Frontiers in Optics + Laser Science 2023 (FiO/LS) © Optica Publishing Group 2023

High-Speed Multidimensional Optical Computing

Alireza Fardoost[1], Fatemeh Ghaedi Vanani[1], Zheyuan Zhu[1], Christopher Doerr[2], Shuo Pang[1], Guifang Li[1]
[1]CREOL, The College of Optics and Photonics, University of Central Florida, Orlando, FL 32816
[2]Aloe semiconductor, 1715 Highway 35, Suit 304 Middletown, NJ 07748
a.fardoost@knights.ucf.edu, Li@ucf.edu

Abstract: A coherent multi-dimensional photonic tensor accelerator performing high-speed matrix-matrix multiplication is proposed and demonstrated. A pattern recognition experiment is demonstrated at a 25Gbps modulation speed exploiting orthogonal dimensions of light including time, wavelength, and spatial mode. © 2023 The Author(s)

1. Introduction

Efficiently performing matrix multiplication is a crucial aspect of hardware accelerators designed to support artificial neural networks (ANNs). So far electronic industries have incorporated parallel computing structures and optimized memory organizations such as graphic processing unit (GPU) and tensor processing unit (TPU) to perform larger amounts of computations. However, the scalability and power consumption of electronic matrix accelerators has become significant challenges due to the limited number of parallelizable degrees of freedom in the two-dimensional plane [1-3]. Here we introduce a photonic tensor accelerator (PTA) that utilizes multidimensional encoding that enables performing matrix multiplication with a single memory access [4]. Utilizing all degrees of freedom of light, PTA holds the potential to deliver significantly higher computing power and energy efficiency compared to both state-of-the-art electronic and photonic accelerators [5, 6]. In addition, reliable communication technologies in modulation, multiplexing, and detection will support computations at 10s of GHz speed with optimized energy efficiency. Here, we experimentally demonstrate a 2×2 matrix-matrix multiplier in free space operating at 25 GBd employing two wavelengths and two spatial modes. As a major application, a pattern recognition scheme based on cross-correlation calculation is presented and the results show a pattern recognition accuracy of 99.99998%.

2. Photonic Tensor Accelerator (PTA)

In the proposed PTA, scalar multiplication is performed via interference and coherent detection. To extend the scalar multiplication to the inner product of two vectors, the weight and input vector elements can be mapped to different wavelengths or spatial modes and, as a result, the BPD output will be $\sum_{n=1}^{N} w_n \times x_n$ where N is the length of the vectors. Combining WDM and MDM with parallelization in space enables matrix-matrix multiplication in a single clock cycle with single memory access as shown schematically in Fig. 1 (a). As shown here, the W matrix is projected on the (mode, space) dimensions and duplicated in the wavelength dimension. Similarly, the X matrix is projected on the (mode, wavelength) dimensions and duplicated in space. The output matrix elements are mapped to different wavelengths and spatial locations. In Fig. 1(b), a free-space implementation of the PTA performing 2×2 matrix-matrix multiplication is illustrated where two modulators operate at 25Gbps and optical delays are used to emulate other elements of the weight and input matrix.

2. PTA for Pattern Recognition

Pattern recognition is a major research topic in image processing. Correlation pattern recognition is based on computing the correlation between an object $O_{I\times J}$ and an image $D_{I\times J}$, $\Gamma_{OD} = \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} D_{ij}$ where I and J are the number of pixels in x and y dimensions, respectively.

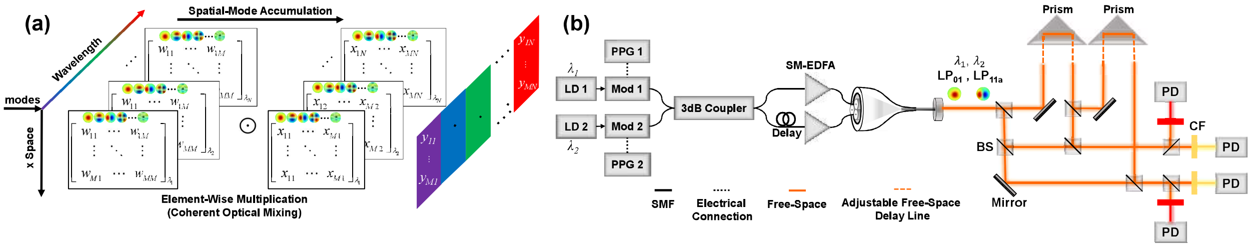

Fig. 1. Photonic Tensor Accelerator (PTA). (a) Schematic mapping of the matrix elements on wavelengths, spatial modes, and space in a matrix-matrix multiplier. (b) the experimental setup for the matrix-matrix multiplication demonstration. LD: Laser Diode, PPG: Pulse Pattern Generator, BS: Beam Splitter, CF: Color Filter, PD: Photodetector.

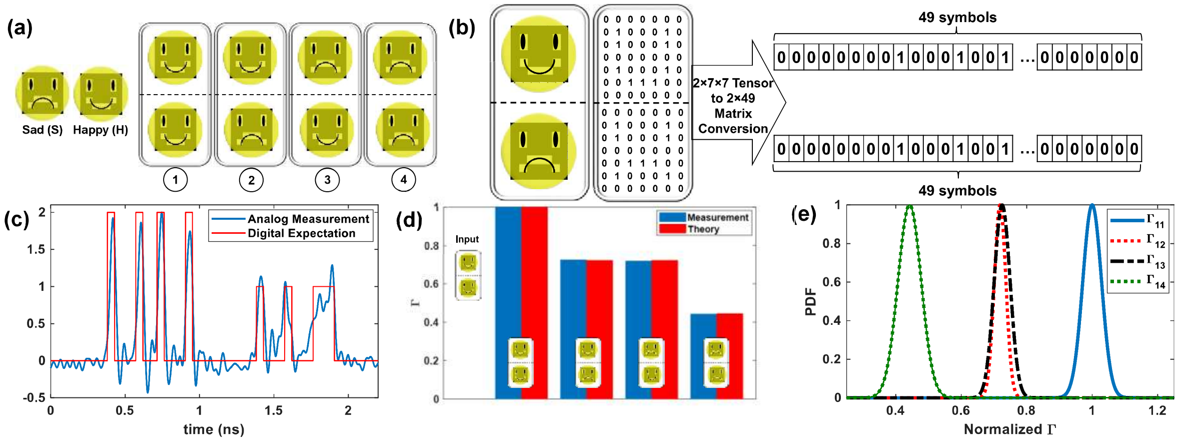

Fig. 2: Pattern recognition experiment and results. (a) Two emojis with happy and sad faces can generate a library of 4 sets of combo cards. (b) A 2×7×7 combo card tensor is reformatted to a 2×49 matrix to be sent as a data on modulators. (c) Measured and Expected output symbols corresponding to the correlation between the pixels of two (H, H) and (H, S) combo cards. (d) Well-match experimental and theoretical normalized average Γ are shown. (e) PDF of each overlap based on 100 measurements to illustrate the accuracy of the pattern recognition process through PTA.

Consider an emoji library shown in Fig. 2(a) consisting of four combinations of the "Happy (H)" and the "Sad (S)" emojis. Given an arbitrary input combo card, we find a match in the library by computing correlations utilizing 2×2 matrix multiplication employing two wavelengths, two spatial modes, and four PDs of the PTA in Fig. 1(b). The matching combo card from the library is the one having the highest Γ_{OD} with the input combo card. Fig. 2(b) illustrates, for the example of a (H, S) combo card, how emojis are pixelated and reformatted from a 2×7×7 tensor into a 2×49 matrix. Accordingly, the modulators in Fig. 1(b) together with the fiber and free-space delays are designed so that each PD receives the top and the bottom emojis of the two cards in the fundamental spatial mode and the second spatial mode, respectively. As a result, the photocurrents produced by the PDs are proportional to the correlations of the input combo card with each of the four combo cards in the library, respectively, one pixel at a time at a symbol/pixel rate of 25 GBd. As an example, the 49 analog output symbols corresponding to the correlation between the pixels of two (H, H) and (H, S) combo cards are depicted in Fig. 2(c) and are compared with their expectations. The overall correlation Γ between the two combo cards is obtained by summing over 49 symbols in the time domain. The time integral decouples the speed of ADC from DACs and consequently saves energy [7].

For the input combo card (H, H), its correlation with each card in the library is repeated experimentally 100 times. The averaged auto- and cross-correlations are plotted in Fig. 2(d), illustrating that experimental measurements and theoretical expectations are well-matched. In the presence of noises, coherent interference, spatial mode mismatch, and other vibrational effects, the experimentally measured correlation values are within 0.5% of theoretical expectations. Furthermore, the probability distribution functions (PDFs) of the four correlations are presented in Fig. 2(e). Using Gaussian estimation, the pattern recognition accuracy is found to be 99.99998%.

3. Conclusion

We demonstrate a PTA that exploits all dimensions of light to perform energy-efficient and high-speed matrix processing. We have illustrated that using only two spatial modes and two wavelengths, the PTA can perform pattern recognition. Employing a larger number of DoFs, more complex images can be processed, and sophisticated tasks such as classification or language processing can be performed.

Funding

This work was supported in part by the Office of Naval Research under contract N00014-20-1-2441, NSF under grant number ECCS-1932858, and the Army Research Office under contract W911NF2110321.

References

1. Manavski, S.A. *CUDA compatible GPU as an efficient hardware accelerator for AES cryptography*. in *2007 IEEE International Conference on Signal Processing and Communications*. 2007. IEEE.
2. Yao, P., et al., *Fully hardware-implemented memristor convolutional neural network.* Nature, 2020. **577**(7792): p. 641-646.
3. Horowitz, M. *1.1 computing's energy problem (and what we can do about it)*. in *ISSCC* 2014. IEEE.
4. Fardoost, A., et al. *Vector-mode Multiplexing For Photonic Tensor Accelerator*. in *2022 27th OECC and 2022 PSC*. 2022. IEEE.
5. Shen, Y., et al., *Deep learning with coherent nanophotonic circuits.* Nature photonics, 2017. **11**(7): p. 441-446.
6. Hamerly, R., et al., *Large-scale optical neural networks based on photoelectric multiplication.* Physical Review X, 2019. **9**(2): p. 021032.
7. Youngblood, N., *Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication.* IEEE Journal of Selected Topics in Quantum Electronics, 2022.