Digital Optical Neural Networks for Large-Scale Machine Learning

Digital Optical Neural Networks for Large-Scale Machine Learning

Liane Bernstein[1], **Alexander Sludds**[1], **Ryan Hamerly**[1], **Vivienne Sze**[1], **Joel Emer**[1,2], **and Dirk Englund**[1]

[1]*Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA*
[2]*NVIDIA, 2 Technology Park Drive, Westford, MA 01886, USA*

lbern@mit.edu

Abstract: We propose a digital incoherent optical neural network architecture using the passive data routing and copying capabilities of optics for artificial neural network acceleration. We demonstrate a proof-of-concept experiment and analyze optimal use cases.

Artificial deep neural networks (DNNs) have revolutionized many fields, including classification, translation and prediction [1]. DNNs' recent surge in popularity is chiefly due to improvements in accuracy achieved thanks to the availability of larger datasets and more compute power. A central challenge now is to reduce energy consumption and increase throughput by developing custom hardware [2]. In datacenters today, a significant amount of DNN tasks revolve around matrix multiplication [3], where the bottlenecks are in data movement and memory access [2].

Optical neural networks (ONNs) have been proposed for efficient matrix multiplication by harnessing the high-speed, low-energy data routing capabilities of optics [4, 5]. However, scaling up the number of neurons in a reconfigurable architecture remains a challenge for ONNs. We have proposed a large-scale, reconfigurable ONN based on homodyne detection (HD-ONN) [6]. By using a combination of massive optical fan-out and interference-based photoelectric multiplication, we estimated that the HD-ONN could achieve potential orders-of-magnitude energy savings and increased throughput compared with state-of-the-art, all-electronic processors. However, error buildup may limit the depth of this system, as well as the other analog ONNs mentioned above [4, 5].

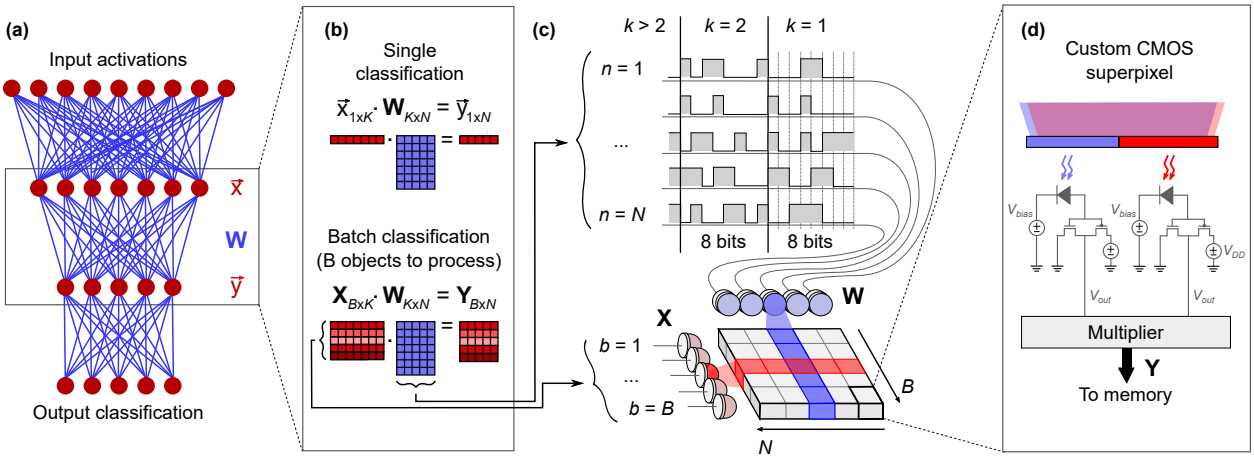

Fig. 1. (a) Fully-connected deep neural network (FC-NN). (b) Batch processing of FC-NN. (c) Example implementation of the digital optical neural network, where bits are streamed in serially, and data delivery to multipliers is achieved with passive optical fan-out. (d) Custom CMOS receiver.

We present a digital ONN (DONN) that solves this error propagation problem with digital encoding into incoherent light for high-efficiency matrix multiplication. As illustrated in Fig. 1, optical elements deliver data from memory to multipliers, where each activation (weight) is passively copied N (B) times, thus greatly reducing data transfer and copying by electrical wire. Contrary to the analog ONNs described above, multiply-accumulate operations (MACs) are performed electronically. To demonstrate the feasibility of optical fan-out, we performed a proof-of-concept experiment shown in Fig. 2. As stand-ins for μLED arrays, a red (blue) LED illuminated one row (column) of a DMD with 10.8 μm-length mirrors to transmit the activations (weights). A cylindrical lens then fanned out each source pixel to a full column (row) of a camera with 10 μm-length square pixels. We ran a custom 3-layer fully-connected network on

the DONN, and found that there was no loss of classification accuracy on the MNIST dataset with respect to a GPU.

In the DONN, potential energy savings arise from the interconnect costs. Interconnect energy per bit is the sum of the energy required to flip a transistor gate and to charge the parasitic capacitances [7]:

$$E = \frac{1}{4}(C_{\text{transistor}} + C_{\text{photodetector}} + \frac{C_{\text{wire}}}{\mu\text{m}} \cdot L_{\text{wire}}) \cdot V_{DD}^2 \tag{1}$$

The experimental electronic values and theoretical approximations for the optical components are reported in Table 1 (shot and thermal noise are negligible here [7], and we assume the frequency of the photons matches the bandgap of silicon and that there is 100% conversion efficiency). The crossover point where the optical interconnect energy drops below the electrical energy occurs when $L_{\text{wire}} \geq 1\ \mu\text{m}$ $(= C_{\text{photodetector}} / \frac{C_{\text{wire}}}{\mu\text{m}})$.

Table 1. Electronic and Optical Interconnect Energies [7]

	$C_{\text{transistor}}$	$C_{\text{wire}}/\mu\text{m}$	$C_{\text{photodetector}}$	V_{DD}	L_{wire}	E
Electrical	0.05 fF	0.1 fF/μm	-	0.7 V	x μm	0.123(0.05 + 0.1 x) fJ
Optical			0.1 fF		small	0.0184 fJ

The electronic multipliers in a large specialized array are separated by roughly 35 μm [3], where charging a long wire that transports data to a long row of multipliers provides an effective electrical fan-out. In this case, the DONN theoretically provides up to an order of magnitude improvement in interconnect energy. That being said, we recognize that this is an optimistic estimate for the optical energy, and furthermore, that interconnects are not always the largest consumers of energy for small matrix multiplication with regular memory-access patterns. However, as DNN sizes are growing at an exponential rate [8], large-scale processing is beginning to require many clusters of multipliers (or chiplets), where chiplet-to-chiplet communication consumes 0.82-1.75pJ/bit [9]. Here, as well as in DNNs with irregular memory access patterns, the DONN becomes even more advantageous, as the energy consumption takes place in the light generation and detection as opposed to the distance travelled.

In conclusion, we have presented a digital incoherent optical neural network and analyzed its energy consumption compared to all-electronic DNN hardware, demonstrating its potential energy efficiency gains at a large scale. Furthermore, as shown in a proof-of-concept experiment, this architecture does not suffer from analog error propagation.

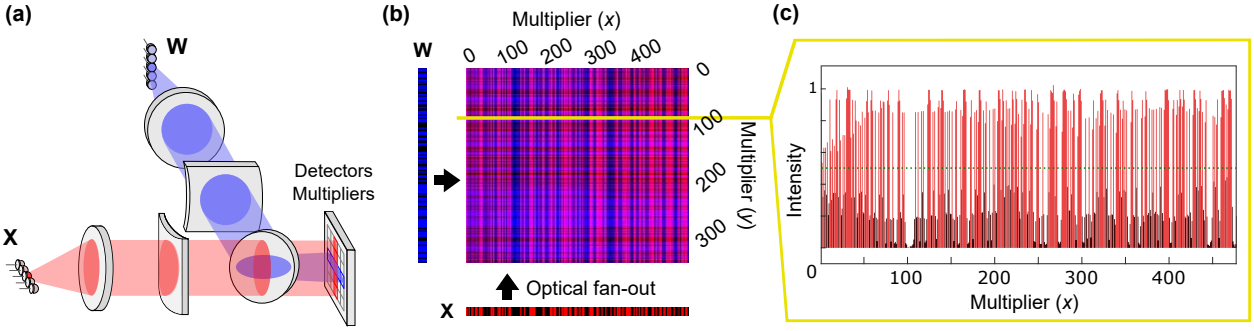

Fig. 2. (a) Experimental implementation of the digital optical neural network with cylindrical lenses for passive fan-out. (b) Image on detector. (c) One line of image showing pixels classified as 1 (red), pixels classified as 0 (black) and threshold (green).

References

1. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**, 436–444, (2015).
2. V. Sze, Y.-H. Chen, T.-J. Yang, J. Emer, Proc. IEEE **105**, 2295-2329, (2017).
3. N. P. Jouppi, C. Young, N. Patil, D. Patterson et al., ISCA, 1-12, (2017).
4. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu et al., Nature Photonics **11**, 441–446, (2017).
5. A. N. Tait, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, J. Light. Technol. **32**, 4029-4041, (2014).
6. R. Hamerly, L. Bernstein, A. Sludds, M. Soljacic, D. Englund, Phys. Rev. X **9**, 021032, (2019).
7. D. A. B. Miller, J. Light. Technol. **35**, 346-396, (2017).
8. D. Amodei, D. Hernandez, https://openai.com/blog/ai-and-compute, (2018).
9. Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer et al., MICRO, 14-27, (2019).