A Hardware-Efficient Silicon Electronic-Photonic Chip for Optical Structured Neural Networks

Shupeng Ning[a], Jiaqi Gu[a,b], Chenghao Feng[a], Rongxing Tang[a], Hanqing Zhu[a], David Z. Pan[a], and Ray T. Chen[a*]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX USA 78758; [b] School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ USA 85281.

ABSTRACT

Optical neural networks (ONNs) have gained significant attention as a promising neuromorphic framework due to their high parallelism, ultrahigh inference speeds, and low latency. However, the hardware implementation of ONN architectures has been limited by their high area overhead. These architectures have primarily focused on general matrix multiplication (GEMMs), resulting in unnecessarily large area costs and high control complexity. To address these challenges, we propose a hardware-efficient architecture for optical structured neural networks (OSNNs). Through experimental validation using an FPGA-based photonic-electronic testing platform, our neural chip demonstrates its effectiveness in on-chip convolution operations and image recognition tasks, which exhibits lower active component usage, reduced control complexity, and improved energy efficiency.

Keywords: optical neural network, integrated photonics, deep learning, hardware efficiency, artificial intelligence

1. INTRODUCTION

Machine learning with deep neural networks (DNNs) is exerting a growing influence across multiple aspects of our lives, encompassing applications such as image recognition, autonomous driving, and medical diagnosis. Furthermore, the emergence of large language models (LLM) has expanded the horizon of potential applications. With the continuous expansion of DNN model sizes and data volumes, there is an increasing demand for hardware accelerators capable of conducting high-speed, energy-efficient, and parallel multiply-accumulate (MAC) operations.[1] Among the various hardware accelerators designed for artificial intelligence (AI), integrated photonics is a promising technology as an efficient solution due to its ultra-high computational speed, low energy consumption, and high parallelism of light using unique multiplexing techniques.[2]

Leveraging integrated photonic platforms, the fundamental operations of DNNs—data transfers and matrix-vector multiplications (MVMs)—can be implemented through the combination of passive optical components and active optoelectronic devices. Specifically, optical signals can be modulated and reconfigured in accordance with the transmission characteristics of photonic integrated circuits (PICs). Based on this approach, a range of ONN tensor cores and architectures have been developed, including the Mach-Zehnder interferometer (MZI) mesh,[3] weight banks,[4] and crossbar arrays based on microring resonators (MRR).[5] However, a significant challenge associated with the deployment of optical tensor cores is their substantial hardware cost. For instance, the implementation of a fully-connected layer with n inputs and m outputs requires $O(m^2+n^2)$ or $O(m\times n)$ active devices for GEMMs. Besides, the electrical components for E-O modulation and AD/DA conversion consume a substantial portion of energy, thus compromising the overall power efficiency of optical computing units. Addressing the aforementioned challenges, this paper introduces a block-circulant optical neuron (BCON) with fewer active components and enhanced hardware efficiency tailored for OSNNs. We experimentally demonstrate on-chip convolution operations and high accuracy on image recognition tasks with fewer trainable optical components. These results highlight the effectiveness and efficiency of the proposed OSNN architecture, offering a promising solution to fully harness the potential of optical computing.

2. DESIGN AND WORKING MECHANISM

Block-circulant-based Structured neural network

Compared to general neural networks with arbitrary weight matrices, structured networks make trade-offs by pruning

* E-mail: chenrt@austin.utexas.edu

Optical Interconnects XXIV, edited by Ray T. Chen, Henning Schröder,
Proc. of SPIE Vol. 12892, 1289204 · © 2024 SPIE
0277-786X · doi: 10.1117/12.3000787

portions of matrix representability and reconfigurability to reduce the number of parameters. Related studies indicate that appropriate pruning does not degrade the model's performance significantly.[6] As a structured architecture, a circulant matrix is defined as a square matrix where all row vectors consist of identical elements, with each row vector being cyclically shifted one element relative to its preceding row. In this work, we propose a block-circulant-based network, wherein the original weight matrix is divided into multiple square sub-matrix blocks, each of which is a circulant matrix, as illustrated in Figure 1.a. For the hardware implementation of the OSNN, the computational and hardware complexities can be reduced during both the training and inference phases, owing to parameter sharing within each circulant block.

Design of electronic-photonic chip for block-circulant OSNN

Based on this approach, an optical tensor core for block-circulant OSNN is developed, as shown in Figure 1.b. The elements of weight vector $\boldsymbol{x}$ are modulated by a series of cascaded MRRs operating at different wavelengths, and then physically multiplying with the input vector $\boldsymbol{x}$ modulated by MZIs. After circulant-matrix-specific reconfiguration by an MRR-based crossbar array, the output signals could be detected by photoreactors. Leveraging WDM technology, on-chip photodetectors can autonomously execute the summation of weight elements, realizing the dot product operation. Compared to conventional ONN architectures, BCON necessitates only n cascaded MRRs for modulating to implement the $n\times n$ weight block, with a constant bias across the crossbar array. This configuration will significantly reduce the control complexity and the number of DACs used for modulation.

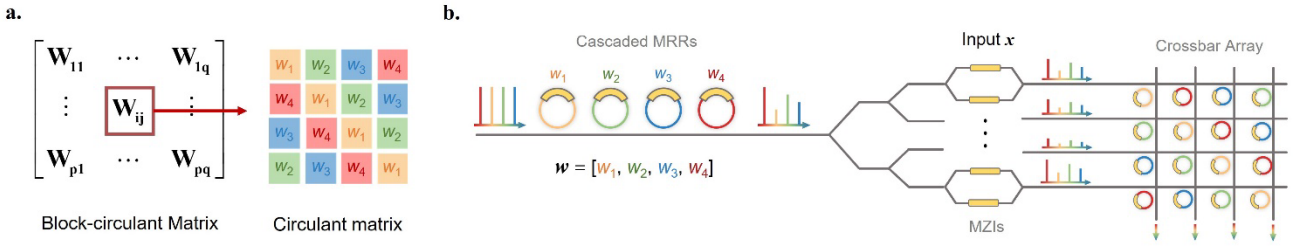

Figure 1 Schematics of block-circulant matrix and BCON. (a). The general structure of block-circulant matrix. (b). Schematic of BCON. Here, each MRR functions as a tunable filter, with varying colors indicating operation at distinct wavelengths.

Based on the schematic, we tape out a 4×4 BCON fabricated by *AIM Photonics*. The chip micrograph, including its electrical and optical packaging, is shown in Figure 2.a. The test flow of BCON is as follows: the large weight matrix is partitioned into our 4×4 circulant-matrix blocks, and we use an FPGA to program the DACs to encode weights and input signal. Furthermore, a customized PCB has been developed for modulation and data collection purposes. The output photocurrents are amplified by a trans-impedance amplifier (TIA) before being transmitted to FPGA-based systems for data processing and subsequent training on a PC (Figure 2.b). To model the no-idealities such as fabrication errors, programming errors, crosstalk and noise, we employ an AI-assisted, hardware-aware training framework for OSNN training, with details disclosed in our previous work.[7]

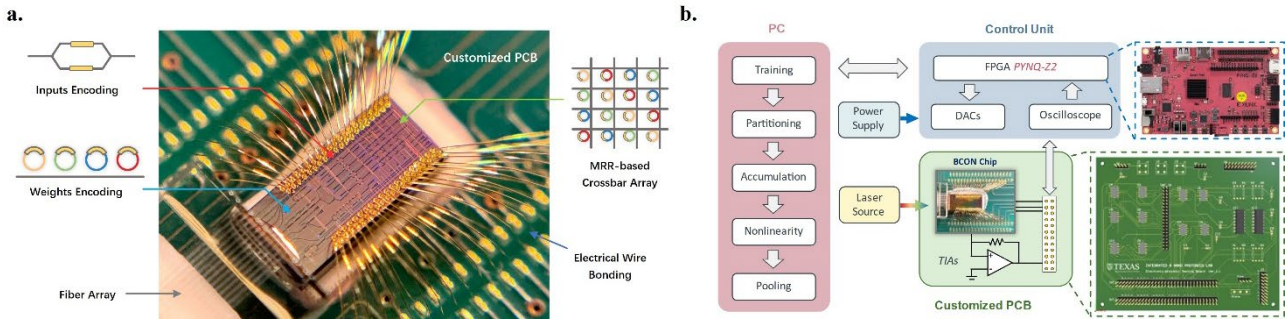

Figure 2 Experimental setup BCON. (a). Micrograph of the wire-bonded photonic chip packaged with fiber array. (b). Schematic of BCON test flow

3. RESULTS AND DISCUSSIONS

For the image convolution operations, we introduce a method similar to the *img2col* technique for conventional GPU acceleration. Specifically, the input image is divided into small patches corresponding to the kernel size and then reshaped

into a 2-D matrix, where each row vector from a single channel corresponds to a patch. Besides, convolution kernels are also reconstructed into a 2-D weight matrix, as shown in Figure 3.a. This approach transforms convolution operations into matrix multiplications, which is more efficient with high parallelism. For OSNN, we constrain the 2-D weight matrix after reconstruction to be a block-circulant matrix, and implement MVMs on BCON. To demonstrate system performance, we conducted on-chip convolution operations for chest X-ray images with a 3×3 Sobel Kernel designed for vertical edge detection (Figure 3.b). It is noteworthy that this system functions through an amplitude tuning mechanism, which poses challenges in achieving a full range of parameters. Hence, we address negative weights by executing two optical convolution processes and then subtracting the results in the electrical domain. The output waveforms and visualization results are shown in Figure 3.b, showcasing minimal noise and errors.

To further evaluation, we performed an image recognition task using the Fashion-MNIST dataset. The network architecture comprises two convolutional layers, an average pooling layer, and a linear classifier with 10 output classes (Figure 3.c). The confusion matrix from the trained network indicates an accuracy of 89.1% with 4-bit control precision for input encoding and 8-bit precision for weight encoding. This preliminary result demonstrates the capacity of the block-circulant ONN and BCON to effectively handle machine learning tasks. Future publications will explore implementations involving more sophisticated network architectures and datasets.

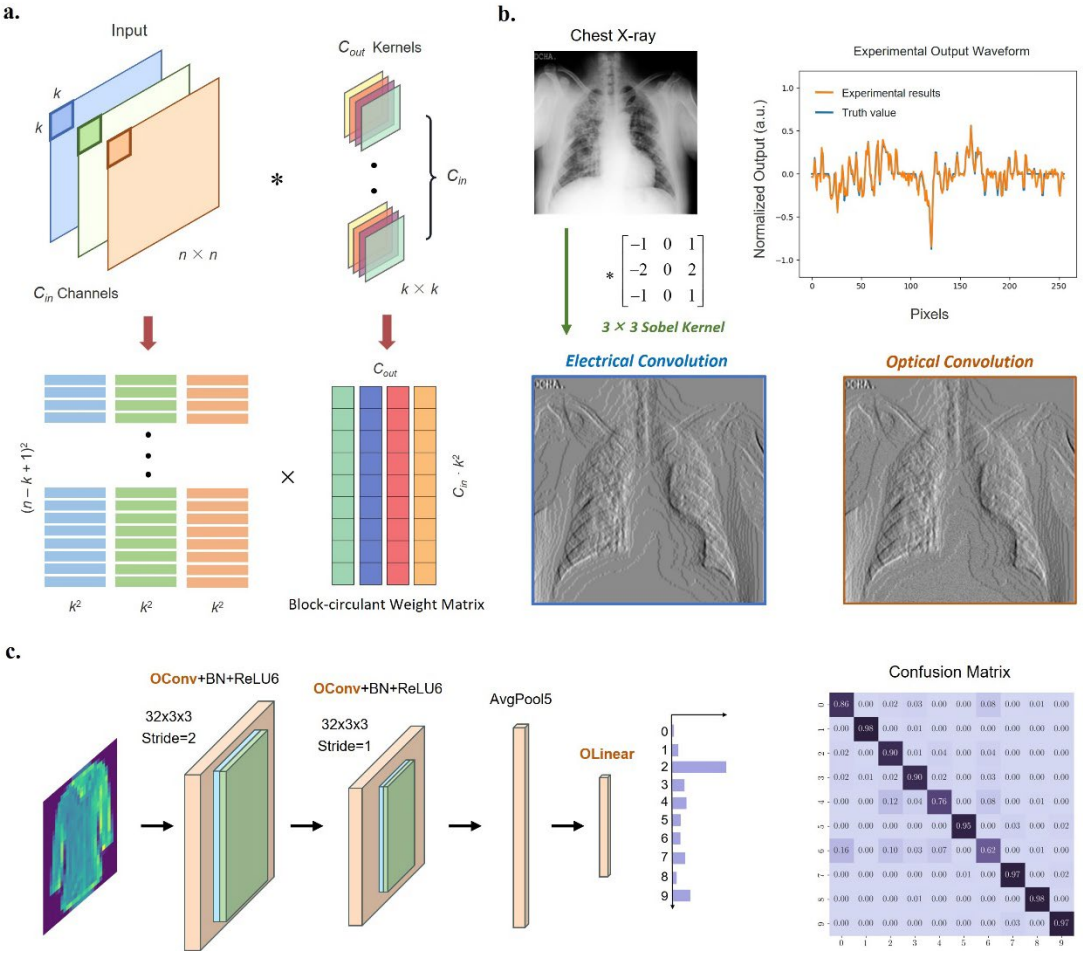

Figure 3 Experimental results of on-chip convolution and image recognition task. (a). 2-D reconstruction of input matrices and kernels for convolution operations. (b). Demonstration of on-chip convolution using BCON. Here, for demonstration purposes, we chose a convolution kernel with tangible physical meaning, corresponding to the first column vector in the 2-D weight matrix. This selection implies redundancy in the other columns (the block-circulant transformation of the first column vector) for this specific task. (c). Architecture of the OSNN and performance in the image recognition task using the FMNIST dataset.

4. CONCLUSION

This work reports a hardware-efficient ONN based on block-circulant structured architecture to implement tensor operations. The architecture features k times fewer active components than a conventional $k \times k$ crossbar array for GEMMs and fewer DACs leading to lower overall power consumption. Besides, an FPGA-based ONN testing platform and hardware-aware training framework are developed. The BCON experimentally demonstrated on-chip optical convolution operations and achieved a measured accuracy of 89.1% on the FMNIST dataset, which presents a compelling avenue for exploiting the advantages of optical computing in the development of next-generation AI accelerators.

REFERENCES

[1] Reuther, Albert, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner. "AI and ML accelerator survey and trends," pp. 1-10, IEEE High Performance Extreme Computing Conference (19 Sep 2022); https://doi.org/10.1109/HPEC55821.2022.9926331

[2] Feng, Chenghao, Shupeng Ning, Jiaqi Gu, Hanqing Zhu, David Z. Pan, and Ray T. Chen. "Integrated Photonics for Computing and Artificial Intelligence," pp. 1-2, IEEE Photonics Society Summer Topicals Meeting Series (17 Jul 2023); https://doi.org/10.1109/SUM57928.2023.10224461

[3] Shen, Yichen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun et al. "Deep learning with coherent nanophotonic circuits," 11, no. 7 (2017): 441-446, Nature photonics (12 Jun 2017); http://dx.doi.org/10.1038/nphoton.2017.93

[4] Tait, Alexander N., Thomas Ferreira De Lima, Ellen Zhou, Allie X. Wu, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal. "Neuromorphic photonic networks using silicon photonic weight banks," 7(1):7430, Scientific reports (7 Aug 2017); https://doi.org/10.1038/s41598-017-07754-z

[5] Ohno, Shuhei, Rui Tang, Kasidit Toprasertpong, Shinichi Takagi, and Mitsuru Takenaka. "Si microring resonator crossbar array for on-chip inference and training of the optical neural network," 9, no. 8: 2614-2622 ACS Photonics, (22 Jul 2022); https://doi.org/10.1021/acsphotonics.1c01777.

[6] Ding, Caiwen, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang et al. "Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices," pp. 395-408, Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (14 Oct 2017); https://dl.acm.org/doi/10.1145/3123939.3123952.

[7] Feng, Chenghao, Jiaqi Gu, Hanqing Zhu, Zhoufeng Ying, Zheng Zhao, David Z. Pan, and Ray T. Chen. "A Compact Butterfly-Style Silicon Photonic–Electronic Neural Chip for Hardware-Efficient Deep Learning," 9, no. 12 (2022): 3906-3916, ACS Photonics (30 Nov 2022); https://doi.org/10.1021/acsphotonics.2c01188