Progress on photonic tensor processors based on time multiplexing and photoelectric multiplication

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Progress on photonic tensor processors based on time multiplexing and photoelectric multiplication

Hamerly, Ryan, Sludds, Alexander, Bernstein, Liane, Sze, Vivienne, Emer, Joel, et al.

Ryan Hamerly, Alexander Sludds, Liane Bernstein, Vivienne Sze, Joel Emer, Marin Soljacic, Dirk Englund, "Progress on photonic tensor processors based on time multiplexing and photoelectric multiplication," Proc. SPIE 11680, Physics and Simulation of Optoelectronic Devices XXIX, 116800E (5 March 2021); doi: 10.1117/12.2576990

SPIE.

Event: SPIE OPTO, 2021, Online Only

Invited Paper

Progress on photonic tensor processors based on time multiplexing and photoelectric multiplication

Ryan Hamerly[1,2], Alexander Sludds[1], Liane Bernstein[1], Vivienne Sze[1], Joel Emer[3,4], Marin Soljacic[1], and Dirk Englund[1]

[1]Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, MA 02139, USA
[2]NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, CA 94085, USA
[3]NVIDIA, 2 Technology Park Drive, Westford, MA 01886, USA
[4]Computer Science and Artificial Intelligence Laboratory, MIT, 32 Vassar St, Cambridge, MA 02139, USA

ABSTRACT

Optical approaches to machine learning rely heavily on programmable linear photonic circuits. Since the performance and energy efficiency scale with size, a major challenge is overcoming scaling roadblocks to the photonic technology. Recently, we proposed an optical neural network architecture based on coherent detection. This architecture has several scaling advantages over competing approaches, including linear (rather than quadratic) chip-area scaling and constant circuit depth. We review the fundamental and technological limits to the energy consumption in this architecture, which shed light on the quantum limits to analog computing, which are distinct from the thermodynamic (e.g. Landauer) limits to digital computing. Lastly, we highlight a recent "digital" implementation of our architecture, which sheds light on the scaling challenges associated with controlling aberrations in the free-space optical propagation.

Keywords: Optical neural network, photonic integrated circuit, machine learning, homodyne detection, cylindrical optics

1. INTRODUCTION

Driven by the recent success of deep learning, the growing computational demand of deep neural networks has motivated development of special-purpose hardware accelerators.[1,2] These accelerators have heretofore been based on digital electronic architectures and are therefore limited by memory access and interconnect energies;[3] consequently, there has been a resurgence of interest in photonic approaches, where the mathematical operations are mapped to the dynamics of optical propagation, i.e. programmable linear optics and nonlinearity. To date, photonic approaches fall into two categories: Free-space systems[4,5] (e.g. diffraction, Fourier optics) boast large numbers of neurons, but suffer limited connectivity. At the other extreme, on-chip approaches based on wavelength multiplexing[6] or beamsplitter meshes[7] can achieve programmable all-to-all coupling, but chip-area constraints make scaling to large numbers of neurons very challenging.

Recently, we proposed a new scheme based on homodyne (coherent) detection.[8] A deep neural network is represented as a sequence of layers (Fig. 1(a)), where each layer is composed of a (linear) tensor product $\vec{x} \to A\vec{x}$ and a (nonlinear) activation function $x_i \to f(x_i)$. Fig. 1(b) schematically illustrates the optically-accelerated tensor core, which performs these operations. The input vector $\vec{x}^{(k)}$ is encoded onto a pulse train, which is fanned out to an array of homodyne detectors. Each detector (inset of Fig. 1(b)) computes the product between $\vec{x}^{(k)}$ and a row of $A^{(k)}$, both encoded on optical pulse trains, by homodyning and time integration. The accumulated charge on the homodyne detector is given by:

$$Q_i = \frac{2\eta e}{\hbar\omega} \int \mathrm{Re}\big[E^{(\mathrm{in})}(t)^* E_i^{(\mathrm{wt})}(t)\big]\mathrm{d}t \propto \sum_j A_{ij}x_j \tag{1}$$

Further author information: (Send correspondence to R.H.)
R.H.: E-mail: rhamerly@mit.edu

Physics and Simulation of Optoelectronic Devices XXIX, edited by Bernd Witzigmann, Marek Osiński, Yasuhiko Arakawa, Proc. of SPIE Vol. 11680, 116800E · © 2021 SPIE
CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2576990

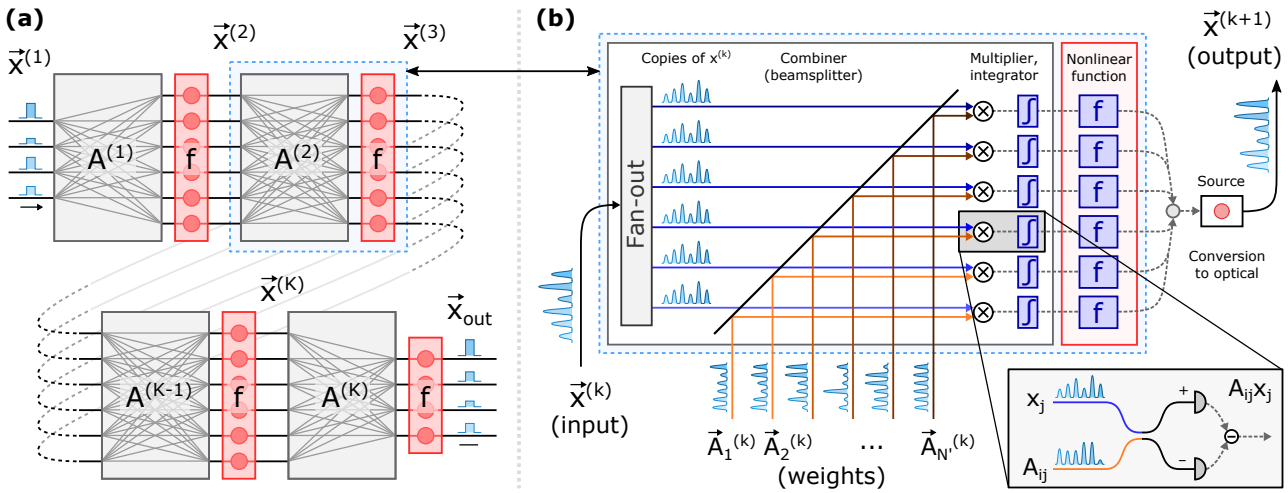

Figure 1. Optical neural network based on coherent detection. (a) A deep neural network is decomposed into a sequence of layers, each consisting of a linear matrix-vector multiplication and a nonlinear activation function. (b) Matrix multiplication realized with time multiplexing and coherent detection. Activations, encoded as a pulse train, are broadcast to an array of homodyne detectors that implement the matrix-vector product. Nonlinearity can be performed subsequently in the electronic domain. Adapted from Ref.[8]

The accumulation requires N time steps, where the pulse rate is limited by the speed of the modulator (which will lead to crosstalk if the rate is too fast[9]).

The detectors here function as quantum photoelectric multipliers, producing a photocurrent proportional to the tensor product $A\vec{x}$. The output is sent through a nonlinearity, serialized, and converted to optical with a modulator. Unlike previous approaches,[6,7] here the weights are encoded optically, allowing the network to be reprogrammed on the fly. Optical weight encoding, plus time multiplexing, significantly reduces the number of photonic components, and therefore chip are, of in this scheme: only $O(N)$ photonic components (modulators, beamsplitters, detectors) are required, in contrast to weight-stationary[1] approaches[6,7] which typically require $O(N^2)$ with an $O(N)$ circuit depth. This greatly reduces the required chip area, a major constraint given the moderate size of low-loss photonic components, and allows scaling to large systems with $N, N' \geq 1000$.

2. DEEP LEARNING AT THE STANDARD QUANTUM LIMIT

A key figure of merit for neural-network accelerators is energy consumption, which limits performance on modern processors due to overheating.[3] Energy consumption can be measured as energy per multiply-and-accumulate (MAC) E_{mac}. For CPUs and GPUs, the figure is about 20 pJ/MAC,[10] although application-specific integrated circuits (ASICs) with reduced precision can push this number down to 1 pJ/MAC, which is considered state of the art.[1,2]

The homodyne-based optical neural network has two limits to E_{mac}: (1) a fundamental standard quantum limit (SQL) set by quantum fluctuations, and (2) limits set by input / output energy consumption, which will decrease as technology improves. The SQL is one of a number of noise-based limits to photonic device performance. It has long been recognized that noise processes can limit the performance of optical logic gates[11–13] and oscillators[14,15] as well as all-optical photodetectors[16] and temperature sensors.[17] Here, noise arises because the photoelectric effect is a stochastic process: quantum-limited detector shot noise will degrade performance at the low-energy (few-photon) level. To study this effect, we perform simulations of the optical neural network trained on the MNIST dataset (Fig. 2(a)) in the presence of shot noise. The input-output relation, including shot noise, for a neural-network layer is given by:

$$x_i^{(k+1)} = f\left(\sum_j A_{ij}^{(k)} x_j^{(k)} + w_i^{(k)} \frac{\|A^{(k)}\| \|x^{(k)}\|}{\sqrt{NN'}} \frac{1}{\sqrt{n_{\text{mac}}}}\right) \tag{2}$$

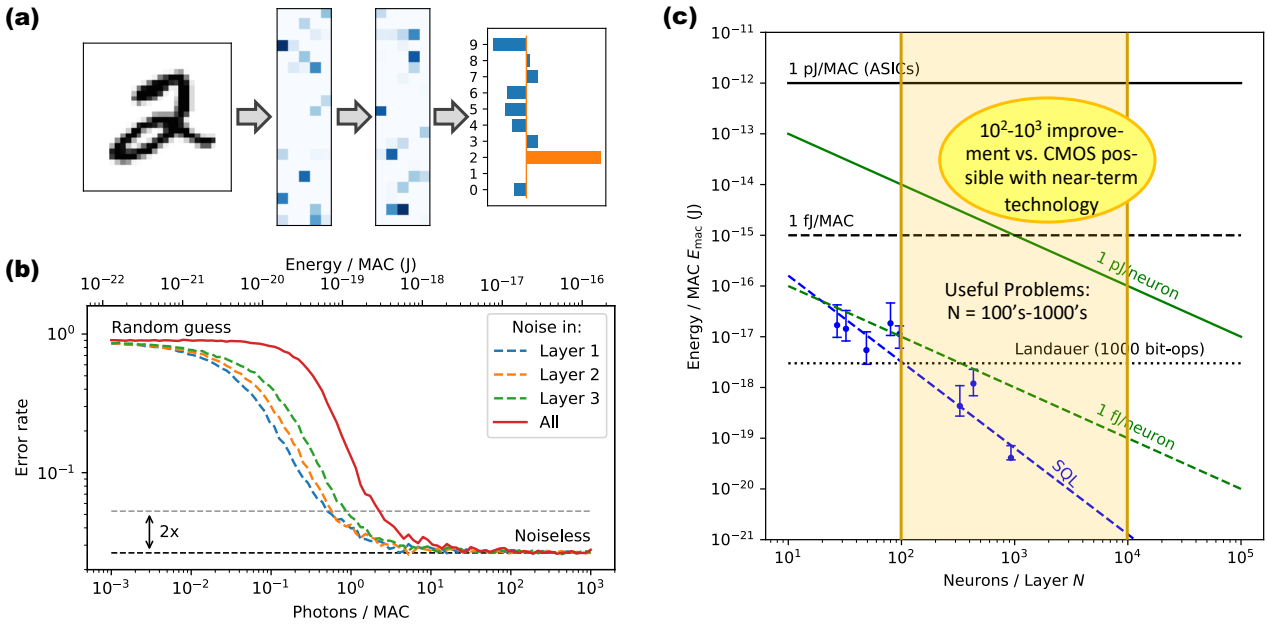

Figure 2. Limits to energy consumption of optical neural network. (a) Illustration of MNIST digit classification with a two-layer perceptron. (b) Error rate due to quantum (shot) noise as a function of optical energy per MAC E_{mac}. (c) Quantum limit relative to other figures for energy consumption. Adapted from Ref.[8]

where $f(\cdot)$ is the activation function, N (N') are the number of input (output) neurons, $w_i^{(k)}$ is a Gaussian random variable with unit variance, and n_{mac} is the number of photons per MAC (so the optical contribution to E_{mac} is $(\hbar\omega)n_{\mathrm{mac}}$). Under shot noise, SNR is proportional to the photon flux. There are two regions of interest, sketched in Fig. 2(b): a noiseless regime $E_{\mathrm{mac}} \gg \hbar\omega$ and a random-guess regime $E_{\mathrm{mac}} \ll \hbar\omega$. The SQL is defined as the crossover point, where the error rate begins to increase significantly compared to its noiseless value. The SQL depends on the neural network being studied. Fig. 2(c) plots the SQL calculated for a number of neural networks trained on the MNIST dataset. For comparison, the Landauer limit[18] for an irreversible digital computer is $N_{\mathrm{gate}} \times k_B T \log(2)$, where N_{gate} is the gate count for the fused multiply-add circuit (since all gates are irreversible). This evaluates to 3 aJ for 32-bit arithmetic ($N \approx 1000$), which is above the SQL for the larger neural networks in Fig. 2(c). This suggests that it is theoretically possible for the optical neural network to operate below the Landauer limit. Since the optical neural network relies on (reversible) optical interference and is an analog system, it does not satisfy the assumptions of Landauer's principle, so beating the Landauer limit is not in contradiction with the laws of thermodynamics.

For near-term devices, the electrical contribution to E_{mac} will dominate. This figure is governed by the readout energy of the detector electronics, the electronic nonlinearity, and the energy required to drive the modulator. A simple approach with existing technology is to amplify and digitize the photocurrent, apply the nonlinearity in digital logic, and convert to optical with a modulator. With existing components,[19,20] this leads to an energy per neuron at the picojoule scale, so $E_{\mathrm{mac}} \sim (1/N)$pJ (solid green curve in Fig. 2(c)). This number can be reduced with advances in modulator and detector technology: to realize on-chip interconnects, femtojoule-scale detectors, tightly integrated to modulators and CMOS logic, are under active development.[21,22] This would push the energy figure to $E_{\mathrm{mac}} \sim (1/N)$fJ (dashed green curve), at which point the SQL becomes a relevant bound to the neural network's performance. Even with near-term energy figures, however, we see the potential to reduce E_{mac} by several orders of magnitude compared to state-of-the-art electronics.

3. GEMM AND CONVOLUTIONS

While the optical unit in Fig. 1(b) performs a matrix-vector product, in practice neural networks achieve high performance with weight reuse (either natively in convolutional layers or through batching). Thus, a practical

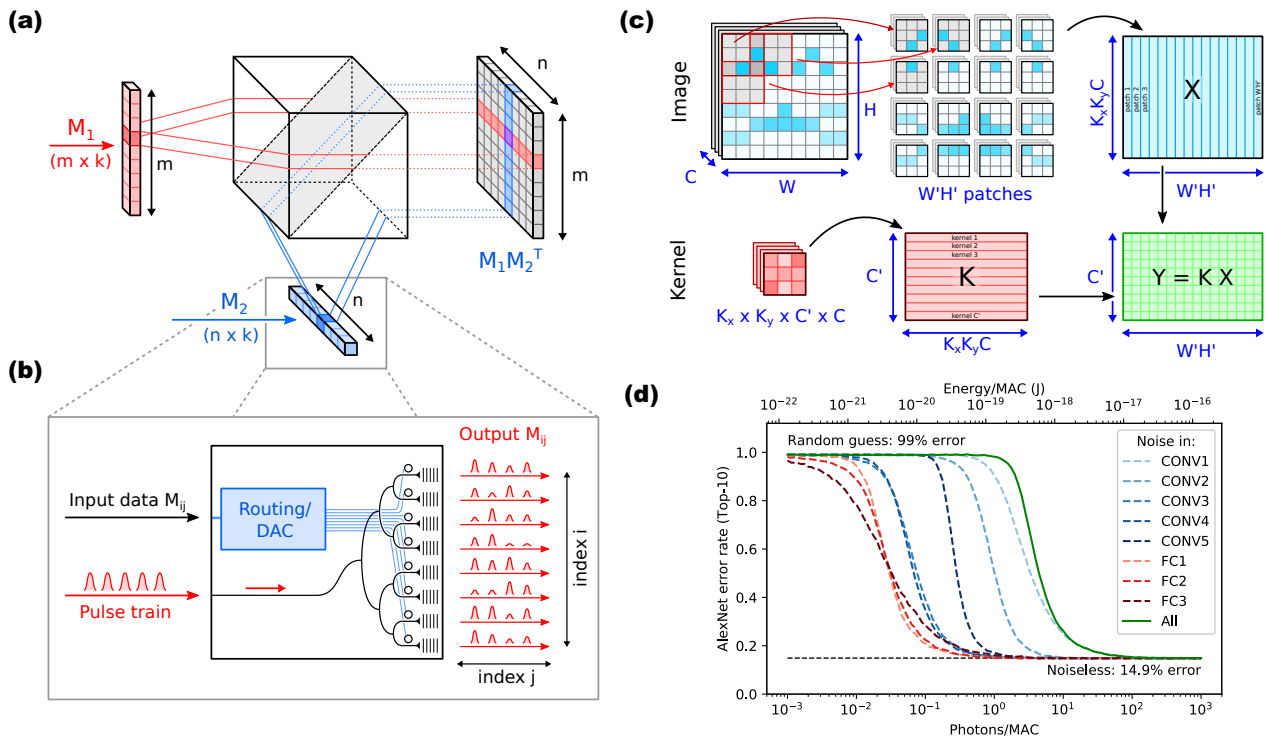

Figure 3. Matrix-matrix optical accelerator. (a) Free-space architecture showing fan-out of optical signals via cylindrical lenses (not shown). (b) Schematic of the implementation of a 1D transmitter array. (c) Convolution recast as a matrix-matrix multiplication via patching. (d) Shot-noise limit for ImageNet classification via AlexNet. Adapted from Ref.[8]

system must implement a general matrix-*matrix* product (GEMM). The coherent-detection architecture can be parallelized to implement GEMM by routing the light out of plane (Fig. 3(a)). The inputs are two matrices $(M_1)_{m\times k}$ and $(M_2)_{n\times k}$, encoded into optical signals on the 1D red (blue) transmitter arrays (Fig. 3(b)), and mapped with cylindrical lenses to rows (columns) of the 2D detector array. From the accumulated charge at each pixel, one can extract the matrix elements of the product $(M_1 M_2^T)_{m\times n}$. This performs $O(mnk)$ MACs at an energy cost that scales as $O(mk)+O(mn)+O(nk)$; note also that the number of modulators still scales only linearly with matrix dimension (a quadratic number of detectors are required, but detector arrays can pack at high densities).

In addition to fully-connected layers, it is also possible to run convolutional layers on the optical GEMM unit by employing a "patching" technique,[23] where the image, divided into patches, is recast as a matrix X (which contains redundant data when patches overlap). The kernel elements can also be rearranged as a matrix K, and the convolution is equivalent to computing the matrix product $Y = KX$ (Fig. 3(c)). Simulations based on the optical architecture of Fig. 3(a) reveal that shot noise also places a limit on the performance of optically accelerated convolutional neural networks. As a benchmark example, consider AlexNet,[24] the first deep neural network to perform competitively at the ImageNet Challenge,[25] which consists of five convolutional layers and three fully-connected layers. Fig. 3(d) plots the network accuracy as a function of $E_{\rm mac}$ in the presence of shot noise. Again, we see the emergence of a quantum limit to classification energy. The SQL obtained for AlexNet ($n_{\rm mac} \gtrsim 20$ or $E_{\rm mac} \gtrsim 3\,$aJ) is slightly larger than that from the MNIST networks in Fig. 2(c).

A key challenge in scaling this scheme will be correcting for optical aberrations that lead to inter-pixel crosstalk. Zemax(R) simulations suggest that crosstalk can be kept to tolerable levels with relatively simple optical designs.[8] Recently, we experimentally demonstrated am incoherent "digital" version of the optical matrix-matrix scheme where data is encoded in bit streams and digital multiplication substitutes for coherent detection.[26,27] For array sizes $N \gtrsim 200$, crosstalk was not found to degrade neural network accuracy. This

suggests that the optical fan-out scheme in Fig. 3(a) is scalable to large arrays without signal degradation.

4. CONCLUSION

We have presented a new architecture for optically accelerated deep learning that is scalable to large problems and can operate at high speeds with low energy consumption. Our approach takes advantage of the opto-electronic nonlinearity in the photoelectric effect, via the relation $I \propto |E|^2$, to compute the desired matrix products without need of an all-optical nonlinearity. Neural-network weights are encoded optically, allowing the network to be rapidly reprogrammed, an important feature for training. Time-multiplexing of data and weights leads to a considerable reduction of the hardware complexity, the number of photonic components scaling linearly with problem size rather than quadratically as in competing nanophotonic schemes.[6,7] This allows scaling to larger problem sizes ($N \geq 1000$) commonly used in neural network layers. Significant reductions in energy consumption are possible compared to state-of-the-art digital electronics, and the quantum limit, set by shot noise, is low enough to suggest that sub-Landauer performance is theoretically possible in such a device. In addition to neural networks, such hardware may find application in combinatorial optimization problems such as Ising and SAT, where optical approaches have proven competitive against both classical heuristics[28] and quantum annealing.[29]

ACKNOWLEDGMENTS

R.H.: IC Postdoctoral Research Fellowship at MIT, administered by ORISE through the U.S. DoE / ODNI. A.S.: NSF Graduate Research Fellowship Program under grant no. 1122374. L.B.: NSERC Postgraduate Scholarship. This research was funded by NTT Research Inc., U.S. ARO through ISN (no. W911NF-18-2-0048), NSF E2CDA (no. 1640012), and NSF EAGER (no. 1946967).

REFERENCES

[1] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE* **105**(12), pp. 2295–2329, 2017.

[2] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*, pp. 1–12, IEEE, 2017.

[3] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pp. 10–14, IEEE, 2014.

[4] E. G. Paek and D. Psaltis, "Optical associative memory using Fourier transform holograms," *Optical Engineering* **26**(5), p. 265428, 1987.

[5] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science* **361**(6406), pp. 1004–1008, 2018.

[6] A. N. Tait, T. F. Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports* **7**(1), p. 7430, 2017.

[7] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics* **11**(7), p. 441, 2017.

[8] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Physical Review X* **9**(2), p. 021032, 2019.

[9] R. Hamerly, A. Sludds, L. Bernstein, M. Prabhu, C. Roques-Carmes, J. Carolan, Y. Yamamoto, M. Soljačić, and D. Englund, "Towards large-scale photonic neural-network accelerators," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 22–8, IEEE, 2019.

[10] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the future of parallel computing," *IEEE Micro* (5), pp. 7–17, 2011.

[11] S. Carter, "Quantum theory of nonlinear fiber optics: Phase-space representations," *Physical Review A* **51**(4), p. 3274, 1995.

[12] C. Santori, J. S. Pelc, R. G. Beausoleil, N. Tezak, R. Hamerly, and H. Mabuchi, "Quantum noise in large-scale coherent nonlinear photonic circuits," *Physical Review Applied* **1**(5), p. 054005, 2014.

[13] R. Hamerly and H. Mabuchi, "Quantum noise of free-carrier dispersion in semiconductor optical cavities," *Physical Review A* **92**(2), p. 023819, 2015.

[14] C. M. Caves, "Quantum-mechanical radiation-pressure fluctuations in an interferometer," *Physical Review Letters* **45**(2), p. 75, 1980.

[15] R. Hamerly and H. Mabuchi, "Optical devices based on limit cycles and amplification in semiconductor optical cavities," *Physical Review Applied* **4**(2), p. 024016, 2015.

[16] C. Panuski, M. Pant, M. Heuck, R. Hamerly, and D. Englund, "Single photon detection by cavity-assisted all-optical gain," *Physical Review B* **99**(20), p. 205303, 2019.

[17] C. Panuski, D. Englund, and R. Hamerly, "Fundamental thermal noise limits for optical microcavities," *Physical Review X* **10**(4), p. 041046, 2020.

[18] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development* **5**(3), pp. 183–191, 1961.

[19] S. Saeedi, S. Menezo, G. Pares, and A. Emami, "A 25 Gb/s 3D-integrated CMOS/silicon-photonic receiver for low-power high-sensitivity optical communication," *Journal of Lightwave Technology* **34**(12), pp. 2924–2933, 2016.

[20] A. H. Atabaki, S. Moazeni, F. Pavanello, H. Gevorgyan, J. Notaros, L. Alloatti, M. T. Wade, C. Sun, S. A. Kruger, H. Meng, *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature* **556**(7701), p. 349, 2018.

[21] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *Journal of Lightwave Technology* **35**(3), pp. 346–396, 2017.

[22] M. Notomi, K. Nozaki, A. Shinya, S. Matsuo, and E. Kuramochi, "Toward fJ/bit optical communication in a chip," *Optics Communications* **314**, pp. 3–17, 2014.

[23] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759* , 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)* **115**(3), pp. 211–252, 2015.

[26] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Freely scalable and reconfigurable optical hardware for deep learning," *Scientific Reports* **11**(1), pp. 1–12, 2021.

[27] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Digital optical neural networks for large-scale machine learning," in *2020 Conference on Lasers and Electro-Optics (CLEO)*, pp. 1–2, IEEE, 2020.

[28] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić, *et al.*, "Accelerating recurrent ising machines in photonic integrated circuits," *Optica* **7**(5), pp. 551–558, 2020.

[29] R. Hamerly, T. Inagaki, P. L. McMahon, D. Venturelli, A. Marandi, T. Onodera, E. Ng, C. Langrock, K. Inaba, T. Honjo, *et al.*, "Experimental investigation of performance differences between coherent ising machines and a quantum annealer," *Science Advances* **5**(5), p. eaau0823, 2019.