Impact on the Accuracy of Aggressive Voltage Underscaling in CNN Accelerators

Yamilka Toca-Díaz[1], Nicolás Landeros Muñoz[2], Alejandro Valero[1], and Rubén Gran-Tejero[1]

[1] Dept. of Computer Science and Systems Engineering, Universidad de Zaragoza, Spain
e-mail: yamilka@unizar.es
University of Zaragoza, María de Luna, Street 50018, Zaragoza, Spain.
[2] Dipt. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

ABSTRACT

Chip designers usually rely on conservative supply voltage (Vdd) guardbands to prevent permanent faults as a consequence of CMOS process variations. Convolutional Neural Networks (CNNs) can be resilient to faults since they usually include significant amounts of data redundancy. This paper shows that the accuracy CNNs, is compromised when the Vdd of a CNN accelerator is reduced.

Framework Overview

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. The Voltage overscaling unnecessarily increases energy consumption. A Convolutional Neural Network (CNN) accelerator is a specific hardware focused on artificial intelligence workloads. These accelerators usually include relatively large and energy-hungry memory buffers that hold either neural network's activations or weights. Focusing on energy savings, an effective solution is to aggressively underscale the Vdd of such buffers beyond Vmin, and somehow hide the permanent faults appearing in vulnerable bitcells as a consequence of reducing the supply voltaje [VSGGT20]. This work characterizes the resiliency of CNN accelerators to such permanent faults.

In this work, we extend the TensorFlow 2.5.0 simulation framework to model a CNN accelerator consisting of an inference processing matrix and on-chip storage. The on-chip storage consists of two input/output buffers for the storage of activations, and another buffer for weights, each of them with a storage capacity of 2 MiB [LVTZ22]. In order to study the impact of voltage underscaling in these buffers, we have assumed the faulty maps of two real FPGA platforms [SSUCK18]. Table 1 shows the number of faulty bits in a 2 MiB memory array varying Vdd from 0.60 to 0.53 V for both FPGAs. We model not only the same number of faults, but also the same spatial distribution of faults. Results show an exponential increase in the number of faults with respect to the voltage underscaling. Notice too that process variations affect much more to the VC_707 FPGA. As benchmarks, we consider two CNNs, AlexNet and SqueezeNet, focusing on image classification. We also 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 [HLM+16]. In particular, AlexNet represents the integer part of both activations and weights with 4 bits, whereas SqueezeNet requires 6 integer bits for activations and weights do not require any integer bit. The remaining bits are used for the sign and fractional part.

Table 1: Number of faulty bits in a 2 MiB memory array varying Vdd for two different FPGAs. N/A: not applicable.

	0.60	0.59	0.58	0.57	0.56	0.55	0.54
VC_707	2	18	90	426	1570	4124	13204
KC_705	0	2	8	28	72	298	802

Experimental Evaluation

This section evaluates the impact of aggressive voltage underscaling on the accuracy of CNN accelerators. Results differentiate between fault

injection in activations or weights. Figures 1 and 2 plot the normalized accuracy degradation for AlexNet and SqueezeNet, respectively, according to the failure bitmaps discussed in the previous section. Remark that the Y-axis ranges from 0.9 to 1.0 for weights. As expected, both fault bitmaps, VC_707 and KC_705, progressively degrade the accuracy of the studied benchmarks as Vdd reduces. Notice that, independently of the fault model, a similar number of faults induces a similar degradation on the accuracy. In other words, the accuracy degradation is similar despite the distribution of faults being different on each model. This is mainly due to activation integer values usually require more bits than weight integer values. That is, there is a higher probability of faulty bits in the integer part, and therefore, a larger range of values to deviate from the fault-free value. This can be appreciated in SqueezeNet, where no weight integer bits are required and the precision does not degrade. Similarly, for activations, there is a higher accuracy drop in SqueezeNet than in AlexNet, since the number of activation integer bits in SqueezeNet (6 bits) is higher than those in AlexNet (4 bits).

However, in the case of Alexnet, activations and weights require the same number of bits but activations are still more vulnerable. In fact, the accuracy slightly improves over the original accuracy for a large number of faulty weights.
Overall, 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. These results point out that aggressively undervolting CNN accelerator buffers beyond Vmin for energy saving purposes requires an additional effort to prevent significant accuracy drops.

Conclusions

Aggressive voltage underscaling below the safe voltage margin is an effective solution to sabe energy in microprocessors. However, such low-power operation modes impose a reliability challenge, since vulnerable transistors experience permanent faults. Experimental results have shown the following outcomes. First, the number of faulty bits has a higher impact on accuracy than the distribution of faults. Second, the number of faulty bitcells that severely degrade the accuracy ranges from tens to thousands and depends on the neural network. Third, 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. Finally, input/output buffers storing activations are more sensitive to faults than weight buffers.

References

[HLM+16] 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.

[LVTZ22] 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.

[SSUCK18] 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.

[VSGGT20] Alejandro Valero, Darío Suárez-Gracia, and Rubén Gran-Tejero. DC-Patch: AMicroarchitectural Fault Patching Technique for GPU Register Files. IEEE Access,8:173276–173288, 2020

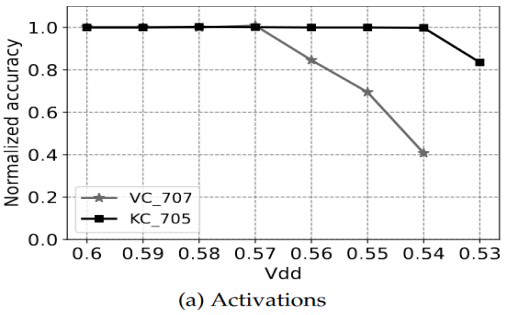

Figure 1: Normalized accuracy of AlexNet with respect to the fault-free accuracy when faults are injected in activations.

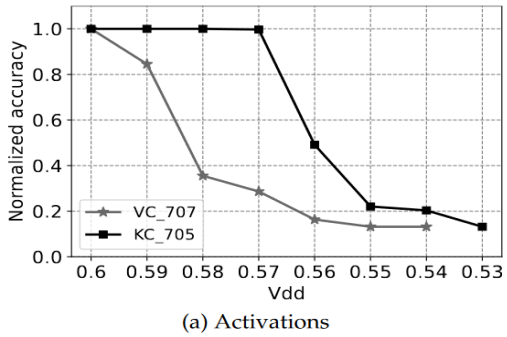

Figure 2: Normalized accuracy of SqueezeNet with respect to the fault-free accuracy when faults are injected in activations.