Training Deep Neural Networks in Situ with Neuromorphic Photonics

Matthew J. Filipovich, Zhimu Guo, Bicky A. Marquez, Hugh D. Morison, and Bhavin J. Shastri
Department of Physics, Engineering Physics and Astronomy,
Queen's University, Kingston, ON K7L 3N6, Canada
mfilipovich@ieee.org

***Abstract*—We discuss an optoelectric architecture for executing the direct feedback alignment algorithm for neural network training. Using 1000 microring resonators, we can theoretically obtain speeds of 9.95 tera operations per second.**

***Index Terms*—Neuromorphic Photonics, Machine Learning, Training Algorithms, Direct Feedback Alignment**

I. Introduction

The emerging field of neuromorphic photonics proposes to implement neuromorphic devices using optoelectronics that are well-suited for machine learning operations [1]. The main benefits of using photonics compared to their electronic counterparts are i) improved energy efficiency for matrix multiplication operations, ii) higher speeds (photonic systems can operate at upwards of 20 GHz), and iii) increased information density [2]. Silicon photonics has shown to be a promising platform for neuromorphic applications due to its compatibility with the mature silicon integrated circuit industry and the availability of high-quality silicon-on-insulator wafers that allow the observation of nonlinear optical interactions [3]. The high refractive index contrast between silicon ($n = 3.45$) and SiO_2 ($n = 1.45$) allows for the manufacturing of photonic devices to the hundreds of nanometer level.

Deep learning algorithms have high computation and memory costs that pose significant challenges to the current hardware platforms executing them [4]. The substantial energy consumption required to train large neural networks using standard von Neumann architectures also presents significant financial and environmental costs [5]. We present an optoelectric analog circuit that executes the direct feedback alignment algorithm on silicon photonic hardware and can be implemented on a photonic integrated chip (PIC).

II. Neuromorphic Photonic Training Architecture

The direct feedback alignment (DFA) algorithm is a supervised learning algorithm for training neural networks that propagates error through fixed random feedback connections directly from the output layer to the hidden layers [6]. The proposed photonic architecture calculates the gradient $\boldsymbol{\delta}^{(k)}$ in situ for each hidden layer k, as defined by

$$\boldsymbol{\delta}^{(k)} = \mathbf{B}^{(k)}\mathbf{e} \odot g'\left(\mathbf{a}^{(k)}\right), \tag{1}$$

where $\mathbf{B}^{(k)}$ is a fixed random weight matrix with appropriate dimensions, $\mathbf{e}$ is the gradient of the cost function in the output layer, $\odot$ is the Hadamard product (element-wise multiplication operator), and g' is the derivative of the activation function with respect to $\mathbf{a}^{(k)}$, which is the sum of the weighted input signals in the kth layer. The network's weights and biases are then updated off-chip for each layer k using the calculated gradient $\boldsymbol{\delta}^{(k)}$.

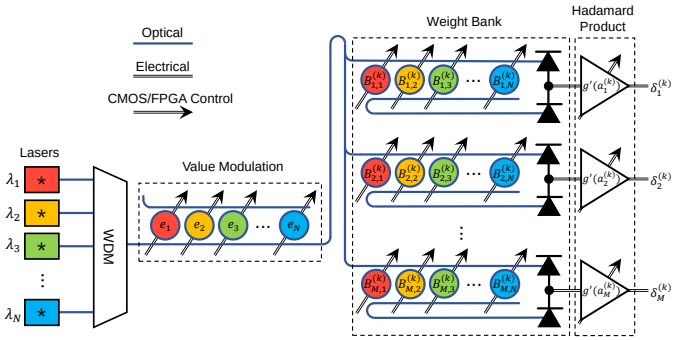

Fig. 1. Photonic architecture implementing the DFA training algorithm. The gradient of the hidden layer k is calculated through the matrix-vector product of the $M \times N$ matrix $\mathbf{B}^{(k)}$ and the output layer gradient vector $\mathbf{e}$ of size N. The Hadamard product $\mathbf{B}^{(k)}\mathbf{e} \odot g'\left(\mathbf{a}^{(k)}\right)$ is executed using TIAs, yielding the desired vector $\boldsymbol{\delta}^{(k)}$ in the electrical domain.

A schematic of the photonic DFA architecture is shown in Fig. 1. Wavelength-division multiplexing (WDM) is used to combine multiple optical signals onto a single waveguide. The gradient vector $\mathbf{e}$ of size N is calculated off-chip and encoded by modulating the amplitude of N distinct input laser wavelengths coupled onto the PIC. An array of N microring resonators (MRRs) is used to modulate the incoming light. Exploiting the plasma dispersion effect, the refractive index of each MRR is adjusted by varying the concentration of carriers through external biasing. Assuming the intensity of all input wavelengths is the same, we use an encoding scheme that linearly maps the desired value in the range [0, 1] to the pass port transmission T_p.

The modulated optical signals representing the vector $\mathbf{e}$ are then coupled into the photonic weight bank [7]. The photonic weight bank consists of M rows of MRR arrays with N MRRs per row. If the size of the photonic weight bank is larger than the dimensions of the matrix $\mathbf{B}$, the redundant MRRs can be tuned with a weighting of zero. The drop and pass ports in each row in the photonic weight bank are connected to a balanced photodetector which sums the signals.

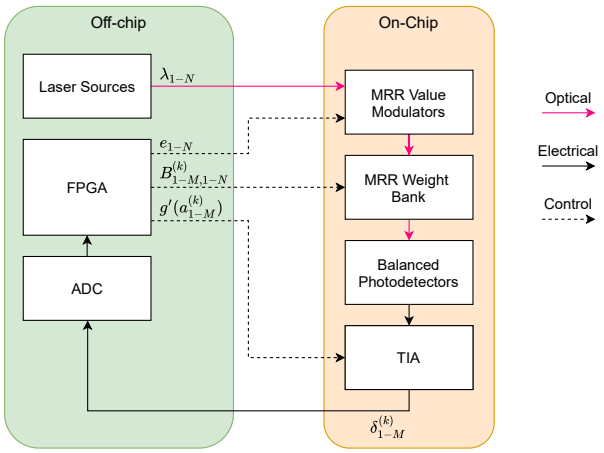

Fig. 2. Data pipeline of the neuromorphic photonic DFA architecture.

The Hadamard product $\mathbf{B}^{(k)}\mathbf{e}\odot g'\left(\mathbf{a}^{(k)}\right)$ is performed using a set of transimpedance amplifiers (TIAs) where each balanced photodetector output is connected to a TIA. The vector $\mathbf{a}^{(k)}$ is encoded onto voltage signals that set the gains of the TIAs, which can be manufactured and integrated onto PICs using standard CMOS processes [8].

The DFA architecture requires a control source off-chip to tune the active components on the PIC. The control source is connected to i) the MRRs that modulate the incoming laser light with the $\mathbf{e}$ values, ii) the MRRs in the weight bank that execute the matrix-vector multiplication operation, and iii) the TIAs that implement the Hadamard product. A diagram of the architecture's data pipeline is shown in Fig. 2.

Summing the two transmission ports in the electrical domain allows the MRRs to be encoded with a weighting W in the range [-1,1], assuming there is minimal loss in the system:

$$W\left(\phi\right) = 2T_p\left(\phi\right) - 1, \tag{2}$$

where the pass port transmission T_p is a function of the round trip phase shift ϕ. The relationship between the applied bias to the MRR and the change in refractive index must be determined experimentally. The weights can then be determined from (2) using ϕ as a function of the applied bias. This is possible since the wavelength and MRR radius are constant, so ϕ is only dependent on the refractive index.

The size of the photonic weight bank is physically bounded by the dimensions of the PIC and the maximum number of supported WDM channels in a single waveguide. However, the dimensions of the photonic weight bank do not restrict the size of the neural network being trained; if the size of the matrix $\mathbf{B}^{(k)}$ is larger than the dimensions of the photonic weight bank, the product can be determined over multiple clock cycles by calculating a subset of the output vector at each cycle. Thus, the computation of the hidden layer gradients using the photonic architecture is $\mathcal{O}(n)$ with respect to both the number of hidden layers and the ceiling function of the ratio between the matrix $\mathbf{B}^{(k)}$ size and the photonic weight bank dimensions.

The theoretical speed bottleneck of the system is the throughput of the ADC, which has been shown to operate at 5 GS/s [9]. An estimate of the average operations per second (OPS) for one training example can be calculated using the operational limit from the ADC:

$$\text{OPS} = \frac{5\cdot 10^9}{l}(2N-1)\left\lceil\frac{N}{d_R}\right\rceil^{-1}\left(\sum_{k=1}^{l} M_k\left\lceil\frac{M_k}{d_C}\right\rceil^{-1}\right), \tag{3}$$

where l is the number of hidden layers, M_k is the size of the kth hidden layer, N is the size of the output layer, d_R and d_C are the maximum row and column dimensions of the weight bank (related to the size of the PIC and the WDM channel limit), and $\lceil\ \rceil$ is the ceiling function. The performance of the photonic architecture while training different sized networks using 1000 MRRs is shown in Table I.

TABLE I
ESTIMATED OPERATIONS PER SECOND EXECUTED BY THE PHOTONIC DFA ARCHITECTURE WHILE TRAINING NEURAL NETWORKS OF VARIOUS SIZES.

Network Size	d_R	d_C	OPS
$784\times 800\times 10$	10	100	9.40×10^{12}
$784\times 400\times 100$	100	10	9.91×10^{12}
$100\times 8000\times 10000\times 100$	10	100	9.95×10^{12}

III. Conclusion

We present a neuromorphic photonic architecture that performs the fundamental operations of the DFA algorithm. The system is cascadable for training neural networks of various sizes and computes $\mathcal{O}(n)$ with respect to both the number of hidden layers and the ceiling function of the ratio between the matrix $\mathbf{B}^{(k)}$ size and the photonic weight bank dimensions. The architecture can compute 9.95 tera operations per second using 1000 MRRs.

References

[1] P. R. Prucnal *et al.*, *Neuromorphic photonics*. 2017.

[2] M. A. Nahmias *et al.*, "Photonic multiply-accumulate operations for neural networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, p. 1–18, Jan 2020.

[3] B. Jalali *et al.*, "Silicon photonics," *Journal of Lightwave Technology*, vol. 24, p. 4600–4615, Dec 2006.

[4] S. K. Esser *et al.*, "Convolutional networks for fast, energy-efficient neuromorphic computing," *Proceedings of the National Academy of Sciences*, vol. 113, p. 11441–11446, Oct 2016.

[5] E. Strubell *et al.*, "Energy and policy considerations for deep learning in nlp," in *ACL*, 2019.

[6] A. Nøkland, "Direct feedback alignment provides learning in deep neural networks," in *Advances in Neural Information Processing Systems 29*, p. 1037–1045, Curran Associates, Inc., 2016.

[7] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports*, vol. 7, p. 7430, Aug 2017.

[8] H. Zheng *et al.*, "A linear and wide dynamic range transimpedance amplifier with adaptive gain control technique," *Analog Integrated Circuits and Signal Processing*, vol. 90, p. 217–226, Jan 2017.

[9] J. Fang *et al.*, "A 5-gs/s 10-b 76-mw time-interleaved sar adc in 28 nm cmos," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, p. 1673–1683, Jul 2017.