Silicon Photonics for Neuromorphic Computing and Artificial Intelligence: Applications and Roadmap

B. J. Shastri[1;2;3], C. Huang[2], A. N. Tait[2], T. Ferreira de Lima[2], and P. R. Prucnal[2]

[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]Vector Institute for Artificial Intelligence, Toronto, ON M5G 1M1, Canada

Abstract |

high parallelism and speed of photonics to bring the same neuromorphic algorithms to applications requiring multiple channels of multi-gigahertz analog signals, which digital processing struggles to process in real-time.

By combining the high bandwidth and parallelism of photonic devices with the adaptability and complexity attained by methods similar to those seen in the brain, photonic neural networks (PNNs) have the potential to be orders of magnitude faster than state-of-the-art electronic processors while consuming less energy per computation [5]. The goal of neuromorphic photonic processors is not to replace conventional computers, but to enable applications that are unreachable at present by conventional computing technology — those requiring low latency, high bandwidth and low energies [6, 7]. As shown in Figure 1, examples of applications for ultrafast neural networks include: 1) Enabling fundamental physics breakthroughs: qubit read-out classification, high-energy-particle collision classification, fusion reactor plasma control; 2) Nonlinear programming: solving nonlinear optimization problems (robotics, autonomous vehicles, predictive control) and partial differential equations; 3) Machine learning acceleration: vector–matrix multiplications, deep learning inference, ultrafast or online learning; 4) Intelligent signal processing: wideband radio-frequency signal processing, fibre-optic communication.

2. NEUROMORPHIC PHOTONICS APPROACHES

Neuromorphic photonic [6, 8] approaches can be divided into two main categories (Figure 2): coherent (single wavelength) and incoherent (multiwavelength) approaches. Neuromorphic systems based on reservoir computing [9–11] and Mach-Zehnder interferometers [12, 13] are example of coherent approaches. In reservoir computing the predefined random weights of their hidden layers cannot be modified. An alternative approach uses silicon photonics to design fully programmable neural networks [14, 15], with a so-called broadcast-and-weight protocol [16]. In this architecture, photonic neurons output optical signals with unique wavelengths. These are multiplexed into a single waveguide and broadcast to all others, weighted, and photodetected. Each connection between a pair of neurons is configured independently by one microring resonator (MRR) weight, and the wavelength division multiplexed (WDM) carriers do not mutually interfere when detected by a single photodetector. Consequently, the physics governing the neural computation is fully analog and does not require any logic operation or sampling, which would involve serialization and here solve that problem by using optoelectronic components (O/E/O), which can be mated with standard electronics providing reconfigurability. However, neuromorphic photonic circuits are challenging to scale up because they do not benefit from digital information, memory units and a serial processor, and therefore requires a physical unit for each element in a neural network, increasing size, area and power consumption. Here, integration costs must also be considered, since the advantages of using analog photonics (high parallelism and high bandwidth) must outweigh the costs of interfacing it with digital electronics (requiring both O/E and analog/digital conversion).

3. VISION OF A NEUROMORPHIC PROCESSOR

Recently, in our tutorial, Ref. [17], we proposed a vision for a neuromorphic processor. We discussed how such a neuromorphic chip could potentially be interfaced with a general-purpose computer (Figure 3), i.e., a CPU, as a coprocessor to target specific applications. In general, there are two levels of complexity associated with co-integrating a general-purpose electronic processor with an application-specific optical processor. Firstly, a CPU processes a series of computation instructions in an undecided amount of time and is not guaranteed to be completed. Neural networks, on the other hand, can process data in parallel and in a deterministic amount of time. CPUs have a concept of a `fixed' instruction set on top of which computer software can be developed. However, a neuromorphic processor would require a hardware description language (HDL) because it describes the intended behavior of a hardware in real-time. Secondly, seamlessly interfacing a photonic integrated circuit with an electronic integrated circuit will take several advances in science and technology including on-chip lasers and amplifiers, co-integration of CMOS with silicon photonics, system packaging, high-bandwidth digital-to-analog converters (DAC) and analog-to-digital converters (ADCs).

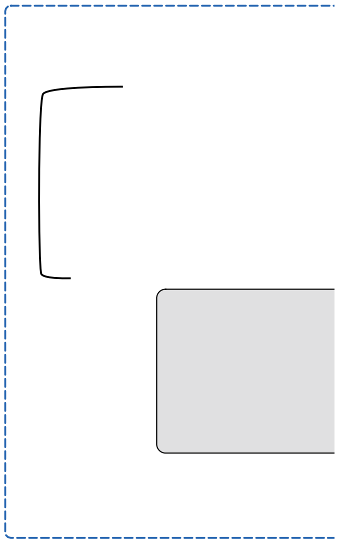

Figure 3: Simplified schematics of a neuromorphic processor. Thanks to integrated laser sources and photodetectors, it can input and output RF signals directly as an option to optically modulated signals. The waveform generator allows for programming arbitrary stimulus that can be used as part of a machine learning task. Reproduced from [17].

4. MACHINE LEARNING APPLICATION: FIBER NONLINEARITY IMPAIRMENT COMPENSATION

The world is witnessing an explosion of internet traffic. The global internet traffic has reached 5.3

on the consistent improvement in CMOS technology [19]. However, the exponential hardware scaling of ASIC based DSP chips, which is embodied in Moore's law as other digital electronic hardware, is fundamentally unsustainable. In parallel, many efforts are focused on developing new DSP algorithms to minimize computational complexity, but usually at the expense of reducing transmission link performances [20].

Instead of embracing such a complexity-performance trade-off, an alternative approach is to explore new hardware platforms that intrinsically offer high bandwidth, speed, and low power consumption [6, 21, 22]. Machine learning algorithms, especially neural networks, have been found effective in performing many functions in optical networks, including dispersion and nonlinearity impairments compensation, channel equalization, optical performance monitoring, traffic prediction, etc [23].

PNNs are well suited for optical communications because the optical signals are processed directly in the optical domain. This innovation avoids prohibitive energy consumption overhead and speed reduction in ADCs, especially in data center applications. In parallel, many PNN approaches are inspired by optical communication systems, making PNNs naturally suitable for processing optical communication signals. For example, we proposed synaptic weights and neuron networking architecture based on the concept of WDM to enable fan-in and weighted addition [14]. This architecture can provide a seamless interface between PNNs and WDM systems, which can be applied as a front-end processor to address inter-wavelength or inter-mode crosstalks problems that DSP usually lacks the bandwidth or computing power to process (e.g., fiber nonlinearity compensation in WDM systems). Moreover, PNNs combine high-quality waveguides and photonic devices that have been initially developed for telecommunications. Therefore, PNNs, by default, can support fiber optic communication rates and enable real-time processing. For example, the scalable silicon PNN proposed by the authors is composed of microring resonator (MRR) banks for synaptic weighting and O/E/O neurons to produce standard machine learning activation functions. The MRR weight bank is inspired by WDM filters, and the O/E/O neurons use typical silicon photodetector and modulator. Therefore, the optimization of associated devices in PNNs can utilize the fruits of the entire silicon photonic ecosystem that is driven by telecommunications and data center applications.

Figure 4: (a) Concept of training and implementing photonic neural networks. Inset shows a transfer function of the photonic neural network measured with real-timepgmassreal-ecaiscati8ngy-olonloptimizat8nganction392Tf042.19-11.9

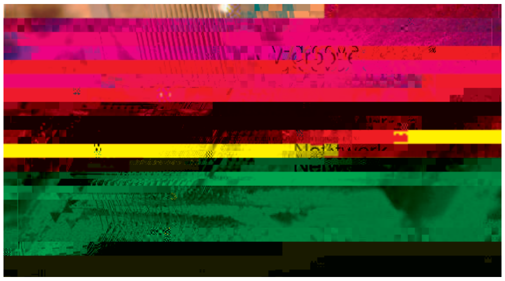

Figure 5: Micrograph image of a wirebonded photonic neural network chip from Princeton University's Lightwave Lab.

capacity and transmission distance. One reason is that the nonlinear interplay between signal, noises, and optical fibers negates the accuracy of conventional nonlinear compensation algorithms based on digital backpropagation. Another reason is, the implementation of most nonlinear compensation algorithms in DSP chips demands excessive resources. In contrast, the neural network approach can learn and approximate the nonlinear perturbation from the abundant training data, rather than solely relying on the physical fiber model (known as stochastic nonlinear Schrodinger equation). Based on the perturbation methods, the derived neural network algorithm has enabled compensating the nonlinear distortion in a 10800 km fiber transmission link with 32 Gbaud signals [24]. In Ref. [15], we developed a PNN platform based on the so-called \neuromorphic" approach, aiming to map physical models of optoelectronic systems to abstract models of neural networks (which differs from the reservoir approaches). By doing so, the PNN system can leverage existing machine learning algorithms (i.e., backpropagation) and map training results from simulations to heterogeneous photonic hardware. The concept is shown in Figure 2. A proof-of-concept experiment demonstrates the real-time implementation of a trained neural network model using an integrated silicon PNN chip [22]. In this work, the authors experimentally demonstrated that the silicon PNN can produce a similar Q factor improvement compared to the simulated neural network for nonlinear compensation as shown in Figure 4, but it promises to process the communication data in real-time and with high bandwidth and low latency.

We also proposed a photonic architecture enabling all-to-all continuous-time recurrent neural networks (RNN) [15]. Recurrent neural networks can resemble optical fiber transmission systems: the linear neuron-to-neuron connections with internal feedback is analog to linear multiple-input multiple-output (MIMO) fiber channel with dispersive memory. With neuron nonlinearity, RNNs can be ideally used to approximate all types of linear and nonlinear effects in a fiber transmission system and compensate for different transmission impairments. RNNs, consisting of many feedback connections, are computationally expensive for digital hardware and require at least milliseconds to conduct a single inference. Contrarily, in photonic RNN, the feedback operations are simply done by busing the signals on photonic waveguides, allowing photonic hardware capable of converging non-critical. Therefore, traditional computers are not appropriate to implement algorithms depending on QP for high-speed applications such as signal processing and control systems. In machine learning, many algorithms, such as SVM, require offline training because of the computational complexity of QP, but would be much more effective were they trained online.

with incongruous fabrication processes (silicon-on-insulator, CMOS, FinFETs). Silicon photonics is becoming an ideal platform for integrating these devices while offering a combination of foundry compatibility, device compactness, and cost that enables the creation of scalable photonic systems on chip.

Materials: Energy efficient and fast switching optical and electro-optical materials are needed for non-volatile photonic storage and weighting, as well as high-speed optical switching and routing, with low power consumption. Neural non-linearities are already possible on mainstream platforms using electrooptic transfer functions [16], but new materials promise significant performance op-