European Conference on Optical Communication © Optica Publishing Group 2022

Silicon Photonics for Machine Learning: Training and Inference

B. J. Shastri[(1)(2)], M. J. Filipovich[(1)], Z. Guo[(1)], P. R. Prucnal[(2)], C. Huang[(2)], A. N. Tait[(1)], S. Shekhar[(3)], and V. J. Sorger[(4)]

[(1)] Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada, shastri@ieee.org
[(2)] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
[(3)] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[(4)] Department of Electrical and Computer Engineering, George Washington University, Washington, DC V6T 1Z4, USA

Abstract *Photonics neural networks employ optical device physics for neuron models, and optical interconnects for distributed, parallel, and analog processing for high-bandwidth, low-latency, and low-switching energy applications in AI and neuromorphic computing. We discuss silicon photonics for machine learning acceleration for inference and in situ training.*

Advancements in machine learning (ML) and artificial intelligence (AI) technologies have enabled numerous applications, including sophisticated recommendation models, natural language processing, computer vision, augmented reality, and so on [1], [2]. The heavy dependence of ML algorithms training on large data sets has enabled The groundbreaking progress of these AI applications in different fields. The interconnection of neurons in artificial neural networks (ANNs) can be described by a matrix, with the processed data represented as a vector. Training on large data sets with deep neural networks results in large-scale dense matrix-vector multiplications. The improvement in the performance (i.e., accuracy) of many ML applications comes at the cost of higher computational power requirements [3]. There has been significant progress in the development of digital electronic application-specific integrated circuits (ASICs) known as AI accelerators that are dedicated to dense matrix computations [4], [5]. However, modern AI accelerators have seen two significant bottlenecks in energy efficiency: data transfer to and from memory and large matrix-vector multiplications. Both have imposed strict physical limitations on the scalability and performance of digital electronic AI accelerators.

Integrated photonic processors enabled by silicon photonics have shown promising capabilities in accelerating tensor (i.e., multidimensional vector and matrix) operations [6]–[9] by exploiting the high bandwidth of photonic devices (modulators and photodetectors), low latency, and minimal energy-delay product due to passive optical waveguides [10]. Some of these processors [7]–[9] are scalable and use the parallel nature of light through wavelength-division multiplexing (WDM) to achieve large-scale interconnects and massively parallel data processing and transfer. Recent developments have shown that the wavelength-multiplexed silicon photonic platforms operate with up to 7-bit precision [11] and, most recently, 9-bit precision [12] on each multiplication unit. However, recent studies in these photonic processors have also seen an increasing demand for a rigorous photonic programming scheme to facilitate efficient communication between photonic hardware and its control system [6], [7], [10], [13].

Over the ten years, several photonic neural networks [10] [14] approaches have been proposed. This can be divided into feedforward and recurrent (including random recurrent, i.e., reservoir computing [15]–[17]), or coherent (single wavelength) [6], [18] and multiwavelength [7], [9], [19]–[22] approaches, or continuous-time networks and spiking networks, or integrated approaches and free-space. In this talk, we will briefly highlight some of these.

An area of machine learning that would benefit from the low power consumption and high information processing bandwidth enabled by photonics is the training of large neural networks. Several photonic architectures have been proposed for executing in-memory computation of neural network inference [6], [7], [19]. However, for the neural network to perform a practical task, the optimal network parameters (weights and biases) must first be determined using deep learning training algorithms. These algorithms have high computation and memory costs that challenge the current hardware platforms executing them [23]. The substantial energy required to train large neural networks using standard von Neumann architectures presents a high financial and environmental cost

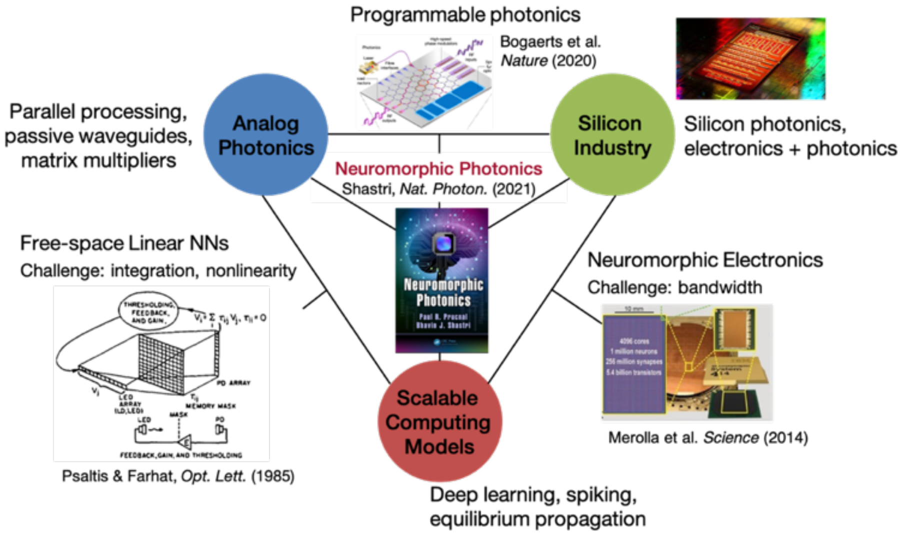
Programmable photonics
Bogaerts et al. *Nature* (2020)
Parallel processing, passive waveguides, matrix multipliers
Analog Photonics
Silicon Industry
Silicon photonics, electronics + photonics
Neuromorphic Photonics
Shastri, *Nat. Photon.* (2021)
Free-space Linear NNs
Challenge: integration, nonlinearity
Neuromorphic Electronics
Challenge: bandwidth
Neuromorphic Photonics
Paul R. Prucnal
Bhavin J. Shastri
THRESHOLDING, FEEDBACK, AND GAIN
LED ARRAY
PD ARRAY
MEMORY MASK
FEEDBACK, GAIN, AND THRESHOLDING
Scalable Computing Models
Psaltis & Farhat, *Opt. Lett.* (1985)
Merolla et al. *Science* (2014)
Deep learning, spiking, equilibrium propagation

Fig. 1: The advent of neuromorphic photonics is due to the convergence of recent advances in photonic integration technology, the resurgence of scalable computing models (e.g., spiking, deep neural networks), and a large-scale silicon industrial ecosystem.

[24].

The recently proposed direct feedback alignment (DFA) supervised learning algorithm [25] has gathered interest as a bio-plausible alternative to the popular backpropagation training algorithm [26]. The DFA algorithm is a supervised learning algorithm that propagates the error through fixed random feedback connections directly from the output layer to the hidden layers during the backward pass [26]. Unlike backpropagation, the DFA algorithm does not require the network layers to be updated sequentially during the backward pass, enabling the algorithm to be a suitable candidate for efficient parallelization using photonics. The training algorithm has been used to train neural networks using the MNIST, CIFAR-10, and CIFAR-100 datasets and yields comparable performance to backpropagation [26]. The DFA algorithm has also been shown to obtain performances comparable to fine-tuned backpropagation in applications requiring state-of-the-art deep learning networks, including natural language processing and neural view synthesis [27]. A recent theory suggests that training shallow networks with the DFA algorithm occurs in two steps: the first step is an alignment phase where the weights are modified to align the approximate gradient with the actual gradient of the loss function, which is followed by a memorization phase where the model focuses on fitting the data [28].

This talk will summarize our recently proposed silicon photonic architecture that uses an electro-optic circuit to calculate the gradient vector of each neural network layer in situ, the most computationally expensive operation performed during the backward pass. The proposed architecture exploits the speed (10s of GHz range in photonics but only 100s of MHz in electronics) and energy advantages of photonics to determine the gradient vector of each neural network layer in a single operational cycle.

The renaissance of neuromorphic photonics is enabled by the confluence of three areas (Fig. 1): technological advances in integrated photonics due to silicon photonics, algorithmic advances in machine learning algorithms, and advances in analog photonic signal processing. In the recent roadmap articles [10], [29], [30], we outlined some scientific and technological advances necessary to meet the challenges of envisioning a practical neuromorphic processor.

References

[1] Zhang, Shuai, et al. "Deep learning based recommender system: A survey and new perspectives." *ACM Computing Surveys (CSUR)* 52.1 (2019): 1-38.

[2] Young, Tom, et al. "Recent trends in deep learning-based natural language processing." *IEEE Computational intelligenCe magazine* 13.3 (2018): 55-75.

[3] Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." *arXiv preprint arXiv:1605.07678* (2016).

[4] Markidis, Stefano, et al. "Nvidia tensor core programmability, performance & precision." *2018 IEEE international parallel and distributed processing symposium workshops (IPDPSW)*. IEEE, 2018.

[5] Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." *Proceedings of the 44th annual international symposium on computer architecture*. 2017.

[6] Shen, Yichen, et al. “Deep learning with coherent nanophotonic circuits.” Nat. Photon. 11.7 (2017): 441-446.

[7] Feldmann, Johannes, et al. "Parallel convolutional processing using an integrated photonic tensor core." *Nature* 589.7840 (2021): 52-58.

[8] Miscuglio, Mario, and Volker J. Sorger. "Photonic tensor cores for machine learning." *Applied Physics Reviews* 7.3 (2020): 031404.

[9] Bangari, Viraj, et al. "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)." *IEEE Journal of Selected Topics in Quantum Electronics* 26.1 (2019): 1-13.

[10] Shastri, Bhavin J., et al. “Photonics for artificial intelligence and neuromorphic computing.” Nat. Photon. 15.2 (2021): 102-114.

[11] Huang, Chaoran, et al. "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits." *APL Photonics* 5.4 (2020): 040803.

[12] Zhang, Weipeng, et al. "Microring weight banks control beyond 8.5-bits accuracy." *arXiv preprint arXiv:2104.01164* (2021).

[13] Prucnal, Paul R., and Bhavin J. Shastri. Neuromorphic photonics. CRC Press, 2017.

[14] Huang, Chaoran, et al. "Prospects and applications of photonic neural networks." *Advances in Physics: X* 7.1 (2022): 1981155.

[15] Brunner, Daniel, et al. “Parallel photonic information processing at gigabyte per second data rates using transient states.” Nat. Commun. 4.1 (2013): 1-7.

[16] Vandoorne, Kristof, et al. “Experimental demonstration of reservoir computing on a silicon photonics chip.” Nat. Commun. 5.1 (2014): 1-6.

[17] Larger, Laurent, et al. “Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing.” Opt. Express 20.3 (2012): 3241-3249.

[18] Hughes, Tyler W., et al. “Training of photonic neural networks through in situ backpropagation and gradient measurement.” Optica 5.7 (2018): 864-871.

[19] Tait, Alexander N., et al. “Neuromorphic photonic networks using silicon photonic weight banks.” Sci. Rep. 7.1 (2017): 1-10.

[20] Tait, Alexander N., et al. “Silicon photonic modulator neuron.” Phys. Rev. Appl. 11.6 (2019): 064043.

[21] Tait, Alexander N., et al. “Broadcast and weight: an integrated network for scalable photonic spike processing.” J. Lightwave Technol. 32.21 (2014): 4029-4041.

[22] Huang, Chaoran, et al. "A silicon photonic–electronic neural network for fibre nonlinearity compensation." *Nature Electronics* 4.11 (2021): 837-844.

[23] Esser, S., et al. "Convolutional networks for fast, energy-efficient neuromorphic computing. arXiv 2016." *arXiv preprint arXiv:1603.08270*.

[24] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for deep learning in NLP." *arXiv preprint arXiv:1906.02243* (2019).

[25] Nøkland, Arild. "Direct feedback alignment provides learning in deep neural networks." *Advances in neural information processing systems* 29 (2016).

[26] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323.6088 (1986): 533-536.

[27] Launay, Julien, et al. "Direct feedback alignment scales to modern deep learning tasks and architectures." *Advances in neural information processing systems* 33 (2020): 9346-9360.

[28] M. Refinetti, S. d’Ascoli, R. Ohana, and S. Goldt, “The dynamics of learning with feedback alignment,” arXiv:2011.12428 [cond-mat, stat], Nov 2020.

[29] Berggren, Karl, et al. “Roadmap on emerging hardware and technology for machine learning.” Nanotechnology 32.1 (2020): 012002.

[30] Ferreira De Lima, Thomas, et al. “Progress in neuromorphic photonics.” Nanophotonics 6.3 (2017): 577-599.