# Training Deep Neural Networks in Situ with Neuromorphic Photonics

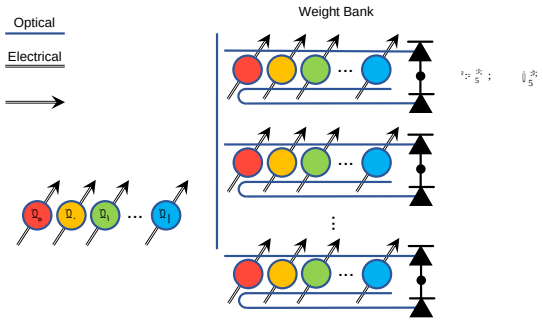
Matthew J. Filipovich, Zhimu Guo, Bicky A. Marquez, Hugh D. Morison, and Bhavin J. Shastri
*Department of Physics, Engineering Physics and Astronomy,*
*Queen's University, Kingston, ON K7L 3N6, Canada*
mfilipovich@ieee.org

***Abstract***

**. 1000 , 9.95 .**

***Index Terms***

**, , A , A**




## 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 C

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

The Hadamard product $\mathbf{B}^{(\ )}\mathbf{e} \odot g'\left(\mathbf{a}^{(\ )}\right)$ is performed using a set of transimpedance amplifiers (TIAs) where each balanced photodetector output is connected to a TIA. The vector $\mathbf{a}^{(\ )}$ 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:

$$W(\varphi) = 2T\ (\varphi) - 1, \tag{2}$$

where the pass port transmission $T$ is a function of the round trip phase shift $\varphi$. 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 $\varphi$ as a function of the applied bias. This is possible since the wavelength and MRR radius are constant, so $\varphi$ 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 $\mathbf{B}^{(\ )}$ 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 $\mathbf{B}^{(\ )}$ size and the photonic weight bank dimensions.