Silicon Photonic Neural Network Accelerators: Opportunities and Challenges

Mahdi Nikdast[1], Sudeep Pasricha[1], and Krishnendu Chakrabarty[2]

[1] *Electrical and Computer Engineering Department, Colorado State University, Fort Collins, CO 80523*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281*

[*] *mnikdast@colostate.edu*

Abstract:
Silicon photonic neural network accelerators (SPNNAs) offer chip-scale and light-speed computation and communication to boost AI inferencing and training performance. In this invited paper, we discuss some of the benefits and challenges of implementing SPNNAs.

1. Introduction

Recent years have seen a push towards domain-specific hardware deep-neural-network (DNN) accelerators with support for custom memory hierarchies, variable precision, and optimized matrix-vector multiply-and-accumulate (MAC) operations. Such DNN accelerators (e.g., Google's TPU and Amazon's Inferentia) have shown superior energy-efficiency (≈MAC/sec/Watt) and footprint efficiency (≈MAC/sec/mm^2) compared to GPUs for DNN inference tasks and training [1]. However, such accelerators cannot keep pace with the ever growing complexity of DNN applications with DNN sizes doubling almost every 3.4 months [2]. To address the limitations facing conventional electronic accelerators, researchers in both academia and industry are investigating the implementation of photonic neural network accelerators. By communicating and processing data in the optical domain, silicon photonic neural network accelerators (SPNNAs) offer the promise of providing very high footprint efficiencies in the hundreds of TeraMAC/sec/mm^2 with energy efficiencies of sub-femtoJoule/MAC [3]. Not only can SPNNAs address the fan-in and fan-out problems with linear algebra processors, their operational bandwidth can match that of the photodetection rate (typically ≈100 GHz), which is at least over an order of magnitude faster than electronic counterparts that are restricted to a clock rate of a few GHz [4]. Considering such benefits, there have been many implementations of coherent and noncoherent SPNNAs [4–12]. Nevertheless, there exist several roadblocks to the further advancement of SPNNAs and scaling them to satisfy the performance requirements of rapidly growing DNNs. In particular, SPNNA performance can be highly impacted by the optical losses and crosstalk noise accumulating when cascading photonic devices [13], susceptibility to fabrication-process variations and thermal crosstalk [14–17], and high cost and low performance of optical storage and optical nonlinear activations [18], just to name a few. In this invited paper, we present a brief overview of some of our prior work on the implementation of noncoherent SPNNA architectures, analysis of coherent SPNNAs under uncertainties, and design optimization techniques to improve robustness and power efficiency in SPNNAs.

2. High-Performance Noncoherent SPNNAs

When designing an SPNNA, it is important to co-design and co-optimize components at the device, circuit, architecture, and application level for more holistic optimization of the entire system. In [8], we showed the promise of hardware-software cross-layer co-design and co-optimization in SPNNAs. Our proposed SPNNA, called CrossLight, involves device-level engineering for resilience to fabrication-process variations and thermal crosstalk, circuit-level tuning enhancements for inference latency reduction, and an optimized architecture-level design that also integrates the device- and circuit-level improvements to enable higher resolution, better energy-efficiency, and improved throughput compared to prior efforts on photonic accelerator design. We demonstrated 9.5× lower energy-per-bit and 15.9× higher performance-per-watt compared to state-of-the-art photonic DNN accelerators. To further reduce SPNNA implementation overhead (e.g., for resource-constrained platforms), we can target simpler binary/ternary quantized models. Accordingly, we developed a novel optical-domain binarized-neural-network (BNN) accelerator, called ROBIN, in [9] and a sparse SPNNA, called SONIC, in [10]. Both ROBIN and SONIC showed considerable improvements in performance-per-watt and energy-per-bit compared to existing state-of-the-art electronic and photonic accelerators. In addition, we proposed an SPNNA to accelerate both homogeneously quantized and heterogeneously quantized convolutional-neural-network (CNN) models [11],

which have lower memory footprint and computational complexity. In [12], we proposed an SPNNA for accelerating any combination of simple recurrent neural networks (RNNs) and the newer RNN variants, including GRUs and LSTMs. As many of these applications are employed in real-time scenarios, SPNNAs can help efficiently accelerate RNN/LSTM/GRU inference. Our work in this area shows the importance of cross-layer co-design and co-optimization to realize high-performance, scalable, and robust SPNNAs.

3. SPNNA Design Challenges and Optimization

Despite the potential benefits of SPNNAs compared to their electronic counterparts, there are several challenges that must be addressed for further advancement and scaling of SPNNAs. For example, SPNNA performance is highly impacted by the intrinsic optical loss and crosstalk noise in underlying photonic devices and uncertainties due to fabrication-process and thermal variations. In [13], we presented a comprehensive analysis of the impact of optical loss and coherent crosstalk noise in coherent SPNNAs and showed considerable scalability constraints, drop in network accuracy (≈84% reported in [13]), and power penalty due to such inefficiencies. In [14], we studied the impact of nonuniform insertion loss across devices, quantization errors, and optical-phase and splitting-ratio noise in coherent SPNNAs, and showed up to 46% drop in the network accuracy due to such imperfections. We found that devices with higher adjusted phase settings (e.g., Mach–Zehnder interferometers in coherent SPNNAs) are more susceptible to uncertainties, based on which we proposed a method to minimize the required phase shifts, which in turn determine tuning power consumption, in coherent SPNNAs without impacting the network accuracy [19]. To improve SPNNAs' performance under uncertainties and reduce their power and area consumption, we proposed a hardware-aware pruning technique based on lottery ticket hypothesis [20] and magnitude-based pruning [21], both being able to prune SPNNAs by more than 85% without any loss in the inferencing accuracy. We also characterized coherent SPNNAs under correlated optical lithography imperfections (e.g., silicon-on-insulator thickness and etch-depth variations) in [22], and showed a significant drop in the network accuracy (to below 10%) due to such variations. Moreover, we proposed a design optimization solution based on using shallow-etched ridge waveguides in the design of underlying devices in coherent SPNNAs to improve their robustness under fabrication-process variations, and achieved on average a 50% increase in the network accuracy. Our work in this area shows the severe impact of uncertainties in SPNNAs and the critical need for low-cost design-time and run-time optimization solutions to realize robust SPNNAs under uncertainties.

References

1. Y. E. Wang et al. Benchmarking TPU, GPU, and CPU platforms for deep learning. *arXiv:1907.10701*, 2019.
2. D. Amodei and D. Hernandez. AI and compute, 2022.
3. M. A. Nahmias et al. Photonic multiply-accumulate operations for neural networks. *IEEE JSTQE*, 26(1):1–18, 2020.
4. F. Sunny et al. A survey on silicon photonics for deep learning. *ACM JETC*, 17(4), 2021.
5. X. Xu et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature*, 589(7840):44–51, 2021.
6. Y. Shen et al. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441–446, 2017.
7. F. Ashtiani et al. An on-chip photonic deep neural network for image classification. *Nature*, 606(7914):501–506, 2022.
8. F. Sunny et al. CrossLight: A cross-layer optimized silicon photonic neural network accelerator. In *DAC*, 2021.
9. F. Sunny et al. ROBIN: A robust optical binary neural network accelerator. *ACM TECS*, 20(5s):1–24, 2021.
10. F. Sunny et al. SONIC: A sparse neural network inference accelerator with silicon photonics for energy-efficient deep learning. In *ASP-DAC*, 2022.
11. F. Sunny et al. A silicon photonic accelerator for convolutional neural networks with heterogeneous quantization. In *GLSVLSI*, 2022.
12. F. Sunny et al. RecLight: A recurrent neural network accelerator with integrated silicon photonics. In *ISVLSI*, 2022.
13. A. Shafiee et al. LoCI: An analysis of the impact of optical loss and crosstalk noise in integrated silicon-photonic neural networks. In *GLSVLSI*, 2022.
14. S. Banerjee et al. Characterizing coherent integrated photonic neural networks under imperfections. *IEEE JLT*, 2022.
15. S. Banerjee et al. On the impact of uncertainties in silicon-photonic neural networks. *IEEE D&T*, 2022.
16. S. Banerjee et al. Towards functionally robust AI accelerators. In *IEEE MDTS*, pages 1–6, 2021.
17. S. Banerjee et al. Modeling silicon-photonic neural networks under uncertainties. In *DATE*, 2021.
18. C. Demirkiran et al. An electro-photonic system for accelerating deep neural networks. *arXiv:2109.01126*, 2021.
19. S. Banerjee et al. Optimizing coherent integrated photonic neural networks under random uncertainties. In *OFC*, 2021.
20. S. Banerjee et al. Pruning coherent integrated photonic neural networks using the lottery ticket hypothesis. In *IEEE ISVLSI*, 2022.
21. S. Banerjee et al. CHAMP: Coherent hardware-aware magnitude pruning of integrated photonic neural networks. In *OFC*, 2022.
22. A. Mirza et al. Characterization and optimization of coherent MZI-based nanophotonic neural networks under fabrication non-uniformity. *IEEE TNANO*, 2022.

The research summarized in this paper was supported by the National Science Foundation (NSF) under grants CCF-1813370, CCF-2006788, and CNS-2046226.