Neuromorphic Photonics for Deep Learning

V. Bangari[1], B. A. Marquez[1], A. N. Tait[2,3], M. A. Nahmias[3], T. Ferreira de Lima[3], H.-T. Peng[3], P. R. Prucnal[3], and B. J. Shastri[1,3]

[1]Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada
[2]National Institute of Standards and Technology, Boulder, CO 80305, USA
[3]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
shastri@ieee.org

Abstract: Co-integrated neuromorphic photonic and electronic processors promise orders of magnitude improvements in both speed and energy efficiency over purely digital electronic approaches. We discuss neuromorphic photonic systems and their application to deep convolutional neural networks inference.

Machine learning (ML) based on deep neural networks have demonstrated, in some cases, super-human performance in several complex tasks [1]. The rise in ML over the last decade can be attributed to: 1) algorithmic innovations [2]; 2) the Internet: an inexhaustible source of millions of training examples; and 3) new hardware: specifically, graphical processing units (GPUs) [3]. Deep neural networks are based on convolutional neural networks (CNNs) which are powerful and highly ubiquitous tools for extracting features from large datasets for applications such as computer vision and natural language processing. The success of CNNs for large-scale image recognition has stimulated research in developing faster and more accurate algorithms for their use. However, CNNs are computationally intensive and therefore results in long processing latency. One of the primary bottlenecks is computing the matrix multiplication required for forward propagation. In fact, over 80% of the total processing time is spent on the convolution [4]. Therefore, techniques that improve the efficiency of even forward-only propagation are in high demand and researched extensively [5]. Central processing units (CPUs) are inefficient at evaluating neural network models because they are centralized and instruction-based, whereas networks are distributed and capable of adaptation without a programmer. GPUs are more parallel, but, today, even they have been pushed to their limits [6].

Recently, there has been much investigation of neural networks implemented with integrated photonics—an emerging field called *neuromorphic photonics* [7]. By combining the high bandwidth and efficiency of photonic devices with the adaptive, parallelism and complexity attained by methods similar to those seen in the brain, photonic processors have the potential to be at least ten thousand times faster than electronic processors while consuming less energy per computation. In summary, the renewed interest in neuromorphic photonics has been heralded by advances in photonic integration technology [8], roadblocks in conventional computing performance [9], the return of neuromorphic electronics [10-13], and the inundation of ML with neural models [14].

Here, we will present a digital electronic and analog photonic (DEAP) architecture capable of performing efficient CNNs for image recognition. The competitive MNIST handwriting dataset [15] is used as a benchmark test for our DEAP CNN. At first, we train a standard two-layer CNN offline, after which network parameters are uploaded to the DEAP CNN. Our scope is limited to the forward propagation but includes power and speed analyses of our proposed architecture.

Photonic neural networks can be divided into two main categories: coherent (single wavelength) and incoherent (multiwavelength) approaches. Neuromorphic systems based on reservoir computing [16]–[18] and Mach-Zehnder interferometers [19], [20] are example of coherent approaches. Note, 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 [21], with a so-called broadcast-and-weight protocol [22]–[24]. This protocol is capable of implementing reconfigurable, recurrent and feed- forward neural network models, using a bank of tunable silicon microring resonators (MRRs) that recreate on-chip synaptic weights. Therefore, such a protocol allows it to emulate physical neurons. This architecture employs wavelength-division multiplexing (WDM) for scalability; here, each neuron in the network is associated with a unique wavelength of light while the reconfigurable tunable filters determine the neuron interconnectivity. The advantage of this approach over the aforementioned approaches is that it has already demonstrated fan-in, inhibition, time-resolved processing, and autaptic cascadability [25,26].

We have recently proposed a photonic network, DEAP, suited for convolutional neural networks. DEAP was estimated to perform convolutions between 2.8 and 14 times faster than a GPU while roughly using 0.75 times the energy consumption. A linear increase in processing speeds corresponds to a linear increase in energy consumption, allowing for DEAP to be as scalable as electronics. Our DEAP CNN design is compatible with mainstream silicon photonic device platforms. This approach leverages the advances in silicon photonics that have recently progressed to

978-1-7281-0615-1/19/$31.00 ©2019 IEEE

the level of sophistication required for large-scale integration. Furthermore, this proposed architecture allows the implementation of multi-layer networks to implement the deep learning framework.

In this talk we will provide an overview of neuromorphic photonic systems and their application to machine learning and specifically deep learning inference with DEAP. We will discuss the physical advantages of photonic processing systems and describe the underlying device models that allow practical systems to be constructed. We will discuss scalability in the context of designing a full-scale neuromorphic photonic processing system, considering aspects such as signal integrity, noise, and hardware fabrication platforms.

References

[1] B. Rajendra et al. arXiv preprint arXiv:1901.03690 (2018).
[2] Y. LeCun, Y. Bengio, and G. Hinton, *Nature* **521**, 436 (2015).
[3] V. K. Pallipuram, M. Bhuiyan, and M. C. Smith, *J. Supercomput.* **61**, 673 (2012).
[4] X. Li et al. *2016 45th International Conference on Parallel Processing (ICPP)* (2016, Aug.)
[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016).
[6] Diamond, T. Nowotny, and M. Schmuker, *Front. Neurosci.* **9**, 491 (2016).
[7] P. R. Prucnal and B. J. Shastri. *Neuromorphic Photonics* (CRC Press, 2017).
[8] D. Thomson et al. *J. Opt.* **18**, 073003 (2016).
[9] J. Hasler and H. B. Marr, *Front. Neurosci.* **7**, 118 (2013).
[10] P. A. Merolla et al. *Science* **345**, 668 (2014).
[11] S. B. Furber et al. *Proc. IEEE* **102**, 652 (2014).
[12] B. Benjamin et al. *Proc. IEEE* **102**, 699 (2014).
[13] M. Davies et al. *IEEE Micro.* **38**, 82 (2018).
[14] J. Schmidhuber, *Neural Netw.* **61**, 85 (2015).
[15] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. Available: http://yann.lecun.com/exdb/mnist/
[16] D. Brunner et al. *Nat. Commun.* **4**, 1364 (2013).
[17] K. Vandoorne et al. *Nat. Commun.* **5**, 3541 (2014).
[18] L. Larger et al. *Opt. Express* **20**, 3241 (2012).
[19] Y. Shen et al. *Nat. Photon.* **11**, 441 (2017).
[20] T. W. Hughes et al. *Optica* **5**, 864 (2018).
[21] T. Ferreira de Lima et al. *J. Lightwave Technol.* **37**, 1515 (2019).
[22] A. N. Tait et al. *J. Lightwave Technol.* **32**, 4029 (2014).
[23] A. N. Tait et al. *Sci. Rep.* **7**, 7430 (2017).
[24] A. N. Tait et al. *Opt. Express* **26**, 26422 (2018).
[25] P. R. Prucnal et al. *Adv. Opt. Photon.* **8**, 228 (2016).
[26] A. N. Tait et al. arXiv preprint arXiv:1812.11898 (2018).