Wavelength-Space Domain High-Throughput Artificial Neural Networks by Parallel Photoelectric Matrix Multiplier

Mehmet Berkay On, Hongbo Lu, Humphry Chen, Roberto Proietti and S. J. Ben Yoo
Department of Electrical and Computer Engineering, University of California, Davis, One Shields Ave., Davis, California 95616 USA
sbyoo@ucdavis.edu

Abstract: We propose a massively parallel neural network architecture with photonic matrix-vector multiplication in the wavelength and space domains with balanced photodetectors and nonlinear transfer functions in MZI modulators. An experimental proof-of-principle demonstration is also discussed. © 2020 The Authors

OSIC Code: (200.0200) Optics in computing, (200.4260) Neural Networks, (200.4700) Optical neural systems, (250.0250) Optoelectronics, (250.5300) Photonic integrated circuits

1. Introduction

Machine learning has become an integral part of our daily lives, however, its scalability, energy-efficiency, and throughputs are limited by electronic processing. For instance, even the state-of-the-art GPU systems with high-bandwidth-memories (HBM2) integrated on interposers and consuming 250 Watts [1], are not expected to meet the real-time pattern and image classification requirements of future autonomous vehicles equipped with LiDARs due to their limited throughput. On the other hand, typical machine learning systems spend more than 90% of the energy and runtime on matrix multiplication. For instance AlexNet [2] is an 8-layer network with five convolutional layers and three fully connected layers of artificial neurons [3], and ResNet-50 has 152 or even 1202 convolutional layers. AlexNet nearly equals and ResNet outperforms a human being on certain image recognition tasks.

Photonic neural networks are attractive because they support high-throughput and matrix multiplication by imposing matrix weight values onto passive optical components such as Mach-Zehnder interferometers [4, 5], micro-resonators [6], and couplers without limitations due to electrical impedance or crosstalk. However, the scalability of such systems are limited by the number of the passive optical components that get cascaded in series in $O(N^2)$. A recent demonstration by Hamerly et al [7] proposed a matrix multiplier that overcomes this scalability problem by introducing time division multiplexed processing with multi-spatial ports. However, this solution requires accurate time synchronization and high-speed A/D and D/A converters. This paper proposes a new wavelength-space domain parallel processing of high-dimensional matrix multiplications that can be updated on a frame-by-frame basis in the time domain. We further propose an artificial neural network (ANN) consisting of the photonic matrix multiplier utilizing balanced detectors and Mach-Zehnder modulators for implementing a nonlinear transfer function. The benefit of scalability and throughput compounds greatly by utilizing spatial and wavelength parallelism on a photonic integrated circuit, and with rapid updates on a frame-by-frame basis.

2. Photonic Matrix Multiplication in Space-Wavelength domains by AWG-MMI coupler pairs

Consider the photonic neural network in Figure 1 and assume that the k-th layer of N neurons are connected to M neurons at the $(k+1)$-th layer. Required calculations for that structure is: $\boldsymbol{x}^{\boldsymbol{k+1}}_{\boldsymbol{M\times 1}} = f_{NL}(\boldsymbol{y}^{\boldsymbol{k}}_{\boldsymbol{M\times 1}})$, and $\boldsymbol{y}^{\boldsymbol{k}}_{\boldsymbol{M\times 1}} = \boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{M\times N}}\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N\times 1}}$where $f_{NL}(*)$ is a nonlinear activation function applied to each element of $\boldsymbol{y}^{\boldsymbol{k}}_{\boldsymbol{1\times M}}$. Matrix $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{M\times N}}$ can be represented as M number of row vectors; $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{M\times N}} = [\boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{1}} \quad \boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{2}} \quad \dots \quad \boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{M}}]$.

Hence, the m-th element of the output vector at k-th layer and the m-th element of input vector to $(k+1)$-th layer become

$y^k_m = \boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{m}}\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N\times 1}} = \sum_{j=1}^{N} W^k_{m\,\lambda_j}\, x^k_{\lambda_j}$ and $x^{k+1}_m = f_{NL}(y^k_m)$, respectively. Each entry of $\boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{m}}$ and $\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N\times 1}}$ are coded in the amplitude of wavelength λ_j. Here, we have utilized wavelength domain vectors instead of time domain vectors used in Reference [6] to offer N degrees of parallelism and speed up. A single unit of AWG-MMI coupler Balanced detector can successfully perform matrix multiplication by using wavelength division multiplexing and coherent homodyne detection scheme. Each entry of weight vector, W^m_j and x_j are amplitude modulated at the same wavelength λ_j. Figure 1 shows $(M+1)*N$ Mach-Zehnder based intensity modulators, N for input vector $\boldsymbol{x}^{\boldsymbol{k}}_{\boldsymbol{N\times 1}}$and fanout for each M spatial location to perform parallel M vector-by-vector multiplication. Remaining $M*N$ modulators encode each entry of $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{M\times N}}$matrix. Then the E fields at the PD1 and PD2 (${E_{PD1}}^k_m$, ${E_{PD2}}^k_m$) and photocurrent output (y^k_m) are given as:

$$E_{PD1_m}^{\ k} = \frac{1}{2}\left(\sum_{j=1}^{N} W_{m\,\lambda_j}^{k} + \sum_{j=1}^{N} x_{\lambda_j}^{k}\right) \qquad E_{PD2_m}^{\ k} = \frac{1}{2}\left(\sum_{j=1}^{N} W_{m\,\lambda_j}^{k} - \sum_{j=1}^{N} x_{\lambda_j}^{k}\right)$$

$$\begin{aligned} y_m^k &\propto \left|E_{PD1_m}^{\ k}\right|^2 - \left|E_{PD2_m}^{\ k}\right|^2 \\ &\propto \left(\sum_{j=1}^{N}\left|W_{m\,\lambda_j}^{k}\right|^2 + \sum_{j=1}^{N}\left|x_{\lambda_j}^{k}\right|^2 + 2Re\left(\sum_{j=1}^{N}\sum_{j'=1}^{N} W_{m\,\lambda_j}^{k} x_{\lambda_{j'}}^{k*}\right)\right) \\ &\quad - \left(\sum_{j=1}^{N}\left|W_{m\,\lambda_j}^{k}\right|^2 + \sum_{j=1}^{N}\left|x_{\lambda_j}^{k}\right|^2 - 2Re\left(\sum_{j=1}^{N}\sum_{j'=1}^{N} W_{m\,\lambda_j}^{k} x_{\lambda_{j'}}^{k*}\right)\right) \propto \left(Re\left(\sum_{j=1}^{N} W_{m\,\lambda_j}^{k} x_{\lambda_j}^{k*}\right)\right) \end{aligned}$$

where we assumed that the wavelength channel spacing is far beyond the balanced detector detection bandwidth at the final step above.

Transimpedance amplifiers after the balanced detection drive the Mach-Zehnder modulators with these electrical signals y_m^k and will produce $x_m^{k+1} = f_{NL}(y_m^k)$ for the k+1 th layer input as desired.

Spatially located M vector-to-vector multipliers can perform matrix-vector multiplication at each k-th frame. Frame-by-Frame update of matrix multiplication is possible at each time window, T which can be determined by the modulator speed, detector electrical bandwidth, and D/A converter speed used to input the weight matrix values.

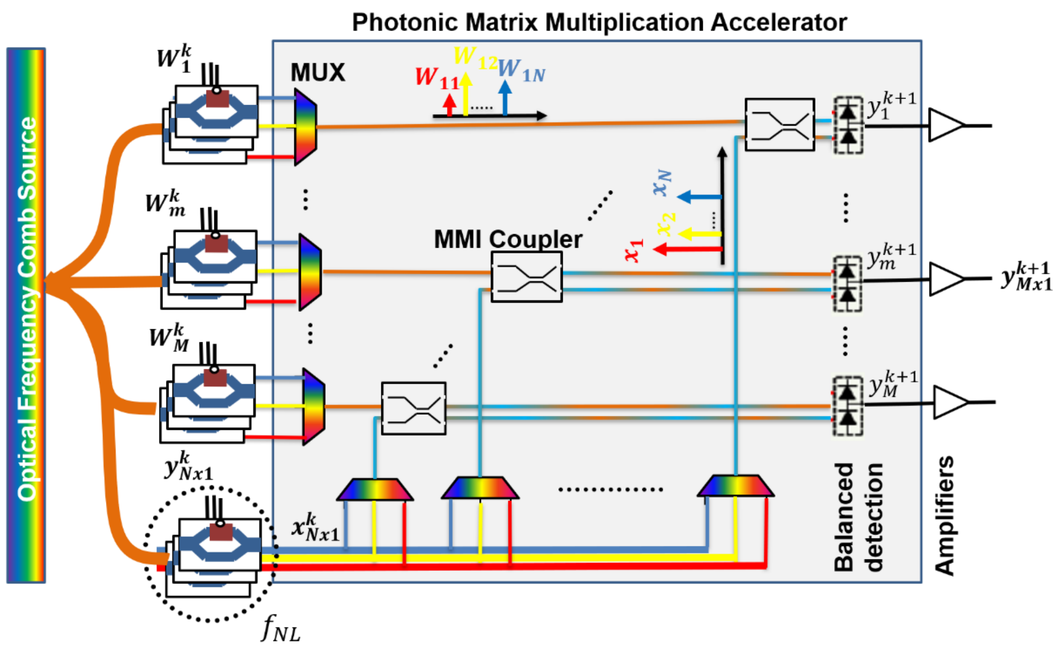


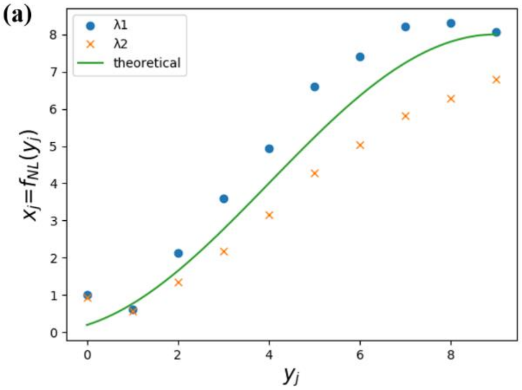
Figure 1. Cascadable analog feed-forward artificial neural network structure with photonic matrix-vector multiplier circuit and Mach-Zehnder modulator nonlinearity.

Figure 1 shows one-layer of the proposed artificial neural network scheme where the optical frequency comb source provides N number of wavelengths. A integrated device such as the SPIDER photonic integrated circuit demonstrated in [8] can achieve $12x18$ matrix by $18x1$ vector multiplication for the proposed matrix multiplication method introduced here.

3. Experimental Results & Discussions

As a proof of principle demonstration, we conducted experiments using the scheme in Figure 1 for two wavelengths (N=2) and two vectors, $\boldsymbol{W_{1x2}}$ and $\boldsymbol{x_{2x1}}$. In this simple case, the wavelength multiplexing operations of the arrayed waveguide gratings (AWG) are performed by 2:2 interferometric coupler. As fibers and bulk interferometric couplers introduce phase fluctuation and distort coherency between multiplied entries of vectors, we overcame this problem in the experimental setup by using distinct lasers with $\Delta\lambda$ spacing (the frequency comb source of Figure 1 was implemented by four lasers at $\lambda_1, \lambda_1 + \Delta\lambda, \lambda_2, \lambda_2 + \Delta\lambda$). In a future system implementation of Figure 1, a single tone optical frequency shifter [9] driven by an RF frequency at Δf can be inserted in tandem to the modulators for $\boldsymbol{y}_{\boldsymbol{Nx1}}^{\boldsymbol{k}}$ shown in Figure 1.

(a)



(b)

$\boldsymbol{W_{1x2}}$ \ $\boldsymbol{x_{2x1}}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$
[1 1]	5.300mV	2.700mV	2.825mV	0.375mV
[1 0]	2.750mV	2.475mV	0.275mV	0.150mV
[0 1]	2.925mV	0.200mV	2.750mV	0.100mV
[0 0]	0.200mV	0.100mV	0.050mV	-0.075mV

Figure 2. (a) Experimental measurement of modulator nonlinearity, (b) Experimental binary multiplication results with detected voltage values.

The proof-of-principle experiment included detection and recording of envelope waveform amplitudes on an oscilloscope. In the real system, this can be replaced by an RF amplitude detection.

Figure 2 (a) shows experimentally tested nonlinear function of the Mach Zehnder intensity modulator. By applying DC voltages on modulators, we amplitude modulated the generated wavelength pulses. The highest power output was recorded as ‘1’ while the lowest power as ‘0’. Figure 2 (b) shows the detected voltage levels representing the results of the multiplication operation between the two vectors (note that lasers power outputs and the modulators losses must be equalized to avoid a bias in the multiplication results in this proof-of-principle experiment).

4. Conclusion

In this work, an analog feed-forward artificial neural network structure with photonic matrix-vector multiplier circuit and Mach-Zehnder modulator nonlinearity is proposed and demonstrated with a proof-of-principle binary 2-by-1 vector dot-product experiment. Future studies will aim at simulating and verifying that performance of this modulator-based activation function and this matrix-vector multiplication scheme can achieve accurate results on ANN tasks.

5. References

[1] “Introducing NVIDIA TITAN V: The World's Most Powerful PC Graphics Card,” *NVIDIA*. [Online]. Available: https://www.nvidia.com/en-au/titan/titan-v/.

[2] Krizhevsky, A., Sutskever, I., and Hinton, G.E.: ‘Imagenet classification with deep convolutional neural networks’, in ‘Book Imagenet classification with deep convolutional neural networks’ (2012, edn.), pp. 1097-1105

[3] Pineda, F.J.: ‘Generalization of back-propagation to recurrent neural networks’, Physical Review Letters, 1987, 59, (19), pp. 2229-2232

[4] Hughes, T.W., Minkov, M., Shi, Y., and Fan, S.: ‘Training of photonic neural networks through in situ backpropagation and gradient measurement’, Optica, 2018, 5, (7), pp. 864-871

[5] Miller, D.A.: ‘Self-aligning universal beam coupler’, Optics express, 2013, 21, (5), pp. 6360-6370

[6] Tait, A.N., Zhou, E., de Lima, T.F., Wu, A.X., Nahmias, M.A., Shastri, B.J., and Prucnal, P.R.: ‘Neuromorphic Silicon Photonics’, arXiv preprint arXiv:1611.02272, 2016

[7] Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M., and Englund, D.: ‘Large-Scale Optical Neural Networks Based on Photoelectric Multiplication’, Physical Review X, 2019, 9, (2), pp. 021032

[8] T. Su, G. Liu, K. E. Badham, S. T. Thurman, R. L. Kendrick, A. Duncan, D. Wuchenich, C. Ogden, G. Chriqui, S. Feng, J. Chun, W. Lai, and S. J. B. Yoo, “Interferometric imaging using Si3N4 photonic integrated circuits for a SPIDER imager,” *Optics Express*, vol. 26, no. 10, p. 12801, Mar. 2018.

[9] Qin, C., Lu, H., Ercan, B., Li, S., and Yoo, S.J.B.: ‘Single-Tone Optical Frequency Shifting and Nonmagnetic Optical Isolation by Electro-Optical Emulation of a Rotating Half-Wave Plate in a Traveling-Wave Lithium Niobate Waveguide’, IEEE Photonics Journal, 2017, 9, (3), pp. 1-13

This work was supported in part by AFOSR grant FA9550-181-1-0186.