BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Hardware Accelerator of Neural Network Inference Supporting Fault Injections

Filip Masár*

Abstract

Neural networks (NNs) are becoming increasingly popular. Inference is now performed not only on high-end GPUs, but also on low-power embedded systems. Small faults in the hardware can lead to critical failures. This paper explores ways to test fault tolerance on the hardware accelerator of neural networks. We propose the use of FPGAs to increase the performance of fault tolerance experiments. To achieve this goal, an open source NN accelerator was used and modified to support fault injection. This solution provides high speed evaluation without loss of accuracy, as is the case with approximation-based simulations. Furthermore, an analysis of the fault tolerance of ResNet-18 is presented to demonstrate the proposed solution.

*xmasar18@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Overview

Neural networks (NNs) are becoming increasingly popular and are used in many systems today. Inference is performed not only on high-end GPUs and CPUs, but also in low-power embedded systems. Embedded systems can be sensitive to faults, and in some areas, such as autonomous driving or automotive systems, a fault can critical failure. Therefore, the fault tolerance of such systems is worth investigating. In the literature, this property is typically analysed in simulators that lack simulation quality or speed [1] [2] [3]. The aim of this work is to create a real NN inference accelerator in hardware that supports the emulation of fault injections.

2. Architecture

The architecture of the accelerator follows the patterns of state-of-the-art accelerators such as EyeRiss [4] or Google TPU [5]. The NVIDIA Deep Learning Accelerator (NVDLA) [6] was our choice for the upgrade. It is a configurable accelerator based on a grid of processing elements (PEs) that perform the multiply-accumulate (MAC) operation that is crucial for NN inference. The architecture of the accelerator is shown in Figure 1 and 2.

The target platform chosen was the Zynq UltraScale+ XCZU7EV SoC. This hardware limits us to the small configuration. However, it is still possible to use the advanced large configuration for more powerful FPGAs such as Virtex UltraScale. The small configuration allows us to perform inference with 8-bit data precision, which seems sufficient for the hardware implementations and is widely supported in SoC accelerators. In addition to an advanced caching system, the large configuration supports some techniques such as data reshaping, Winograd convolution or weight compression.

The original NVDLA design targets the ASIC flow, and some modifications have been made to support configurable FPGAs. The most advanced part of the accelerator design was a software stack for the Linux subsystem. In this work, we adapted the kernel to support the proposed accelerator. We also used the Tengine framework [7] to compile the pre-trained neural network coming from the Caffe tool.

A comparison of ResNet-18 inference performance on the processor and accelerator is shown in Table 1. The NN has been quantized to 8 bits. The results on NVDLA are three times faster than on the CPU.

3. Fault injection

Faults can enter the system in a number of ways. The memory can be corrupted. This can be easily emulated by changing a few bits in memory at the software kernel level. It is more difficult to emulate the fault injection in the computational path. We decided to extend the architecture of the PEs to integrate the fault injector. There is a grid of 8x8 multipliers that can produce faulty output. We support the following fault types: stuck-at (0, 1) and pulse. The system is able to change the injected value at runtime without having to restart the system. However, users can easily implement their own fault injector, e.g. based on a certain fault probability. The level of fault injection can be controlled using special configuration wires (red wires in Figure 4).

4. Results

The ResNet-18 trained on CIFAR-10 was used to test the fault tolerance of the NVDLA. The neural network was trained with an accuracy of 75.1%. The Figure 5 shows the loss of accuracy as a function of the number of multipliers affected. For example, the 6th and 7th MAC units are more sensitive to error when only one multiplier is affected by an error that changes their output to zero.

The faults can only change individual bits at the output of multipliers. The Figure 6 shows that the accuracy is negligibly reduced if the LSBs up to the 8th bit are affected. In addition, the neural networks are more sensitive to changes from 0 to 1 than from 1 to 0.

5. Conclusion

The proposed accelerator supports real-time fault injection into the NN inference. Although there are some limitations in the basic pre-trained NNs coming from the older version of the Caffe framework, the proposed accelerator can emulate the faults in real hardware. It allows researchers to investigate the parameters of the NNs and can help them to make the NNs more robust.

Acknowledgements

I would like to thank my supervisor Ing. Vojtěch Mrázek Ph.D. for his help.

References

[1] Ahmadilivani, M. H.; Barbareschi, M.; Barone, S.; Bosio, A.; Daneshtalab, M. et al. *Special Session: Approximation and Fault Resiliency of DNN Accelerators*. 2023.

[2] Taheri, M.; Daneshtalab, M.; Raik, J.; Jenihhin, M.; Pappalardo, S. et al. *SAFFIRA: a Framework for Assessing the Reliability of Systolic-Array-Based DNN Accelerators*. 2024.

[3] Pappalardo, S.; Ruospo, A.; O'Connor, I.; Deveautour, B.; Sanchez, E. et al. A Fault Injection Framework for AI Hardware Accelerators. In: *2023 IEEE 24th Latin American Test Symposium (LATS)*. March 2023, p. 1–6. ISSN 2373-0862.

[4] Chen, Y.-H.; Krishna, T.; Emer, J. S. and Sze, V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, 2017, vol. 52, no. 1, p. 127–138.

[5] Jouppi, N. P.; Kurian, G.; Li, S.; Ma, P.; Nagarajan, R. et al. *TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings*. 2023.

[6] NVIDIA Corporation. *NVIDIA Deep Learning Accelerator* online. 2018. Available at: http://nvdla.org/. [cit. 2024-04-02].

[7] OPEN AI LAB. *Tengine* online. Available at: https://github.com/OAID/Tengine. [cit. 2024-04-02].