Instituto Universitario de Investigación
de Ingeniería de Aragón
Universidad Zaragoza

Impact on the Accuracy of Aggressive Voltage Underscaling in CNN Accelerators

Universidad Zaragoza
1542

Yamilka Toca Díaz
Nicolás Landeros Muñoz
Rubén Gran Tejero
Alejandro Valero

INTRODUCTION

- In the last decades, CMOS fabrication technology has evolved into smaller node sizes, which in turn has allowed higher integration densities and capabilities. However, energy and power requirements of new technology generations have not evolved so positively with respect to the overall energy constraints for computing systems.
- Variations in the manufacturing process of CMOS technology impose conservative operation margins that impact the efficiency of the entire system. An example is the transistor's supply voltage (V_{dd}) [3] .
- This voltage is conservatively set above the safe voltage (V_{min}) margin imposed by the worst-case transistor to ensure a reliable operation of all transistors.
- Aggressive supply voltage underscaling below V_{min} causes permanent failures in vulnerable transistors.
- In this work we aim to characterize the resilience of CNN accelerators to such permanent failures.

CNN Accelerator Architecture

- A Convolutional Neural Network (CNN) accelerator is a specific hardware focusing on artificial intelligence workloads. It usually includes relatively large and energy-hungry memory buffers that hold either neural network's activations or weights.
- Our accelerator [*Fig. 1*] consists of of a 16×16 Processing Element (PE) array, on-chip memory storage to reduce costly off-chip memory accesses, dispatchers for every memory, and a control unit [2].
- On-chip intermediate storage includes a pair of 2 MiB activation memories and a 2 MiB weight memory.
- Activation memories swap their roles after the computation of every network layer. On the other hand, the weight memory caches weights to be issued in the proper order by the dispatcher to the PE array.

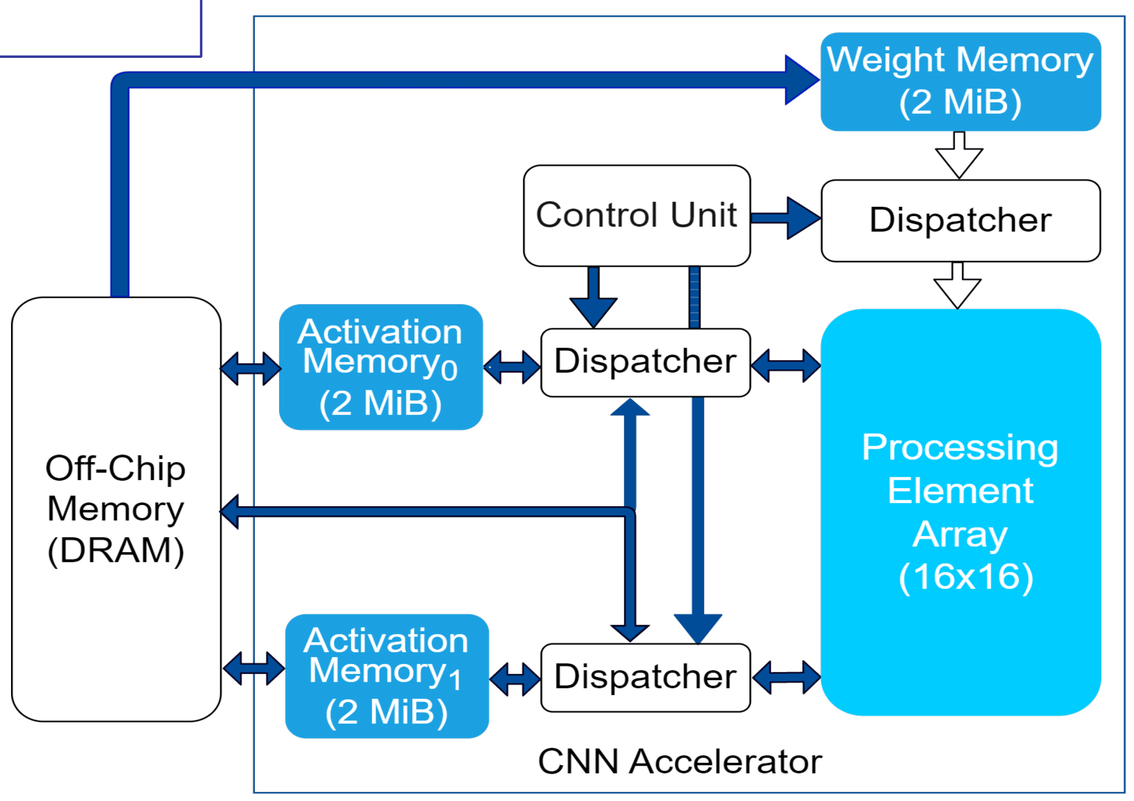

Figure 1: Overview of the baseline CNN accelerator.

RELIABILITY MODEL

	0.60 V	0.59 V	0.58 V	0.57 V	0.56 V	0.55 V	0.54 V	0.53 V
VC_707	2	18	90	426	1570	4124	13204	N/A
KC_705	0	2	8	28	72	298	802	2640

Table. 1: Number of faulty bits in a 2 MiB memory array varying V_{dd} for two different FPGAs (Virtex-7 and Kintex-7). N/A: Not Applicable (FPGA stops operating)

- We model the same number and the spatial distribution of failures as in [3].
- The results show an exponential increase in the number of faults with respect to the voltage.

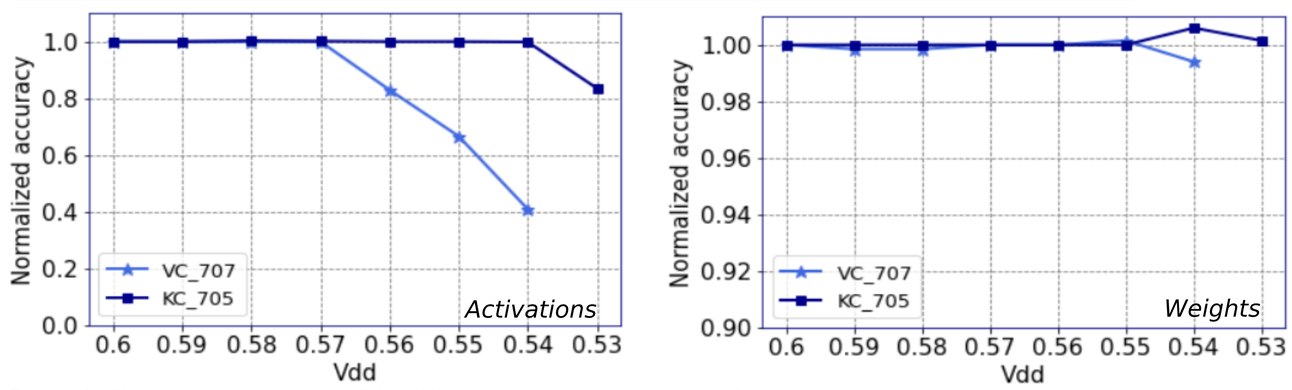

Figure 2: Normalized accuracy of AlexNet with respect to the fault-free accuracy when faults are injected in activations or weights

We assume a 16-bit fixed-point representation for both activations and weights, adjusting the number of integer and fractional bits for each benchmark to ensure the same accuracy as obtained with a IEEE-754 floating-point data representation [1].

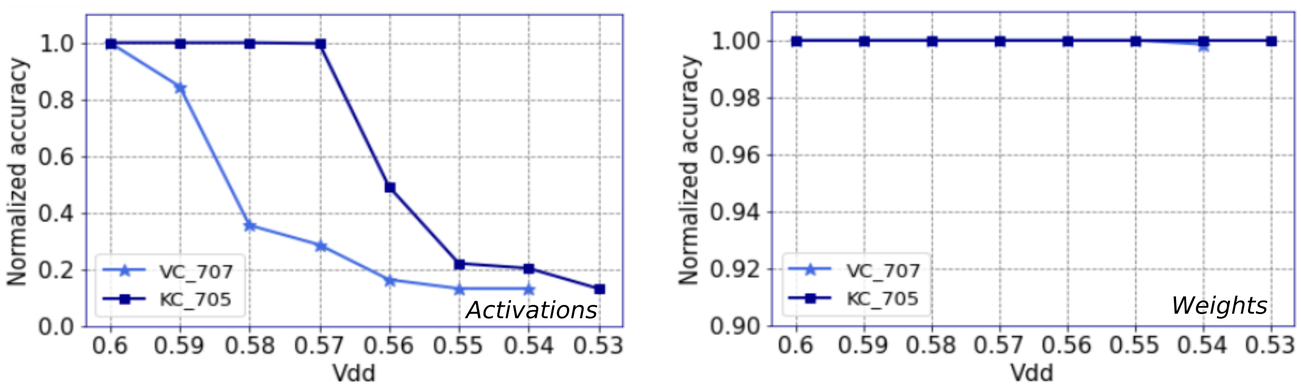

Figure 3: Normalized accuracy of SqueezeNet with respect to the fault-free accuracy when faults are injected in activations or weights

EXPERIMENTAL EVALUATION

Impact of aggressive voltage underscaling on the accuracy of CNN accelerators:

- Both fault bitmaps, VC_707 and KC_705, progressively degrade the accuracy of the studied benchmarks as V_{dd} decreases [*Fig. 2 and Fig. 3*].
- Regardless of the fault model, a similar number of faults induces a similar degradation of the accuracy for each benchmark.
- Both CNNs are more sensitive to faulty bitcells in activations than in weights.
- For activations, there is a higher accuracy drop in SqueezeNet [*Fig. 3*] than in AlexNet [*Fig. 2*], since the number of activation integer bits in SqueezeNet (6 bits) is higher than those in AlexNet (4 bits).
- In Alexnet, activations and weights require the same number of bits but activations are still more vulnerable.

There are two main reasons that could explain why activations are more vulnerable than weights:

- An inference pass consists of a sum of weighted activations. Therefore, a faulty weight has a limited effect on the final output activation value.
- The *Batch Normalization and Rectified Linear* units in the processing matrix could reduce large deviations in output activations as a consequence of faulty weights.

These results point out that aggressively undervolting CNN accelerator buffers beyond V_{min} for energy saving purposes requires an additional effort to prevent significant accuracy drops.

CONCLUSIONS

- Aggressive voltage underscaling below the safe threshold voltage margin is an effective solution to save energy in microprocessors.
- Convolutional Neural Networks(CNNs) have been identified as resilient to faults.

Experimental results have shown the following outcomes:

- The number of faulty bits has higher impact on accurancy than the faulty bitmap distribution.
- The number of faulty bits that severely degrade the accurancy ranges from tens to thousands.
- Despite the inherent resiliency of CNNs to faults, supply voltages of 0.54 V or 0.58 V, depending on the CNN, result in unacceptable accuracy degradation.
- Memories storing the activations are more sensitive to faults than weights.
- Depending on the benchmark, the accuracy reduction drops as much as 60% and 88% with respect to the original accuracy for a sufficient number of faults.

REFERENCES

[1] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In Proceedings of the 43rd International Symposium on Computer Architecture, page 243–254, 2016.

[2] Nicolás Landeros Muñoz, Alejandro Valero, Rubén Gran Tejero, and Davide Zoni. Gated-CNN: Combating NBTI and HCI aging effects in on-chip activation memories of Convolutional Neural Network accelerators. Journal of Systems Architecture, 128:1–13, 2022.

[3] Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman. Comprehensive Evaluation of Supply Voltage Underscaling in FPGA on-Chip Memories. In Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture, pages 724–736, 2018.

[4] Alejandro Valero, Darío Suárez-Gracia, and Rubén Gran-Tejero. DC-Patch: A Microarchitectural Fault Patching Technique for GPU Register Files. IEEE Access, 8:173276–173288, 2020.

This work has been partially supported by Agencia Estatal de Investigación (AEI) under Grant PID2019-105660RB-C21/AEI/10.13039/50110001103 3, and by Dept. of Science, University, and Knowledge Society, Government of Aragon, under Grant gaZ: T5820R research group. Y. Toca Díaz is supported by Universidad de Zaragoza-Banco Santander Ph.D. scholarship.
The authors thank HiPEAC for the given grant to participate in the event.