Realization of a Compact Photoelectric Platform for Optical Convolution Processing

Shupeng Ning,[1] Hanqing Zhu,[1] Chenghao Feng[1], Christian Uselton[1], Jiaqi Gu,[1,2], Rongxing Tang,[1], David Z. Pan[1] and Ray T. Chen[1,3,*]

[1] *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78758, USA*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA*
[3] *Omega Optics, Inc., 8500 Shoal Creek Blvd., Bldg. 4, Suite 200, Austin, TX 78757, USA*

[*] *chenrt@austin.utexas.edu*

Abstract: We presents a hardware-efficient optical computing architecture for structured neural networks (OSNNs). The performance of our neural chip was validated on a photonic-electronic testing platform experimentally, demonstrating reduced optical component utilization and small deviation.

1. Introduction

Deep neural networks (DNNs) have consistently exhibited exceptional performance across a diverse range of intelligent tasks. With the rapid expansion of DNN model sizes and data volumes, the demand for hardware accelerators capable of executing high-speed, energy-efficient, and parallel multiply-accumulate (MAC) operations is steadily increasing. Among the various technologies under exploration, optical neural networks (ONNs) have emerged as a promising solution for accelerating artificial intelligence tasks, owing to their low latency, broad bandwidth, and the inherent parallelism of light. [1, 2]

In addition to advancements in hardware implementation, recent progress in neural architecture design and network compression techniques has substantially reduced computational complexity. One approach involves the utilization of structured neural networks (SNNs), which efficiently lower computational requirements while also enhancing hardware implementation and scalability. [3, 4] In this study, we introduces an innovative photonics circuit, illustrated by a convolutional neural network with block-circulant weight matrices, and provide validation of the effectiveness of our neural chip based on a FPGA-based photoelectric testing platform.

2. Design and Working Mechanism

2.1. Block-circulant-based structured network

Structured networks boost hardware efficiency via pruning, with fine-grained pruning effectively balancing accuracy and compression. Specifically, larger block sizes yield higher compression ratios but may compromise accuracy, while smaller blocks maintain better accuracy at the cost of increased hardware resource usage. [5]

The core concept of block-circulant-based network is to partition the original $M \times N$ weight matrix into $P \times Q$ blocks of sub-matrices, and each sub-matrix is a circulant matrix with size of $k \times k$ (Fig. 1.a). Correspondingly, the input $\mathbf{x}$ is also partitioned as $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_Q^T]^T$. During the inference phase, the computational and storage complexities of this architecture are $O(PQ \cdot k \log k)$ and $O(PQk)$, respectively.

2.2. Design of silicon photoelectric chip

In this work, a circulant matrix with $k = 4$ is implemented on a silicon electronic-photonic chip, incorporating thermal-optic Mach-Zehnder interferometers (MZIs) and microring resonators (MRRs) as shown in Fig. 1.b. The elements of weight vector w_k are modulated by 4 MRRs at different wavelengths, and then, after multiplying with the input vector $\mathbf{x}_j$ modulated by 4 MZIs. The partial outputs from each column are subsequently converged at four photodetectors through a 4×4 MRR-based switch array, aligning with the block-circulant structure.

3. Results and Discussions

The computing performance of the platform is demonstrated and evaluated through the edge detection of handwritten digits with 28×28 pixels and 4-bit resolution from the MNIST dataset (Fig. 2.a). n images are first flattened into an $4 \times (n \times 27^2)$ vector using the *im2col* transform, where 4 and 27 correspond to the size of the kernel and output image, respectively. Subsequently, this input vector is transformed into a serial electrical waveform through a digital-to-analog converter (DAC) controlled by FPGA. The analog waveform is fed into MZIs to modulate the