Advanced Photonics Congress (BGPP, IPR, Networks, NOMA, PVLED, SOF, SPPCom) © Optica Publishing Group 2022

# Silicon Photonics Neural Networks for Training and Inference

**Bhavin J. Shastri[1,2], Matthew J. Filipovich[1], Zhimu Guo[1], Paul R. Prucnal[2], Sudip Shekhar[3], and Volker J. Sorger[4]**

*[1]Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada*
*[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*
*shastri@ieee.org*

**Abstract:** Deep learning hardware accelerators based on analog photonic networks are trained on standard digital electronics. We discuss on-chip training of neural networks enabled by a silicon photonic architecture for parallel, efficient, and fast data operations.

Analog photonic computing has garnered significant interest as an alternative to conventional electronic computer architectures [1,2]. The emerging field of neuromorphic (i.e., neuron-isomorphic) photonics [3,4] proposes implementing high-performance neural networks and related machine learning algorithms using electro-optic circuits. Many neuromorphic photonic systems [5] have been proposed as accelerators for machine learning inference [6-8], neuromorphic computing [4, 9-12], and tensor cores for matrix multiplications [13-15]. 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 [16]. The substantial energy required to train large neural networks using standard von Neumann architectures presents a high financial and environmental cost [17]. Training large neural networks is an area of machine learning that would benefit from photonics' low power consumption and high information processing bandwidth.

The recently proposed direct feedback alignment (DFA) supervised learning algorithm [18] has gathered interest as a bio-plausible alternative to the popular backpropagation training algorithm [19]. 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 [19]. 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 [19]. The DFA algorithm has also been shown to perform fine-tuned backpropagation in applications requiring state-of-the-art deep learning networks, including natural language processing and neural view synthesis [20]. 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 [21].

This talk will summarize our recently proposed silicon photonic architecture [22] 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.

While practical neuromorphic processors may be years away, we have outlined in [3], [23] some scientific and technological advances necessary to meet the challenges.

## References

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

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

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

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

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

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

[7] Lin, Xing, et al. "All-optical machine learning using diffractive deep neural networks." Science 361.6406 (2018): 1004-1008.

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

[9] Shi, B., Calabretta, N. & Stabile, R. Deep neural network through an InP SOA-based photonic integrated cross-connect. IEEE J. Sel. Top. Quantum Electron. 26, 7701111 (2020).

[10] Mourgias-Alexandris, G. et al. An all-optical neuron with sigmoid activation function. Opt. Express 27, 9620–9630 (2019).

[11] Peng, H. T. et al. Neuromorphic photonic integrated circuits. IEEE J. Sel. Top. Quant. Electron. 24, 6101715 (2018).

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

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

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

[15] 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.

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

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

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

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

[20] 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.

[21] 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.

[22] Filipovich, Matthew J., et al. "Monolithic Silicon Photonic Architecture for Training Deep Neural Networks with Direct Feedback Alignment." arXiv preprint arXiv:2111.06862 (2021).

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