%, and this fault rate has a negligible negative effect on the training process.

We found that with the higher fault rates of at least 25%, the DNN accuracy can be affected by 6.25% (with the same number of iterations). This gap can be filled with more training iterations. It should be mentioned that the accuracy of the network with the Hyperbolic Tangent (Tanh) activation function has been less affected by increasing the fault rate.

In a nutshell, we evaluate the resilience of DNN training in the presence of FPGA undervolting faults. More specifically, the contributions of this paper are listed below:

- The DNN training is inherently robust for undervolting-related faults, evaluated on the fault maps of real FPGA fabrics that are publicly available. This observation is due to the relatively low fault rate for modern FPGAs that is

measured up to 0.1%.

- We generate higher fault rates with uniform distribution to complete our experiments. For the LeNet-5 network, the fault rate of at least 25% can significantly affect the DNN accuracy.

The rest of the paper is organized as follows. The experimental methodology is introduced in Section II. The obtained results are presented and discussed in Section III. We review the related work in Section IV and finally the paper is concluded in Section V.

## II. EXPERIMENTAL METHODOLOGY

Our experiments are based on injecting undervolting-related faults into the inputs, weights, and intermediate values generated during the DNN training phase. To evaluate the resilience behavior, we compare the accuracy and loss of the faulty DNNs with the baseline without any faults. Below, we elaborate on the experimental methodology, including the fault and the DNN models, as well as the overall experimental setup.

### A. Fault Model

Voltage undervolting below the minimum safe voltage level, i.e.,  $V_{min}$ , can result in timing faults. In [15], this technique is investigated for modern FPGAs, specifically on SRAM-based on-chip memories. Reference [15] reports that the fault rate increases exponentially when the voltage is decreased below  $V_{min}$ . At the lowest voltage level that could be practically undervolting, i.e.,  $V_{crash}$ , (almost half of the default voltage level, i.e.,  $V_{nominal}$ ), the maximum fault rate observed is less than 0.1%. Also, it has been shown that the faults show a permanent behavior for a specific device, and their location does not typically change at a fixed voltage level. The undervolting fault maps, i.e., the distribution of faults in physical locations of memories at different voltage levels below the  $V_{min}$ , are released in [29] publicly. The undervolting fault maps are unique and per-FPGA, due to the process variation effects, demonstrated on VC707 and KC705 in [15]. More specifically, it has been explored in [15] that faults appear in  $[V_{min} = 0.6V, V_{crash} = 0.54V]$  and  $[V_{min} = 0.59V, V_{crash} = 0.53V]$  for a maximum of 23706 and 2274 faults for VC707 and KC705, respectively. It should be noted that in both FPGAs, the  $V_{nominal} = 1V$ . We utilize the fault map from [29] for each FPGA that precisely illustrates the location of flipped bits in each FPGA under undervolting.

We use these publicly-available undervolting fault maps and inject them into the DNN training and monitor the accuracy. Note that the total size of the available FPGA on-chip memories is limited, e.g., 4.5 MB, and 1.9 MB for VC707 and KC705, respectively. We use this memory to store inputs of the network, weights of the network, and intermediate values of computations, i.e., values of losses in each iteration during the training. So, memories are crucial components in implementing DNNs, and a significant fraction of the total power consumption of the whole system is related to these components. Therefore decreasing the power consumption in

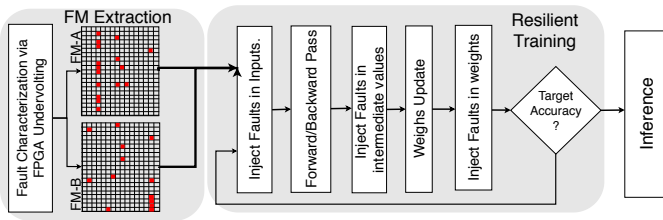


Fig. 2. Overall methodology

TABLE I  
THE DETAILED ARCHITECTURE OF LeNET-5. [30]

Layer type	Kernel size	Stride	# of channels	Activation function
Conv	$5 \times 5$	1	6	Relu or Tanh
Avg. pooling	$2 \times 2$	2	–	–
Conv	$5 \times 5$	1	16	Relu or Tanh
Avg. pooling	$2 \times 2$	2	–	–
FC layer	–	–	–	Sigmoid
FC layer	–	–	–	Sigmoid
Softmax classifier	–	–	–	Softmax

block RAMs can be directly translated to overall system power consumption reduction.

### B. DNN Models

To evaluate the impact of voltage underscaling on the training of neural networks, we apply our model on two convolutional neural networks. The first DNN model is LeNet-5 [30]. As illustrated in Table I, the network has two convolutional layers, each of them followed by an average pooling (sub-sampling) layer. Then two fully-connected layers and a softmax layer are placed to generate the desired output. We use the MNIST dataset [31] to train this network. MNIST contains 60000 samples of handwritten digits to train the network and also includes 10000 samples to test the training process. Each sample is a  $28 \times 28$  gray-scale image. We train this network for classification MNIST by more than  $10K$  iterations, and the top-1 classification accuracy achieved is 98.6% (normally, the classification accuracy reaches 99.5%, but more training iterations are needed).

The second network is a special architecture that is designed to classify images of the CIFAR-10 dataset [32], and the details of this network are shown in Table II. The dataset contains 50000 training images with a size of  $32 \times 32$  pixels and also 10000 test images. After more than  $100K$  of training iterations, the top-1 classification accuracy was 85.7% (the reported accuracy is for the case with Rectified Linear Unit (Relu) Activation function). It is constructed from four convolutional layers, one sigmoid fully connected layer, and a softmax layer.

### C. Overall Methodology

The overall experimental methodology is shown in Fig. 2. As seen, we utilize the fault map of the FPGA memories for different chips and inject these faults into the inputs (pixels of input images), weights of the DNN, and all values which

TABLE II  
DETAILED ARCHITECTURE OF THE CNN WHICH IS USED FOR CIFAR-10 DATASET.

Layer type	Kernel size	Stride	# of Channels	Activation function
Conv	$3 \times 3$	1	32	Relu or Tanh
Conv	$3 \times 3$	1	32	Relu or Tanh
Max pooling	$2 \times 2$	–	–	–
0.3 Dropout	–	–	–	–
Conv	$3 \times 3$	1	64	Relu or Tanh
Conv	$3 \times 3$	1	64	Relu or Tanh
Max pooling	$2 \times 2$	–	–	–
0.4 Dropout	–	–	–	–
FC layer	–	–	–	Sigmoid
FC layer	–	–	–	Softmax

are generated in both forward and backward paths of training process based on the location of them on block RAMs. After updating the faulty weights in each iteration, we repeat the process for other iterations, where faults appear at the same location, due to the permanent behavior of the undervolting faults.

In our simulations, after iteration  $i$ , the updated weights are obtained by injecting faults to the values based on the position that values are stored in block RAMs. Weights that are obtained in iteration  $i$  are used for the next iteration of training. So in  $(i + 1)$ -th iteration, the training process of network tries to eliminate the effects of faults in previous iterations of the training. It should be mentioned that the training for all the voltages has been performed in the same iterations, and the loss function in all experiments is categorical cross-entropy.

We use single-precision floating-point numbers (32-bits) to represent inputs, weights, and intermediate variables; unlike inference, the training process typically requires floating-point computation. Each word of block RAMs in employed FPGAs has 16-bits. So, two words of the block RAM are needed to store a single-precision floating-point number. A floating-point number is stored in two words of block RAM in a way that is shown in Fig. 3. Hence, according to the location of reported faults in block RAMs, some bit-flips may occur in sign-bit, mantissa, or exponent of the floating-point number.

Values of input images, weights, and intermediate values are stored in block RAM in the following manner: first, several block RAMs (two block RAMs for MNIST and six for CIFAR-10) of the FPGA are assigned to store values of pixels of the input images. These required block RAMs are selected randomly. According to the architecture of the network, several block RAMs are reserved for storing the latest value of weights. The two employed networks are small enough for all of their weights can be stored in block RAMs offered by FPGAs. These block RAMs are also selected randomly. Intermediate values that are generated in the process of training are written to other parts of block RAMs. When the capacity of the FPGA's block RAMs has been reached, the new intermediate values are substituted by previously generated ones. The replacement policy is First-In-First-Out (FIFO), in which the oldest intermediate value has replaced with the latest

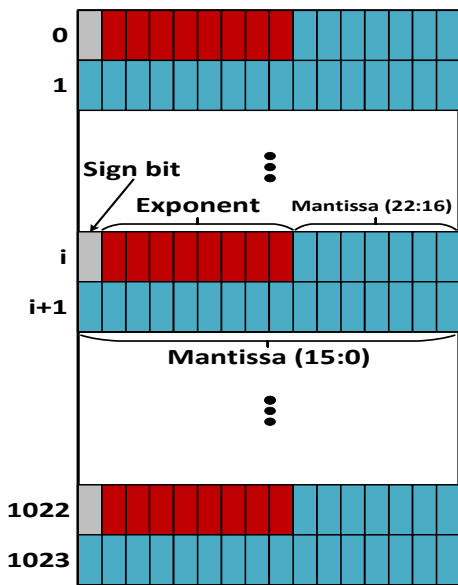


Fig. 3. Storing a 32-bit floating-point number (input, weight, and intermediate data) to two rows of a single block ram with the size of 16K bit, i.e., a matrix of 1024 rows and 16 columns. The intersection of each row and column represents a bitcell.

generated intermediate value.

By writing all values in the above-mentioned order, it is possible to determine the exact position of each variable (Input, weight, intermediate value). So, it can be possible to simulate the impact of voltage underscaling on the values and also the whole network, using the fault map of FPGA presented in [29].

### III. EXPERIMENTAL RESULTS

Fig. 4 shows the Lenet-5 network accuracy when it is used for the classification of the MNIST dataset. Fig. 4 (a) illustrates the accuracy of the network, which is simulated using the VC707 fault map, and the activation function is Relu. As can be seen in this figure, decreasing the voltage results in an increase in the fault rate, which results in a minimal and negligible decrease in the network accuracy. As illustrated in Fig. 4 (b), the accuracy of the network, which is simulated using the VC705 fault map and Relu activation function decreases slightly. Fig. 4 (c) and (d) show the simulations when the activation functions are Tanh. As Fig. 4 shows, when the real fault maps are used for simulations, there is not much difference in the accuracy results when the activation functions for convolutional layers are either Relu or Tanh.

Fig. 5 illustrates the accuracy of the classification of the CIFAR-10 dataset. We use the network that its structure is shown in Table II. When the network is simulated by the VC707 fault map, and activation function is Relu, as shown in Fig. 5 (a), the accuracy of the network is negligibly decreased by supply voltage decrease. Fig. 5 (b) shows a similar trend for VC705 fault map and Relu activation function. The accuracy of the network is reduced when we substitute the Relu activa-

tion function by Tanh. Although, as depicted in Fig. 5 (c) and (d), the trend of accuracy changes is similar to the trend of the Relu activation function in which the accuracy decreases with voltage decrease. In this case, the network with the Relu activation function is more accurate. It should be mentioned that the training for all the voltages has been performed with the same number of iterations (more than 10K iterations for MNIST and more than 100K for the CIFAR-10).

Fig. 6 shows the loss values for the training of these networks for two voltages ( $V_{min}$  and  $V_{crash}$ ). Fig. 6 (a) illustrates the loss of the LeNet-5 network when VC707 is employed. Since the LeNet-5 network is a small network with a low number of parameters, the loss for both the voltages follows the same trend, and there is no significant difference between the loss values.

On the other hand, Fig. 6 (b) shows as the network parameters increase, the gap between the loss values for two different voltages reveals (the network that is used for classification of the CIFAR-10 has more parameters than LeNet-5). This gap can be interpreted as follows: The lower voltage can decrease the convergence rate of the network. In other words, if we decrease the voltage, the training process needs more iterations typically to reach a specific accuracy point. For example, Table III shows the difference between the iterations to reach the accuracy of 98% in classification of MNIST in several cases and the iterations to reach the accuracy of 80% for CIFAR-10. Table III reveals, on average, 10% additional iterations can handle the effect of these faults in the accuracy of the networks. For example, to reach 98% top-1 accuracy in MNIST, approximately 200 more iterations are required.

Reference [15] has shown that in a VC707 FPGA, by decreasing the supply voltage of block RAMs from  $V_{min}$  to  $V_{crash}$ , the power consumption of block RAMs is decreased by 40%; however, this reduction can lead to reliability issues and result in some faults in the content of block RAMs. Our observations show that these faults can be masked in the process of training. As previously mentioned, in modern DNNs, 30%-70% of the total power consumption of the system is related to SRAMs. By combining the two above mentioned facts, it can be inferred that aggressive voltage underscaling can decrease the total power consumption of the system by at least 10%.

The reported fault rate in real FPGA fault maps is under 0.1%. To investigate the resilience of the training process to higher fault rates, we perform an experiment in which we generate fault maps with higher fault rates. Faults are randomly distributed in the whole block RAM spaces with uniform distribution, and simulations are performed for VC707 FPGA block RAM size and Lenet-5 network.

Fig. 7 (a) shows the training accuracy of LeNet-5 in two cases: In the first case, the activation functions of convolutional layers are Relu, and in the second one, the activation functions of convolutional layers are Tanh. In both cases, the accuracy remains high when the injected fault rate is lower than 25%. Then the accuracy decreases, and at a point between 30 and 40%, the accuracy curve breaks. Fig. 7 (b) shows the loss

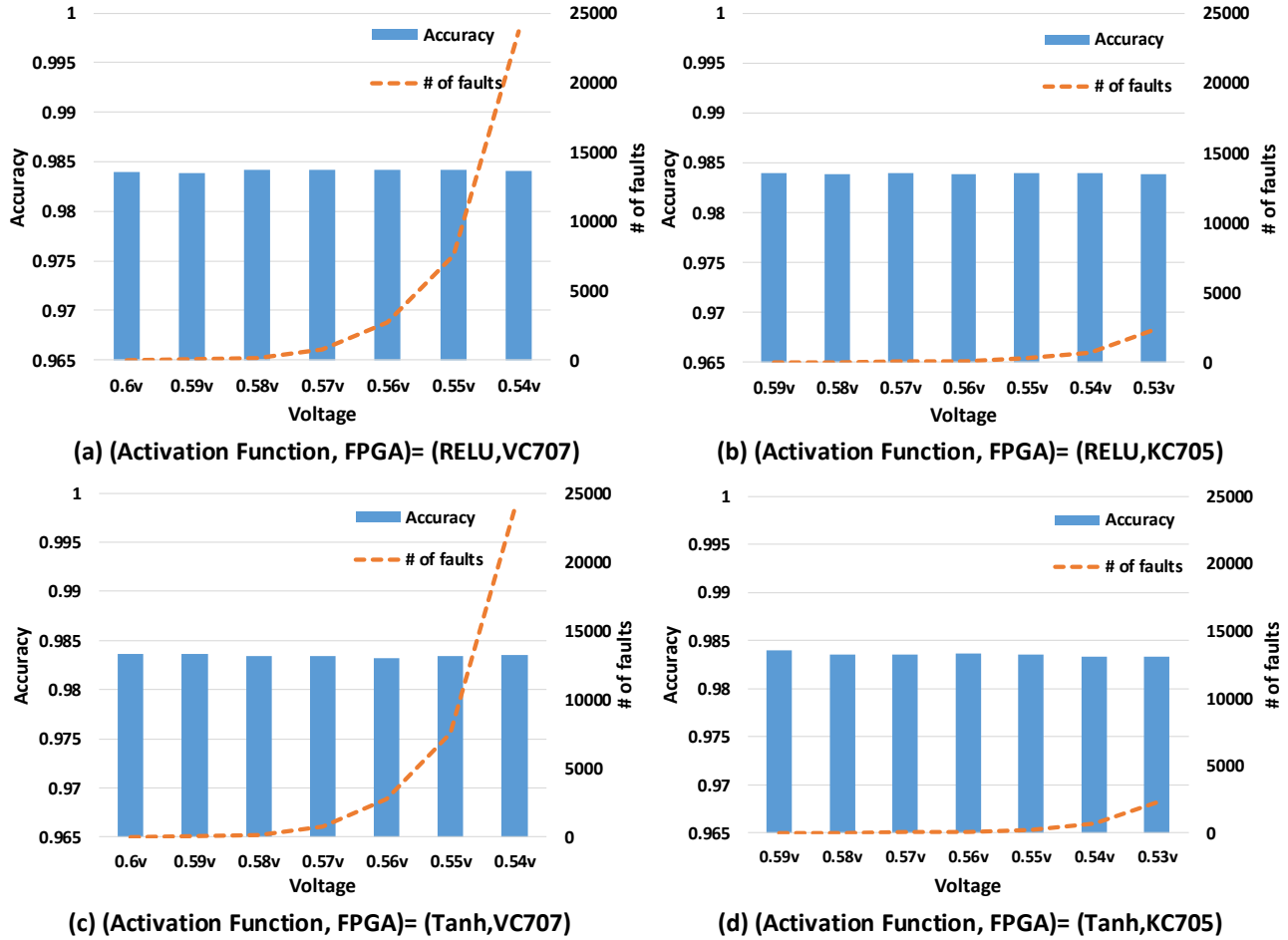


Fig. 4. The training accuracy of the LeNet-5 network in the classification of MNIST dataset under different voltages and with different number of faults for (Activation function, FPGA).

TABLE III  
COMPARISON OF THE NUMBER OF ITERATIONS TO REACH TO A SPECIFIC ACCURACY POINT.

Dataset	Reference accuracy	Device	Voltage	Iterations
MNIST (Relu)	98%	VC707	$V_{min}=0.6V$	4950
			$V_{crash}=0.54V$	5200
		KC705	$V_{min}=0.59V$	4900
			$V_{crash}=0.53V$	5050
CIFAR-10 (Tanh)	80%	VC707	$V_{min}=0.6V$	47800
			$V_{crash}=0.54V$	51200
		KC705	$V_{min}=0.59V$	43800
			$V_{crash}=0.53V$	61000

values for some fault rates. As seen, the network with the Tanh activation function outperforms the one with Relu since the injected fault rate is more than 15%; the Tanh network has 1.09% better accuracy than the Relu one. The gap between these two curves extends to 4.92% when the injected fault

rate increases to 30%. It can be inferred that in situations that fault rate is high, using the Tanh activation function may be helpful.

#### IV. RELATED WORKS

With technology scale developing, the resilience of DNNs can be significantly affected due to the fabrication process uncertainties, soft-errors, harsh and noisy environments, aggressively low-voltage operations, among others. Hence, recently, the resilience of DNNs has been studied in different abstraction levels. A vast majority of the previous works in this area belong to the DNN inference phase, including simulation-based efforts [33]–[36] and works on the real hardware [12], [37]–[39]. The verification of the simulation-based works on the real fabric can be a crucial concern; also, the real hardware works are mostly performed on the customized ASICs, which of course, reproducing those results on the COTS systems is a crucial question.



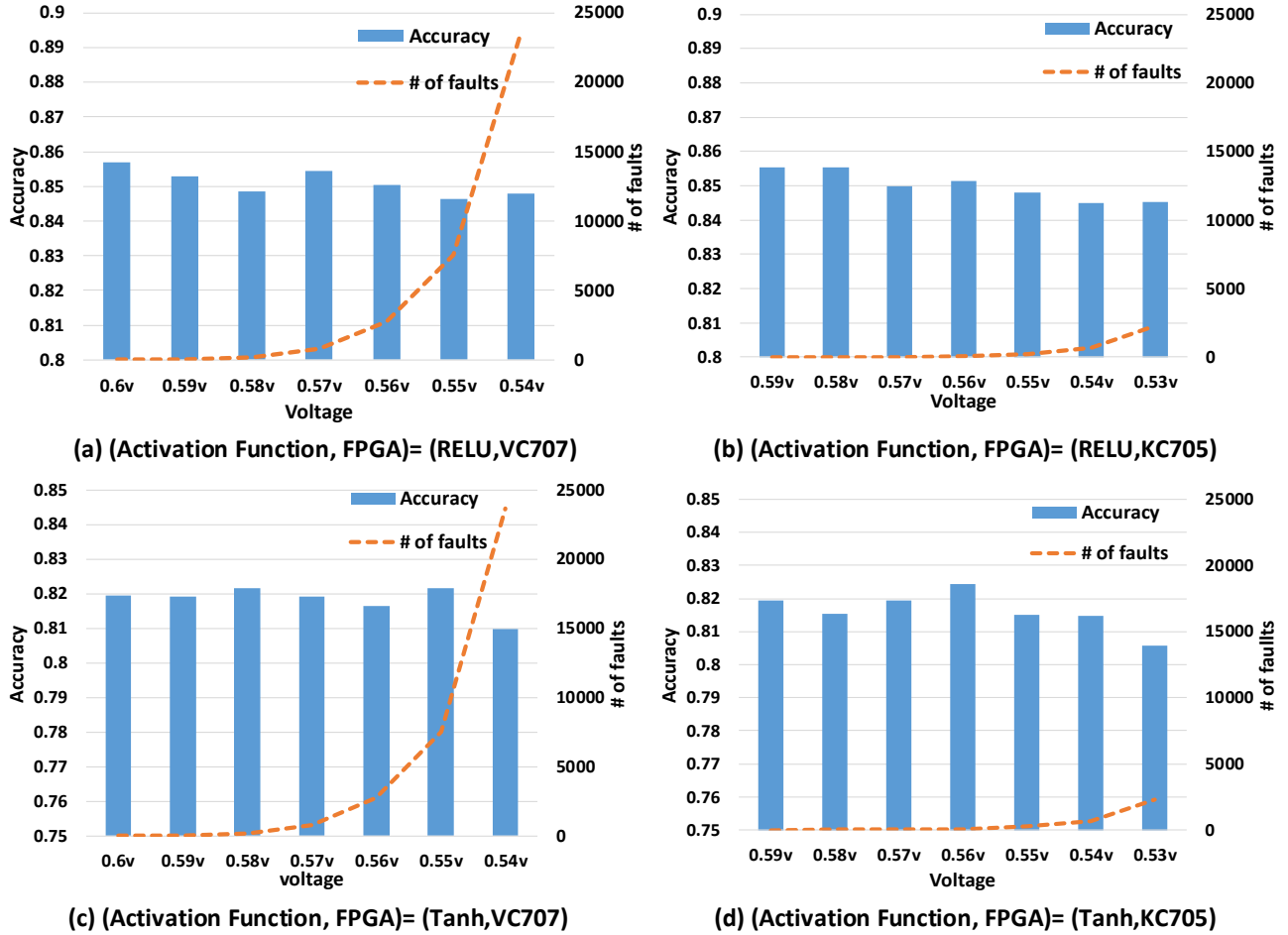


Fig. 5. The training accuracy of the network that is used in the classification of the CIFAR-10 dataset under different voltages and with the different number of faults for (Activation function, FPGA).

On the other hand, there are not thorough efforts on the resilience of the DNN training phase; recent works in part cover the study in this area [24], [25], [40]–[42]. For instance, [41], [42] have analyzed only the fully-connected model of DNNs, [24] carried out the analysis on a customized ASIC model of the DNN, and finally, [25] performed a simulation-based study. Our paper extends the study on the resilience of the DNN training, especially by using the fault maps of low-voltage SRAM-based on-chip memories of real FPGA fabrics.

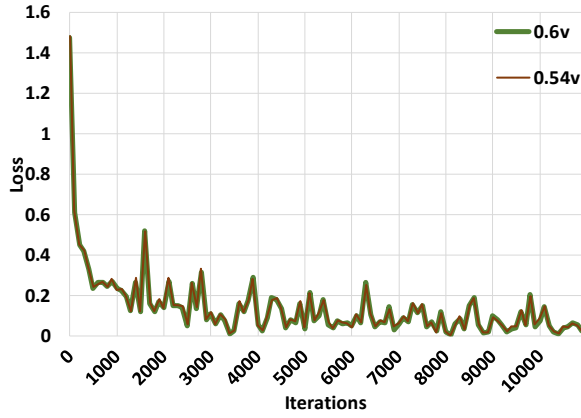
Our experimental methodology is based on emulating the real fault maps of FPGA-based SRAM memories during the DNN training iterations. A similar approach has been considered for real DRAMs, as well [43], [44]. Unlike the fully software-based approaches, our study is based on real fault maps, which can lead to more precise study. Also, unlike the fully hardware-based approach, our study is more facilitated and can be easily expanded for many different applications. In other words, our approach has the advantage of both full software [45], and fully real hardware [46] resilience study

approaches, similar to recent works [43], [47], [48].

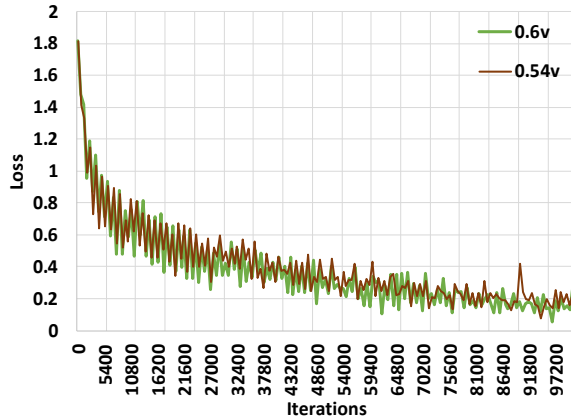
## V. CONCLUSION

In this paper, we experimentally evaluate the effect of aggressive voltage underscaling of FPGA block RAMs on the training phase of deep neural networks. Simulation results show that the training process of deep neural networks is resilient to faults that are generated because of the reduced-voltage supply. We observed that due to the low fault rate of real FPGA fabrics of up to 0.1%, the effect of these faults on the accuracy of the network is negligible and can be compensated, on average, by 10% more iterations in training.

Furthermore, the training process is resilient to fault rate more than the fault rate of real FPGAs. Additionally, our simulations show that with injecting 25% random faults to memory, the accuracy of the LeNet-5 Network in the classification of the MNIST dataset is only 6.25% for Relu activation function, and 2.75% for Tanh activation function lower than the training with no faults (in the same number of iterations).



(a) (Dataset, Activation Function)= (MNIST, RELU)



(b) (Dataset, Activation Function)= (CIFAR-10, RELU)

Fig. 6. The Loss of the networks for training using VC707 fault map (a) MNIST (b) CIFAR-10.

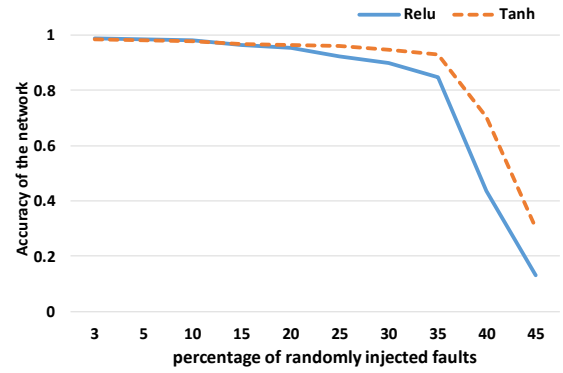
As an ongoing work, we are going to repeat our model on real FPGAs.

## VI. ACKNOWLEDGMENTS

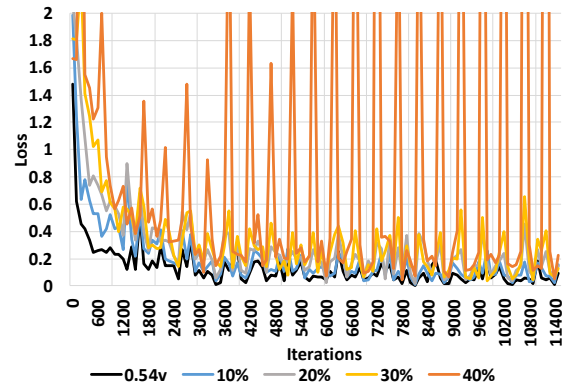
The research leading to these results has received funding from the European Unions Horizon 2020 Programme under the LEGaTO Project ([www.legato-project.eu](http://www.legato-project.eu)), grant agreement n<sup>o</sup> 780681.

## REFERENCES

- [1] O. Arcas-Abella, A. Armejach, T. Hayes, G. A. Malazgirt, O. Palomar, B. Salami, and N. Sonmez, "Hardware acceleration for query processing: leveraging fpgas, cpus, and memory," *Computing in Science & Engineering*, vol. 18, no. 1, p. 80, 2015.
- [2] B. Salami, O. Arcas-Abella, and N. Sonmez, "Hatch: hash table caching in hardware for efficient relational join on fpga," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2015, pp. 163–163.
- [3] B. Salami, G. A. Malazgirt, O. Arcas-Abella, A. Yurdakul, and N. Sonmez, "Axledb: A novel programmable query processing platform on fpga," *Microprocessors and Microsystems*, vol. 51, pp. 142–164, 2017.
- [4] B. Salami, O. Arcas-Abella, N. Sonmez, O. Unsal, and A. C. Kestelman, "Accelerating hash-based query processing operations on fpgas by a hash table caching technique," in *Latin American High Performance Computing Conference*. Springer, 2016, pp. 131–145.



(a) (Dataset, FPGA)=(MNIST, VC707)



(b) (Dataset, Activation Function, FPGA)=(MNIST, Relu, VC707)

Fig. 7. (a) The accuracy comparison of Relu and Tanh activation functions for randomly generated fault maps (b) Comparison of the network loss for several random generated fault maps with 0.54V real fault map for VC707 and LeNet-5 when the activation function is Relu.

- [5] O. Melikoglu, O. Ergin, B. Salami, J. Pavon, O. Unsal, and A. Cristal, "A novel fpga-based high throughput accelerator for binary search trees," *arXiv preprint arXiv:1912.01556*, 2019.
- [6] D. Gizopoulos, G. Papadimitriou, A. Chatzidimitriou, V. J. Reddi, B. Salami, O. S. Unsal, A. C. Kestelman, and J. Leng, "Modern hardware margins: Cpus, gpus, fpgas recent system-level studies," in *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. IEEE, 2019, pp. 129–134.
- [7] E. Nurvitadhi, G. Venkatesh, J. Sim, D. Marr, R. Huang, J. Ong Gee Hock, Y. T. Liew, K. Srivatsan, D. Moss, S. Subhaschandra *et al.*, "Can fpgas beat gpus in accelerating next-generation deep neural networks?" in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2017, pp. 5–14.
- [8] A. Yazdambakhsh, M. Brzozowski, B. Khaleghi, S. Ghodrati, K. Samadi, N. S. Kim, and H. Esmailzadeh, "Flexigan: An end-to-end solution for fpga acceleration of generative adversarial networks," in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2018, pp. 65–72.
- [9] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2015, pp. 161–170.
- [10] S. Li, C. Wu, H. Li, B. Li, Y. Wang, and Q. Qiu, "Fpga acceleration of recurrent neural network based language model," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2015, pp. 111–118.
- [11] A. Shawahna, S. M. Sait, and A. El-Maleh, "Fpga-based accelerators of

- deep learning networks for learning and classification: A review,” *IEEE Access*, vol. 7, pp. 7823–7859, 2018.
- [12] B. Reagen, P. Whatmough, R. Adolf, S. Rama, H. Lee, S. K. Lee, J. M. Hernández-Lobato, G.-Y. Wei, and D. Brooks, “Minerva: Enabling low-power, highly-accurate deep neural network accelerators,” in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2016, pp. 267–278.
- [13] R. Hojabr, K. Givaki, S. Tayaraniyan, P. Esfahanian, A. Khonsari, D. Rahmati, and M. H. Najafi, “Skippynn: An embedded stochastic-computing accelerator for convolutional neural networks,” in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 132.
- [14] E. Nurvitadhi, D. Sheffield, J. Sim, A. Mishra, G. Venkatesh, and D. Marr, “Accelerating binarized neural networks: Comparison of fpga, cpu, gpu, and asic,” in *2016 International Conference on Field-Programmable Technology (FPT)*. IEEE, 2016, pp. 77–84.
- [15] B. Salami, O. S. Unsal, and A. C. Kestelman, “Comprehensive evaluation of supply voltage underscaling in fpga on-chip memories,” in *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2018, pp. 724–736.
- [16] B. Salami, “Aggressive undervolting of fpgas: power & reliability trade-offs,” 2018.
- [17] S. Salamin, H. Amrouch, and J. Henkel, “Selecting the optimal energy point in near-threshold computing,” in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 1691–1696.
- [18] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner *et al.*, “Razor: A low-power pipeline based on circuit-level timing speculation,” in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2003, p. 7.
- [19] K. K. Chang, A. G. Yağlıkçı, S. Ghose, A. Agrawal, N. Chatterjee, A. Kashyap, D. Lee, M. O’Connor, H. Hassan, and O. Mutlu, “Understanding reduced-voltage operation in modern dram devices: Experimental characterization, analysis, and mechanisms,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 1, p. 10, 2017.
- [20] B. Salami, O. S. Unsal, and A. C. Kestelman, “Evaluating built-in ecc of fpga on-chip memories for the mitigation of undervolting faults,” in *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2019, pp. 242–246.
- [21] B. Salami, K. Parasyris, A. Cristal, O. Unsal, X. Martorell, P. Carpenter, R. De La Cruz, L. Bautista, D. Jimenez, C. Alvarez *et al.*, “Legato: Low-energy, secure, and resilient toolset for heterogeneous computing,” *arXiv preprint arXiv:1912.01563*, 2019.
- [22] A. Cristal, O. S. Unsal, X. Martorell, P. Carpenter, R. De La Cruz, L. Bautista, D. Jimenez, C. Alvarez, B. Salami, S. Madonar *et al.*, “Legato: towards energy-efficient, secure, fault-tolerant toolset for heterogeneous computing,” in *CF’18 Proceedings of the 15th ACM International Conference on Computing Frontiers*. Association for Computing Machinery (ACM), 2018, pp. 276–278.
- [23] A. Kumar and S. Mehta, “A survey on resilient machine learning,” *arXiv preprint arXiv:1707.03184*, 2017.
- [24] J. J. Zhang, T. Gu, K. Basu, and S. Garg, “Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator,” in *2018 IEEE 36th VLSI Test Symposium (VTS)*. IEEE, 2018, pp. 1–6.
- [25] G. B. Hacene, F. Leduc-Primeau, A. B. Soussia, V. Gripon, and F. Gagnon, “Training modern deep neural networks for memory-fault robustness,” in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2019, pp. 1–5.
- [26] K. Guo, S. Zeng, J. Yu, Y. Wang, and H. Yang, “A survey of fpga-based neural network accelerator,” *arXiv preprint arXiv:1712.08934*, 2017.
- [27] S. Salamat, B. Khaleghi, M. Imani, and T. Rosing, “Workload-aware opportunistic energy efficiency in multi-fpga platforms,” *arXiv preprint arXiv:1908.06519*, 2019.
- [28] F. Conti, P. D. Schiavone, and L. Benini, “Xnor neural engine: A hardware accelerator ip for 21.6-fj/op binary neural network inference,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2940–2951, 2018.
- [29] “Fpga brams undervoltig study,” <https://github.com/behzadsalami/FPGA-BRAMs-Undervoltig-Study>.
- [30] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [31] “The mnist database of handwritten digits,” <http://yann.lecun.com/exdb/mnist/>.
- [32] “The cifar-10 dataset,” <https://www.cs.toronto.edu/~kriz/cifar.html>.
- [33] J. J. Zhang, K. Basu, and S. Garg, “Fault-tolerant systolic array based accelerators for deep neural network execution,” *IEEE Design & Test*, 2019.
- [34] G. Li, S. K. S. Hari, M. Sullivan, T. Tsai, K. Pattabiraman, J. Emer, and S. W. Keckler, “Understanding error propagation in deep learning neural network (dnn) accelerators and applications,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2017, p. 8.
- [35] W. Choi, D. Shin, J. Park, and S. Ghosh, “Sensitivity based error resilient techniques for energy efficient deep neural network accelerators,” in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 204.
- [36] B. Salami, O. S. Unsal, and A. C. Kestelman, “On the resilience of rtl nn accelerators: Fault characterization and mitigation,” in *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. IEEE, 2018, pp. 322–329.
- [37] J. Zhang, K. Rangineni, Z. Ghodsi, and S. Garg, “Thundervolt: enabling aggressive voltage underscaling and timing error resilience for energy efficient deep learning accelerators,” in *Proceedings of the 55th Annual Design Automation Conference*. ACM, 2018, p. 19.
- [38] P. Pandey, P. Basu, K. Chakraborty, and S. Roy, “Greentpu: Improving timing error resilience of a near-threshold tensor processing unit,” in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 173.
- [39] N. Chandramoorthy, K. Swaminathan, M. Cochet, A. Paidimarri, S. Eldridge, R. Joshi, M. Ziegler, A. Buyuktosunoglu, and P. Bose, “Resilient low voltage accelerators for high energy efficiency,” in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2019, pp. 147–158.
- [40] B. W. Denkinger, F. Ponzina, S. S. Basu, A. Bonetti, S. Balási, M. Ruggiero, M. Peón-Quirós, D. Rossi, A. Burg, and D. Atienza, “Impact of memory voltage scaling on accuracy and resilience of deep learning based edge devices,” *IEEE Design & Test*, 2019.
- [41] S. Kim, P. Howe, T. Moreau, A. Alaghi, L. Ceze, and V. S. Sathé, “Energy-efficient neural network acceleration in the presence of bit-level memory errors,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, no. 99, pp. 1–14, 2018.
- [42] S. Kim, P. Howe, T. Moreau, A. Alaghi, L. Ceze, and V. Sathé, “Matic: Learning around errors for efficient low-voltage neural network accelerators,” in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1–6.
- [43] S. Koppula, P. Orosa, A. G. Yağlıkçı, R. Azizi, T. Shahroodi, K. Kanellopoulos, and O. Mutlu, “Eden: Enabling energy-efficient, high-performance deep neural network inference using approximate dram,” in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 2019, pp. 166–181.
- [44] M. Widmer, A. Bonetti, and A. Burg, “Fpga-based emulation of embedded drams for statistical error resilience evaluation of approximate computing systems,” in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 36.
- [45] C.-K. Chang, W. Yin, and M. Erez, “Assessing the impact of timing errors on hpc applications,” 2019.
- [46] R. Bertran, A. Buyuktosunoglu, P. Bose, T. J. Slegel, G. Salem, S. Carey, R. F. Rizzolo, and T. Strach, “Voltage noise in multi-core processors: Empirical characterization and optimization opportunities,” in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2014, pp. 368–380.
- [47] B. W. Denkinger, F. Ponzina, S. S. Basu, A. Bonetti, S. Balási, M. Ruggiero, M. Peon Quiros, D. Rossi, A. P. Burg, and D. Atienza Alonso, “Impact of memory voltage scaling on accuracy and resilience of deep learning based edge devices,” Tech. Rep., 2019.
- [48] A. Chatzidimitriou, G. Panadimitriou, D. Gizopoulos, S. Ganapathy, and J. Kalamatianos, “Assessing the effects of low voltage in branch prediction units,” in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2019, pp. 127–136.