Photonics for Neuromorphic Computing

Paul R. Prucnal[(1)], Alexander N. Tait[(1)], Mitchell A.Nahmias[(1)], Thomas Ferreira de Lima[(1)], Hsuan-Tung Peng[(1)], Bhavin J. Shastri[(1)]

[(1)] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA, prucnal@princeton.edu

Abstract *An emerging field at the nexus of photonics/neuroscience, neuromorphic photonics combines the advantages of optics/electronics. In this tutorial, we will look at challenges of photonic information processing, describe photonic neural-network approaches, and offer a glimpse at this field's future.*

Introduction

Neuromorphic (i.e., brain-inspired) processors are widely considered as one the next frontiers in computing. The proliferation of microelectronics has enabled the emergence of next-generation industries to support emerging artificial intelligence services and high-performance computing. These data-intensive enterprises rely on continual improvements in hardware. The demand for data will continue to grow as smart gadgets multiply and become increasingly integrated into our daily lives. However, this rapidly expanding space has been subverted by a stark reality: exponential hardware scaling in digital electronics is fundamentally unsustainable.

Neuromorphic photonics (Fig. 1) is an emerging field at the interface of photonics and neuroscience that combines the advantages of optics and electronics to build systems with high efficiency, high interconnectivity and high information density[1,2] . In this tutorial, we will look at some of the traditional challenges of photonic information processing, describe the photonic neural-network approaches being developed by our lab and others, and conclude with a future outlook of neuro-inspired photonic processing

The Emergence of Photonic Neural Networks

Instead of using digital 0's and 1's, neural networks represent information in analog signals, which can take the form of either continuous real number values, or spikes, in which information is encoded in the timing between short pulses[3]. Rather than abiding by a sequential set of instructions, neurons process data in parallel and are programmed by the connections between them (Fig. 2). The input into a particular neuron is a linear combination—also referred to as weighted sum—of the output of other neurons. These connections can be weighted with negative and positive values, respectively, which are called "inhibitory" and "excitatory" synapses. The weight is therefore represented as a real number, and the interconnection network can be expressed as a matrix.

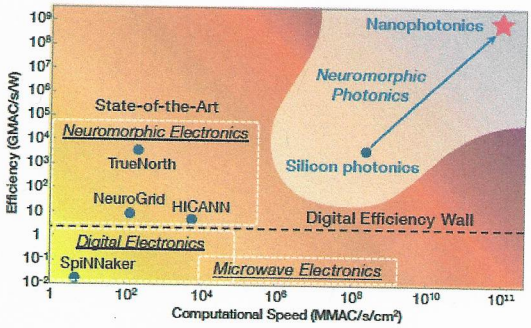

Fig. 1: Performance scaling between electronic and photonic neuromorphic hardware. TrueNorth is a leading electronic neuromorphic platform developed by IBM; Neurogrid: Stanford; HICANN: Heidelburg; SpiNNaker: Manchester.

Photonics is a promising technology to implement neural networks (Fig. 2). The greatest computational burden in neural networks lies with the interconnectivity: in a system with N neurons, if every neuron can communicate with every other (plus itself), this results in N^2 connections. Just one more neuron adds N more connections, which can be prohibitive if N is large. Photonic systems could address this problem in two ways: 1) waveguides can boost interconnectivity by carrying many signals at the same time through optical multiplexing, and 2) low-energy, photonic operations can reduce the computational burden of performing linear functions such as weighted sum. For example, by associating each node with a color of light, a network could support N additional connections without necessarily adding any physical wires. A comparison of the potential speed and efficiency of photonic based systems is shown in Fig 1. These advantages have motivated researchers to investigate a number of photonic neural models that exhibit a large range of interesting properties.

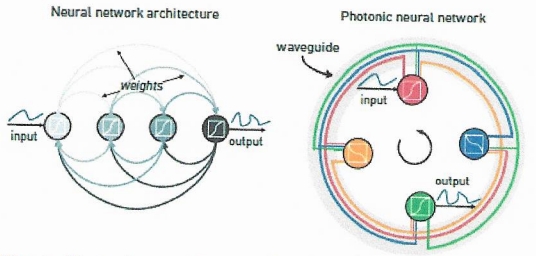

Fig. 2: Photonic neural nets (right) can solve the interconnect bottleneck by using one waveguide to carry signals from many connections (easily $N^2 \sim 10,000$) simultaneously.

Photonic Neuron Implementations

Researchers have engineered dynamical lasers to resemble the biological behavior of neurons[4]; an example of an integrated system currently under investigation at Princeton is shown in Fig. 3. Laser neurons are capable of operating approximately 100 million times the speed of their biological counterparts, owing to the speed of optoelectronic physics over biochemical interactions. They represent neural spikes via optical pulses by operating under a dynamical regime called "excitability." Excitability is a behavior in feedback systems in which small inputs that exceed some threshold cause a major excursion from equilibrium, which in the case of a laser neuron, releases an optical pulse. This event is followed by a recovery back to equilibrium, or refractory period.

We discovered[5] a theoretical link between the dynamics of semiconductor lasers and a common neuron model used in computational neuroscience, and demonstrated how a laser with an embedded graphene section could effectively emulate such behavior[3]. Building from these results, a number of researchers have fabricated, tested, and proposed a variety of laser neurons with various feedback conditions[4]. These include two-section models in semiconductor lasers, photonic crystal nanocavities, polarization sensitive vertical cavity lasers, lasers with optical feedback or optical injection, and linked photodetector-laser systems with receiverless connections or resonant tunneling.

A recently demonstrated[6] approach based on optical modulators has been investigated recently that has the potential to exhibit much lower conversion costs from one processing stage to another. In addition, it would be fully integrated systems on silicon photonic platforms.

Scalable Photonic Neural Networks

Recently, researchers have investigated interconnection protocols that can tune to any desired net-

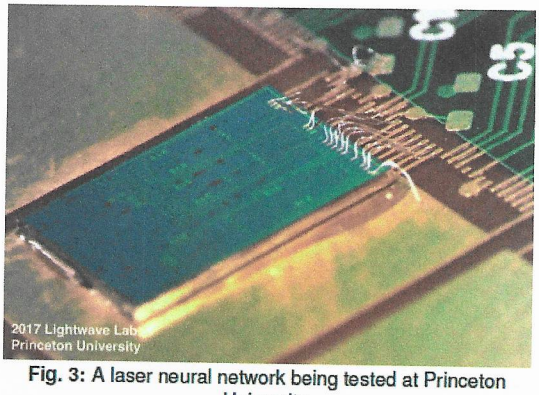

Fig. 3: A laser neural network being tested at Princeton University.

work configuration. Arbitrary weights allow a wide array of potential applications based on classical neural networks. There are several notable approaches in the literature that use complementary physical effects in this regard.

A neural network architecture called "broadcast and weight" uses groups of tunable filters to implement weights on signals encoded onto multiple wavelengths[7]. Tuning a given filter on and off resonance changes the transmission of each signal through that filter, effectively multiplying the signal with a desired weight. The resulting weighted signals travel into a photodetector, which can receive many wavelengths in parallel to perform a summing operation. Broadcast and weight takes advantage of the enormous information density available to on-chip photonics through the use of optical multiplexing, and is compatible with a number of laser neuron models. Filter-based weight banks have also been investigated both theoretically and experimentally in the form of closely packed microring filters, prototyped in a silicon photonic platform. A fully integrated superconducting optoelectronic network was recently proposed to offer unmatched energy efficiency[8]. While based on an exotic superconducting platform, the interconnect architecture could be compatible with broadcast-and-weight.

A "coherent" approach utilizes destructive or constructive interference effects in optical interferometers to implement a matrix-vector operation of incoming signals[9]. There is no need to convert from the optical domain to the electrical domain; interfacing such systems with photonic, nonlinear nodes (i.e., based on the Kerr effect) could allow for energy efficient, passive all-optical processors. However, the coherent approach is limited to only one wavelength and requires devices that are much larger than tunable filters, limiting the information density of the approach in its

current form. In addition, all-optical interconnects must grapple with both amplitude and phase and there is still no proposed solution to prevent phase noise accumulation from one stage to another. Nonetheless, the investigation of large-scale networking schemes is a promising direction for the integration of various technologies in the field towards highly scalable on-chip photonic systems.

A contrasting approach to tunable neural networks, "reservoir computing," extracts useful information from a fixed, possibly nonlinear system of interacting nodes[10]. Reservoirs require far fewer tunable elements than neural network models to run effectively, making them less challenging to implement in hardware; however, they cannot be easily programmed. These systems have utilized optical multiplexing strategies in both time and wavelength. Experimentally demonstrated photonic reservoirs have displayed state-of-the-art performance in benchmark classification problems, such as speech recognition.

Discussion

Although it remains to be seen in what ways photonic processing systems will complement microelectronic hardware, current technological developments point in a promising direction. For example, the fixed cost of electronic to photonic conversion is no longer as energetically unfavorable: a modern silicon photonic link can transmit a photonic signal using only femtojoules of energy per bit of information, while thousands of femtojoules of energy are consumed per operation in even the most efficient digital electronic processors, including IBM's TrueNorth cognitive computing chip and Google's tensor processing unit. This figure will improve as optoelectronic devices are scaled in performance. New modulators or lasers based on plasmonic localization, graphene modulation or nanophotonic cavities have the potential to increase this efficiency. The next generation of photonic devices could potentially consume only hundreds of attojoules of energy per time slot, allowing analog photonic processors to consume even less per operation.

There are many applications of photonic neural network technologies, especially in light of the developments mentioned above. For one, photonic systems can act as a co-processor to perform linear operations—including multiply-accumulate operations, fourier transforms, and convolutions—by implementing them in the photonic domain, potentially decreasing the energy consumption and increasing the throughput of signal processing, high performance computing and artificial intelligence algorithms. This could be a major boon for datacenters, which are increasingly dependent on such operations and have consistently doubled their energy consumption every four years.

Secondly, photonic processors have unmatched speeds and latencies, which make them well-suited for specialized applications requiring either real-time response times or fast signals. One example is a front-end processor in radio frequency transceivers. As the wireless spectrum becomes increasingly overcrowded, the use of large, adaptive phased-array antennas that receive many more radio waves simultaneously may soon become the norm. Photonic neural networks could perform complex statistical operations to extract important data, including the separation of mixed signals or the classification of recognizable radio frequency signatures. A second example is in low-latency, ultrafast control systems. It is well known that recurrent neural networks can solve various problems that involve minimizing or maximizing some known function. A process method known as "Hopfield optimization" requires the solution to such a problem during each step of the algorithm, and could utilize the short convergence times of photonic networks for nonlinear optimization.

Just as fiber optics once rendered copper cables obsolete for long-distance communications, neuromorphic photonic processing has the potential to one day usher a paradigm shift in computing to create a smarter, more efficient world.

References

[1] P. R. Prucnal and B. J. Shastri. Neuromorphic Photonics (CRC Press, 2017).

[2] M. A. Nahmias et al. Opt. Photon. News 29, 34 (2018).

[3] B. J. Shastri et al. Sci. Rep. 5, 19126 (2016).

[4] P. R. Prucnal et al. Adv. Opt. Photon. 8, 228 (2016).

[5] M. A. Nahmias et al. IEEE J. Sel. Top. Quantum Electron. 19, 1800212 (2013).

[6] A. N. Tait et al. Sci. Rep. 7, 7430 (2017).

[7] A. N. Tait et al. J. Lightwave Technol. 32, 4029 (2014).

[8] J. M. Shainline et al. Phys. Rev. Appl. 7, 034013 (2017).

[9] Y. Shen et al. Nat. Photon. 11, 441 (2017).

[10] G. Van der Sande et al. Nanophotonics, 6, 561 (2017).