READ: Reliability-Enhanced Accelerator Dataflow Optimization using Critical Input Pattern Reduction

Zuodong Zhang[1], Meng Li [2,1,⋆], Yibo Lin[1,3], Runsheng Wang[1,3], and Ru Huang[1,3]

[1]*School of Integrated Circuits, Peking University, Beijing, China*
[2]*Institute for Artificial Intelligence, Peking University, Beijing, China*
[3]*Beijing Advanced Innovation Center for Integrated Circuits, Beijing, China*

Deep neural networks (DNNs) have revolutionized different applications ranging from computer vision to natural language processing, and are widely deployed in data centers and edge devices. It can be foreseen that DNNs will be applied in more and more safety-critical applications like autonomous driving and robotics, which typically require highly reliable computing to avoid catastrophic consequences. Therefore, not only the model's robustness against various perturbations like adversarial noise, but also the robustness of the silicon-based accelerators to hardware faults needs to be comprehensively investigated [1], [2].

As the fabrication of DNN accelerators pushes toward nanoscale, the transient soft errors like timing errors that cannot be detected during manufacturing tests have become a more pronounced problem [3], [4]. Timing errors due to the increased path delay usually occur under process, voltage, temperature, and aging (PVTA) variations. Although DNN shows inherent error resilience at the algorithm level, timing errors are shown to cause significant accuracy degradation [5], [6]. This is because, on one hand, timing errors often occur in the most significant bit; while on the other hand, the error will accumulate in each convolution operation and across the whole network.

Several timing error-resilient accelerator designs have been explored from the architecture level to the circuit level. These works either utilize timing error detection and correction (TEDC) schemes to recover the correct value [5]–[7], or algorithm-based fault tolerance (ABFT) techniques to check the correctness of computing [8], [9]. However, these approaches usually compromise network accuracy or introduce large hardware overhead.

In this extended abstract, we provide a promising solution to alleviate the timing error in DNN accelerators from a new perspective. We propose READ, a reliability-enhanced accelerator dataflow optimization technique. READ mitigates the accuracy loss of DNN accelerators due to timing errors by exploiting sequence optimization of dataflow, and it is orthogonal to the previous TEDC and ABFT approaches.

The proposed optimization technique is based on the observation that the most common type of input pattern that causes timing errors is the input that can cause the sign bit flip of partial sum (PSUM). Hence, to minimize the timing error rate, an effective approach is to reduce the sign flip rate of PSUM.

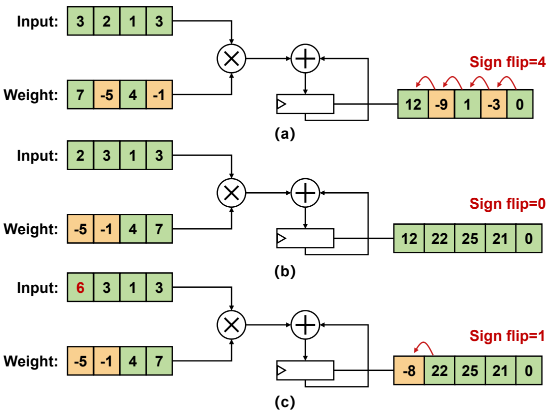

Fig. 1: A 1×4 convolution calculated in different orders. Reordering weights does not change the computing result, but avoids the critical input pattern of MAC.

As rectified linear unit (ReLU) is widely used in modern networks, the input activations of a convolution layer are often non-negative. Hence, the sign flip of PSUM is mainly determined by the sequence of weights for the computation. Fig. 1 gives a simple example to show how weight sequence impacts the sign flip. Based on the observations above, we propose the following heuristic solution: when computing an output activation, arrange the computation sequence so that all MACs with non-negative weights are computed first.

The heuristic solution proposed above can find the optimal sequence for the case of a single output channel. However, to improve throughput and data reuse, DNN accelerators often have more than one column to process multiple output channels simultaneously. To reduce the sign flip of simultaneously processed channels, we propose an input channel reordering algorithm to find the optimal computing sequence and a clustering algorithm that divides the weight matrix into submatrices to improve the reordering flexibility and achieves better reordering results.

The experimental results on VGG-16 and ResNet-18 demonstrate on average 7.8× timing error rate (TER) reduction and up to 37.9× TER reduction for certain layers, which enables the accelerator to maintain accuracy over a wide range of PVTA variations.

[⋆] Corresponding author: meng.li@pku.edu.cn

978-3-9819263-7-8/DATE23/© 2023 EDAA

References

[1] S. Tang *et al.*, “Robustart: Benchmarking robustness on architecture design and training techniques,” *arXiv preprint arXiv:2109.05211*, 2021.

[2] C. Liu, Z. Gao, S. Liu, X. Ning, H. Li, and X. Li, “Fault-tolerant deep learning: A hierarchical perspective,” *arXiv preprint arXiv:2204.01942*, 2022.

[3] P. H. Hochschild, P. Turner, J. C. Mogul, R. Govindaraju, P. Ranganathan, D. E. Culler, and A. Vahdat, “Cores that don’t count,” in *Proc. HotOS*, 2021, p. 9–16.

[4] H. D. Dixit, S. Pendharkar, M. Beadon, C. Mason, T. Chakravarthy, B. Muthiah, and S. Sankar, “Silent data corruptions at scale,” *arXiv preprint arXiv:2102.11245*, 2021.

[5] P. N. Whatmough, S. K. Lee, D. Brooks, and G.-Y. Wei, “Dnn engine: A 28-nm timing-error tolerant sparse deep neural network processor for iot applications,” *JSSC*, vol. 53, no. 9, pp. 2722–2731, 2018.

[6] 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 *DAC*, 2018.

[7] N. D. Gundi, T. Shabanian, P. Basu, P. Pandey, S. Roy, and K. Chakraborty, “Effort: A comprehensive technique to tackle timing violations and improve energy efficiency of near-threshold tensor processing units,” *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 29, no. 10, pp. 1790–1799, 2021.

[8] K. Zhao, S. Di, S. Li, X. Liang, Y. Zhai, J. Chen, K. Ouyang, F. Cappello, and Z. Chen, “Ft-cnn: Algorithm-based fault tolerance for convolutional neural networks,” *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, vol. 32, no. 7, pp. 1677–1689, 2021.

[9] D. Filippas, N. Margomenos, N. Mitianoudis, C. Nicopoulos, and G. Dimitrakopoulos, “Low-cost online convolution checksum checker,” *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 30, no. 2, pp. 201–212, 2022.