


Large-Scale Optical Neural Networks Based on Photoelectric Multiplication

Ryan Hamerly,^{*} Liane Bernstein, Alexander Sludds, Marin Soljačić, and Dirk Englund
Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, Massachusetts 02139, USA

 (Received 12 November 2018; revised manuscript received 21 February 2019; published 16 May 2019)

Recent success in deep neural networks has generated strong interest in hardware accelerators to improve speed and energy consumption. This paper presents a new type of photonic accelerator based on coherent detection that is scalable to large ($N \gtrsim 10^6$) networks and can be operated at high (gigahertz) speeds and very low (subattojoule) energies per multiply and accumulate (MAC), using the massive spatial multiplexing enabled by standard free-space optical components. In contrast to previous approaches, both weights and inputs are optically encoded so that the network can be reprogrammed and trained on the fly. Simulations of the network using models for digit and image classification reveal a “standard quantum limit” for optical neural networks, set by photodetector shot noise. This bound, which can be as low as 50 zJ/MAC, suggests that performance below the thermodynamic (Landauer) limit for digital irreversible computation is theoretically possible in this device. The proposed accelerator can implement both fully connected and convolutional networks. We also present a scheme for backpropagation and training that can be performed in the same hardware. This architecture will enable a new class of ultralow-energy processors for deep learning.

DOI: [10.1103/PhysRevX.9.021032](https://doi.org/10.1103/PhysRevX.9.021032)

Subject Areas: Interdisciplinary Physics,
 Optoelectronics, Photonics

I. INTRODUCTION

In recent years, deep neural networks have tackled a wide range of problems including image analysis [1], natural language processing [2], game playing [3], physical chemistry [4], and medicine [5]. This is not a new field, however. The theoretical tools underpinning deep learning have been around for several decades [6–8]; the recent resurgence is driven primarily by (1) the availability of large training datasets [9] and (2) substantial growth in computing power [10] and the ability to train networks on graphics processing units (GPUs) [11]. Moving to more complex problems and higher network accuracies requires larger and deeper neural networks, which in turn require even more computing power [12]. This motivates the development of special-purpose hardware optimized to perform neural-network inference and training [13].

To outperform a GPU, a neural-network accelerator must significantly lower the energy consumption, since the performance of modern microprocessors is limited by on-chip power [14]. In addition, the system must be fast, programmable, scalable to many neurons, compact, and

ideally compatible with training as well as inference. Application-specific integrated circuits (ASICs) are one obvious candidate for this task. State-of-the-art ASICs can reduce the energy per multiply and accumulate (MAC) from 20 pJ/MAC for modern GPUs [15] to around 1 pJ/MAC [16,17]. However, ASICs are based on CMOS technology and therefore suffer from the interconnect problem—even in highly optimized architectures where data are stored in register files close to the logic units, a majority of the energy consumption comes from data movement, not logic [13,16]. Analog crossbar arrays based on CMOS gates [18] or memristors [19,20] promise better performance, but as analog electronic devices, they suffer from calibration issues and limited accuracy [21].

Photonic approaches can greatly reduce both the logic and data-movement energy by performing (the linear part of) each neural-network layer in a passive, linear optical circuit. This allows the linear step to be performed at high speed with no energy consumption beyond transmitter and receiver energies. Optical neural networks based on free-space diffraction [22] have been reported, but require spatial light modulators or 3D-printed diffractive elements, and are therefore not rapidly programmable. Nanophotonic circuits are a promising alternative [23,24], but the footprint of directional couplers and phase modulators makes scaling to large ($N \geq 1000$) numbers of neurons very challenging. To date, the goal of a large-scale, rapidly reprogrammable photonic neural network remains unrealized.

^{*}rhamerly@mit.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

This paper presents a new architecture based on coherent (homodyne) detection that is fast, low power, compact, and readily scalable to large ($N \gtrsim 10^6$) numbers of neurons. In contrast to previous schemes, here we encode both the inputs and weights in optical signals, allowing the weights to be changed on the fly at high speed. Synaptic connections (matrix-vector products) are realized by the quantum photoelectric multiplication process in the homodyne detectors. Our system is naturally adapted to free-space optics and can therefore take advantage of the massive spatial multiplexing possible in free-space systems [25,26] and the high pixel density of modern focal-plane arrays [27] to scale to far more neurons than can be supported in nanophotonics or electronic crossbar arrays. The optical energy consumption is subject to a fundamental standard quantum limit (SQL) arising from the effects of shot noise in photodetectors, which lead to classification errors. Simulations based on neural networks trained on the Modified NIST (MNIST) dataset [8] empirically show the SQL can be as low as 50–100 zeptojoules (zJ)/MAC. Using realistic laser, modulator, and detector energies, performance at the sub-fJ/MAC level should be possible

with present technology. The optical system can be used for both fully connected and convolutional layers. Finally, backpropagation is straightforward to implement in our system, allowing both inference and training to be performed in the same optical device.

II. COHERENT MATRIX MULTIPLIER

Figure 1 illustrates the device. A deep neural network is a sequence of K layers [Fig. 1(a)], each consisting of a matrix multiplication $\vec{x} \rightarrow A\vec{x}$ (synaptic connections) and an elementwise nonlinearity $x_i \rightarrow f(x_i)$ (activation function); thus the input into the $(k + 1)$ th layer is related to the k th layer input by

$$x_i^{(k+1)} = f\left(\sum_j A_{ij}^{(k)} x_j^{(k)}\right). \quad (1)$$

For a given layer, let N and N' be the number of input and output neurons, respectively. Input (output) data are encoded temporally as N (N') pulses on a single channel as shown in Fig. 1(b). This encoding, reminiscent of

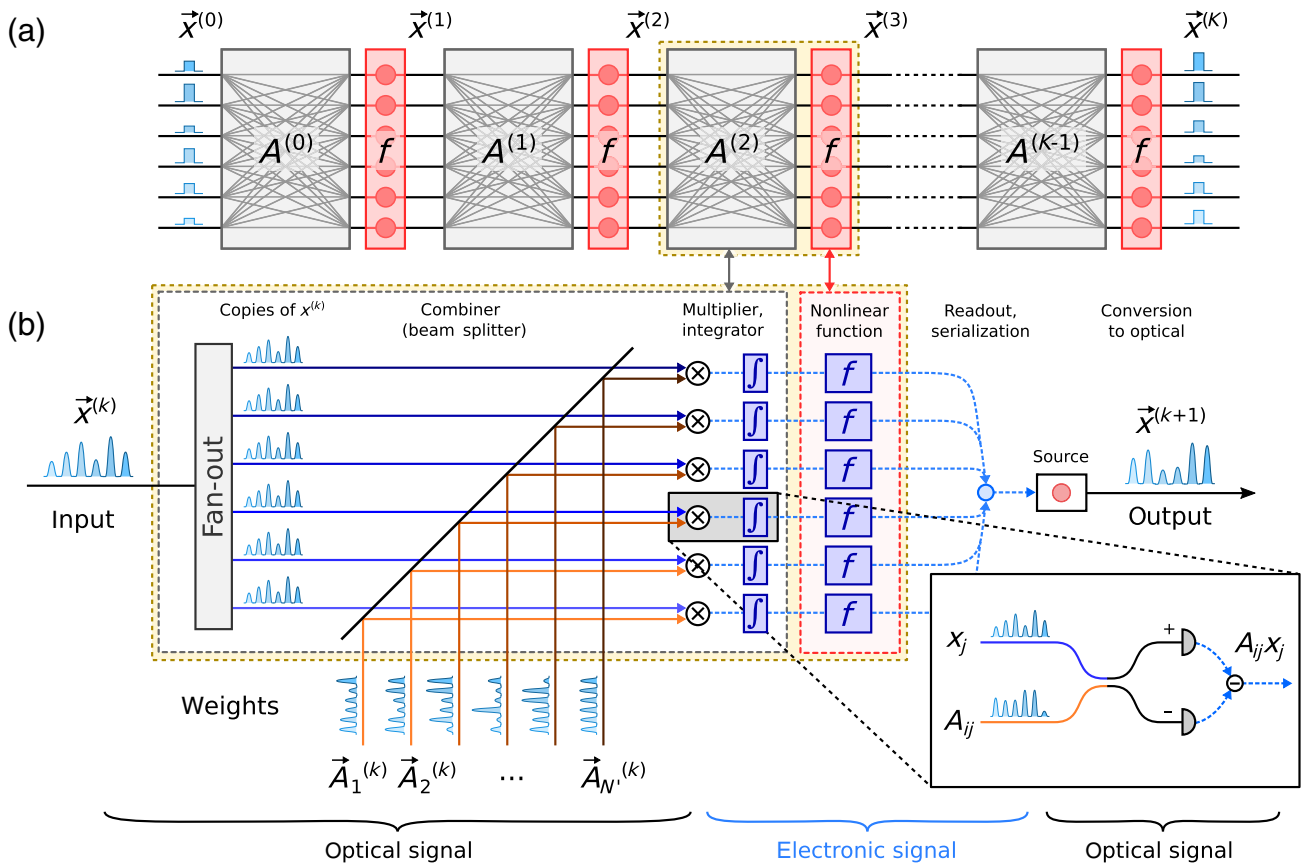


FIG. 1. Schematic diagram of a single layer of the homodyne optical neural network. (a) Neural network represented as a sequence of K layers, each consisting of a matrix-vector multiplication (gray) and an elementwise nonlinearity (red). (b) Implementation of a single layer. Matrix multiplication is performed by combining input and weight signals and performing balanced homodyne detection (inset) between each signal-weight pair (gray box). For details on experimental implementation see Sec. S1 of Supplemental Material [31]. The resulting electronic signals are sent through a nonlinear function (red box), serialized, and sent to the input of the next layer.

the coherent Ising machine [28–30], contrasts with other approaches used for neural networks, which encode inputs in separate spatial channels [22–24]. As there are NN' weights for an $N' \times N$ fully connected matrix, the weights enter on N' separate channels, each carrying a single matrix row encoded in time. Input data are optically fanned out to all N' channels, and each detector functions as a quantum photoelectric multiplier, calculating the homodyne product between the two signals [inset of Fig. 1(b)]. As long as both signals are driven from the same coherent source and the path-length difference is less than the coherence length, the charge Q_i accumulated by homodyne receiver i is

$$Q_i = \frac{2\eta e}{\hbar\omega} \int \text{Re}[E^{(\text{in})}(t)^* E_i^{(\text{wt})}(t)] dt \propto \sum_j A_{ij} x_j. \quad (2)$$

Here $E^{(\text{in})}(t)$ and $E_i^{(\text{wt})}(t)$ are the input and weight fields for receiver i , which are taken to be sequences of pulses with amplitudes proportional to x_j and A_{ij} , respectively ($x_j, A_{ij} \in \mathbb{R}$). Thus, each receiver performs a vector-vector product between \vec{x} and a row \vec{A}_i of the weight matrix; taken together, the N' electronic outputs give the matrix-vector product $A\vec{x}$. Fields are normalized so that power is given by $P(t) = |E(t)|^2$, and η is the detector efficiency. A serializer reads out these values one by one, applies the nonlinear function $f(\cdot)$ in the electrical domain, and outputs the result to a modulator to produce the next layer's inputs.

The balanced homodyne detector in Fig. 1(b) (inset) combines the advantages of optics and electronics: it can process data encoded at extremely high speeds, limited only by the bandwidth of the beam splitter (\gtrsim THz) and the (optical) bandwidth of the photodetectors (typically $\gtrsim 100$ nm, or $\gtrsim 10$ THz). The electrical bandwidth can be much slower, since only the integrated charge is measured. Finally, the present scheme avoids the need for low-power nonlinear optics that is a major stumbling block in all-optical logic [32]: since the output is electrical, the dot product $A_{ij}x_j$ can be computed at extremely low power (sub-fJ/MAC) using standard nonresonant components (photodiodes) that are CMOS compatible and scalable to arrays of millions.

Previous approaches used optoelectronics (photodiodes, lasers, amplifiers) both to sum neuron inputs [24,33] and to generate nonlinearity or spiking dynamics [34–38]; here, thanks to the optical weight encoding, the synaptic weighting itself is performed optoelectronically.

Coherent detection greatly simplifies the setup compared to alternative approaches. With a given set of weight inputs, the network in Fig. 1(b) requires N input pulses and N' detectors to perform a matrix-vector operation with NN' MACs, performing an operation that should scale quadratically with size using only linear resources. This is in contrast to electrical approaches that require quadratic resources (NN' floating-point operations total).

The (optical) energy consumption of nanophotonic systems [23,24] also scales linearly for the same operation; however, the circuit is much more complex, requiring $O(NN')$ tunable phase shifters [39,40] or ring resonators [24], which becomes very challenging to scale beyond several hundred channels and may be sensitive to propagation of fabrication errors. The main caveat to our system is the need to generate the weights in the first place, which imposes an energy cost that does scale quadratically. However, in many cases (particularly in data centers) neural networks are run simultaneously over large batches of data, so with appropriate optical fan-out, the cost of the weights can be amortized over many clients. Put another way, running the neural network on data with batch size B , we are performing a matrix-matrix product $Y_{N' \times B} = A_{N' \times N} X_{N \times B}$, which requires $N'NB$ MACs, with an energy cost that should scale as $O(N'N) + O(N'B) + O(NB)$ rather than $O(N'NB)$.

III. DEEP LEARNING AT THE STANDARD QUANTUM LIMIT

As energy consumption is a primary concern in neuro-morphic and computing hardware generally [14], an optical approach must outperform electronics by a large factor to justify the investment in a new technology. In addition, optical systems must show great potential for improvement, ideally by many orders of magnitude, to allow continued scaling beyond the physical limits of Moore's law. Thus two considerations are relevant: (1) the fundamental, physical limits to the energy consumption and (2) the energy consumption of a practical, near-term device using existing technology.

The fundamental limit stems from quantum-limited noise. In an electrical signal, energy is quantized at a level $E_{\text{el}} = h/\tau_{\text{el}}$, where $\tau_{\text{el}} \sim 10^{-10}$ s is the signal duration. Optical energy is quantized at a level $E_{\text{opt}} = h/\tau_{\text{opt}}$, where $\tau_{\text{opt}} \equiv c/\lambda \sim (2-5) \times 10^{-15}$ s, which is 10^4-10^5 times higher. As a result, $E_{\text{opt}} \gg kT \gg E_{\text{el}}$, and electrical signals can be treated in a classical limit governed by thermal noise, while optical signals operate in a zero-temperature quantum limit where vacuum fluctuations dominate. These fluctuations are read out on the photodetectors, where the photoelectric effect [41] produces a Poisson-distributed photocurrent [42,43]. While the photocurrents are subtracted in homodyne detection, the fluctuations add in quadrature, and Eq. (1) is replaced by (see Sec. S3 of Supplemental Material for derivation and assumptions [31])

$$x_i^{(k+1)} = f\left(\sum_j A_{ij}^{(k)} x_j^{(k)} + w_i^{(k)} \frac{\|A^{(k)}\| \|x^{(k)}\| \sqrt{N}}{\sqrt{N^2 N'} \sqrt{n_{\text{MAC}}}}\right). \quad (3)$$

Here the $w_i^{(k)} \sim N(0, 1)$ are Gaussian random variables, $\|\cdot\|$ is the L_2 norm, and n_{MAC} is the number of photons per MAC, related to the total energy consumption of the layer by $n_{\text{tot}} = NN'n_{\text{MAC}}$.

The noise term in Eq. (3) scales as $n_{\text{MAC}}^{-1/2}$, and therefore the signal-to-noise ratio (SNR) of each layer will scale as $\text{SNR} \propto n_{\text{MAC}}$. Since noise adversely affects the network’s performance, one expects that the energy minimum should correspond to the value of n_{MAC} at which the noise becomes significant. To quantify this statement, we perform benchmark simulations using a collection of neural networks trained on the MNIST (digit recognition) dataset. While MNIST digit classification is a relatively easy task [13], the intuition developed here should generalize to more challenging problems. Data for two simple networks are shown in Figs. 2 and 3, both having a three-layer, fully connected topology [Fig. 2(a)]. In the absence of noise, the networks classify images with high accuracy, as the example illustrates [Fig. 2(b)].

As Fig. 3 shows, the error rate is a monotonically decreasing function of n_{MAC} . The two asymptotic limits correspond to the noiseless case ($n_{\text{MAC}} \rightarrow \infty$, which returns the network’s canonical accuracy) and the noise-dominated case ($n_{\text{MAC}} \rightarrow 0$, where the network is making a random guess). Of interest to us is the cutoff point, loosely defined as the lowest possible energy at which the network returns close to its canonical accuracy (for example, within a factor of $2\times$, see dashed lines in Fig. 3). This is around 0.5–1 aJ (5–10 photons) for the small network (inner layer size $N = 100$), and 50–100 zJ (0.5–1 photon) for the large network (inner layer size $N = 1000$). (Note that this is per MAC; the number of photons per detector Nn_{MAC} is typically $\gg 1$.) This bound stems from the standard quantum limit: the intrinsic uncertainty of quadrature measurements on coherent states [44], which is

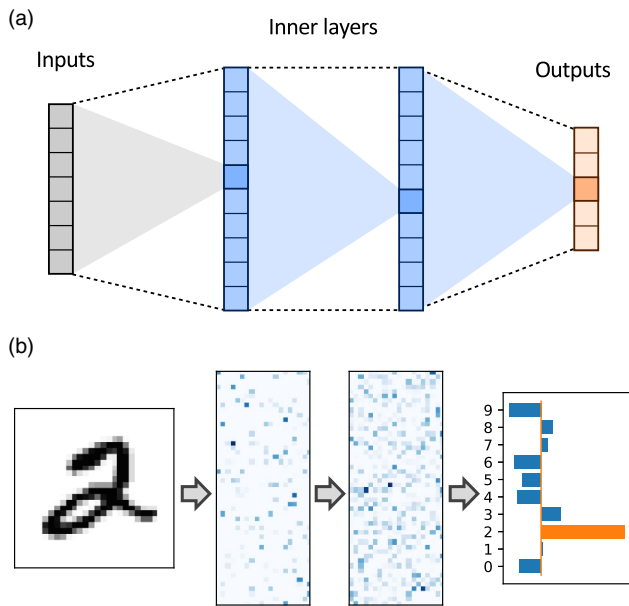


FIG. 2. (a) Illustration of a three-layer neural network with full connectivity. (b) MNIST image classified by network (size 784 → 1000 → 1000 → 10).

temperature and device independent. This should be viewed as an absolute lower bound for the energy consumption of neural networks of this type; although the use of squeezed light allows one to reach sensitivity below the SQL [45,46], this requires squeezing all inputs (including vacuum inputs in optical fan-out), which will likely lead to a net increase in overall energy consumption [squeezing injects an average of $\sinh^2(\eta)$ photons per pulse, where η is the squeezing parameter [42], which will substantially increase n_{MAC}].

The SQL is network dependent, and not all layers contribute equally. For each MAC, we have $\text{SNR} \propto n_{\text{MAC}}$; however, the signal adds linearly while the errors add in quadrature. As a result, the larger network is more resilient to individual errors because each output is averaging over more neurons. Moreover, the solid curves in Fig. 3 are restricted to the case when n_{MAC} is the same for all layers. The dashed lines show the error rate in a fictitious device where quantum-limited noise is present only in a particular layer. For the large network, a smaller n_{MAC} can be tolerated in the second layer, suggesting that better performance could be achieved by independently tuning the energy for each layer. Moreover, just as neural networks can be “codesigned” to achieve high accuracy on limited bit-precision hardware [13], changes to the training procedure (e.g., injecting noise to inner layers, a technique used to reduce generalization error [47,48]) may further improve performance at low powers.

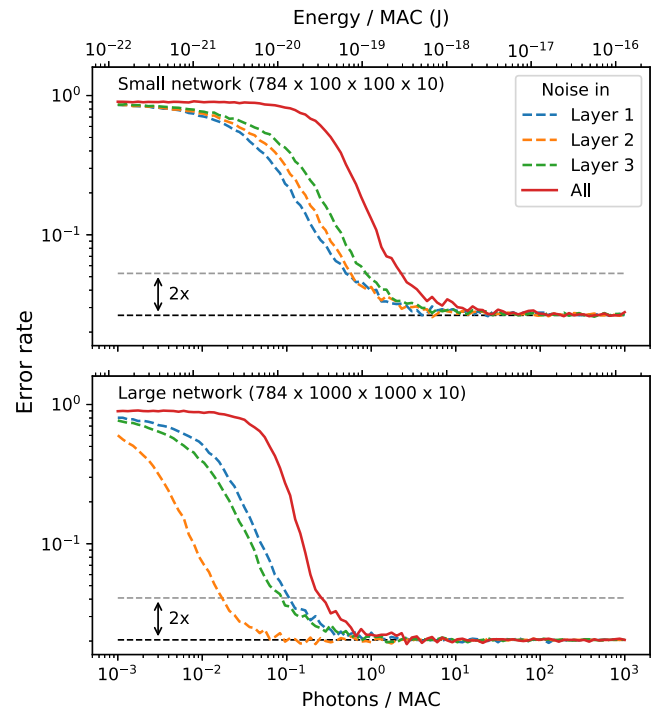


FIG. 3. MNIST digit classification. Error rate for neural-network inference as a function of photons per MAC n_{MAC} (equivalently energy $E_{\text{MAC}} = (hc/\lambda)n_{\text{MAC}}$; here, $\lambda = 1.55 \mu\text{m}$).

Quantum limits to computational energy efficiency in photonics are not unique to neural networks. In digital photonic circuits based on optical bistability [49], vacuum fluctuations lead to spontaneous switching events that limit memory lifetime and gate accuracy [50,51]. However, these effects require bistability at the attojoule scale [50,52], which is well out of the reach of integrated photonics (although recent developments are promising [53–55]). By contrast, neural networks are analog systems, so the quantum fluctuations set a meaningful limit on efficiency even though no attojoule-scale optical nonlinearities are employed.

IV. ENERGY BUDGET

Viewing the neural network as an analog system with quantum-limited performance shifts the paradigm for comparing neural networks. Figure 4(a) shows the standard approach: a scatter plot comparing error rate with number of MACs, a rough proxy for time or energy consumption [12,13]. There is a trade-off between size and accuracy, with larger networks requiring more operations but also giving better accuracy. In the SQL picture, each point becomes a curve because now we are free to vary the number of photons per MAC, and the energy bound is set by the total number of photons, not the number of MACs. Figure 4(b) plots the error rate as a function of photon number for the networks above. While the general trade-off

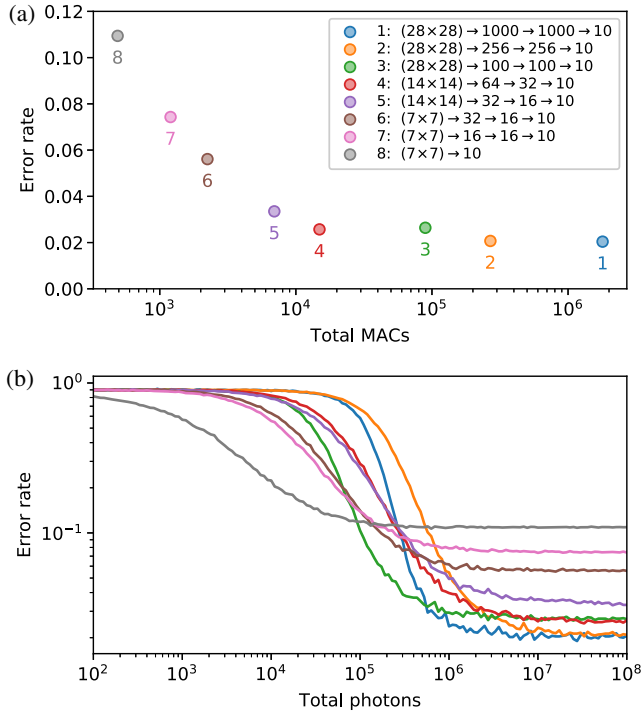


FIG. 4. (a) Conventional picture. Error rate as a function of number of MACs for different fully connected MNIST neural networks. (b) SQL picture. Error rate as a function of total number of photons, for the same networks.

between energy and accuracy is preserved, there are a number of counterintuitive results. For example, according to Fig. 4(a), networks 1 and 2 have similar performance but the first requires $8 \times$ more MACs, so under a conventional analysis, network 2 would always be preferred. However, Fig. 4(b) indicates that network 1 has better performance at all energy levels. This is because network 1 is less sensitive to shot noise due to averaging over many neurons, and therefore can be operated at lower energies, compensating for the increased neuron count. The same apparent paradox is seen with networks 3 and 4. This suggests that, in a quantum-limited scenario, reducing total energy may not be as simple as reducing the number of operations.

The total energy budget depends on many factors besides the SQL. Figure 5 plots energy per MAC as a function of the average number of input neurons per layer N , a rough “size” of the neural network. The SQL data are plotted for the eight networks in Fig. 4, and the corresponding dashed line is an empirical fit. Note that the SQL is an absolute lower bound, assumes perfect detectors, and counts only input optical energy. In a realistic device, this curve is shifted up by a factor $(\eta_d \eta_c \eta_s \beta_{\text{mod}})^{-1}$, where η_d , η_c , and η_s are the detector, coupling, and source (laser) efficiencies and β_{mod} is the modulator launch efficiency [56]; these are all close enough to unity in integrated systems [26,57–59] that the factor is $\lesssim 10$.

Another key factor is the detector electronics. The homodyne signal from each neuron needs to be sent through a nonlinear function $y_i \rightarrow f(y_i)$ and converted to the optical domain using a modulator [Fig. 1(b)]. The most obvious

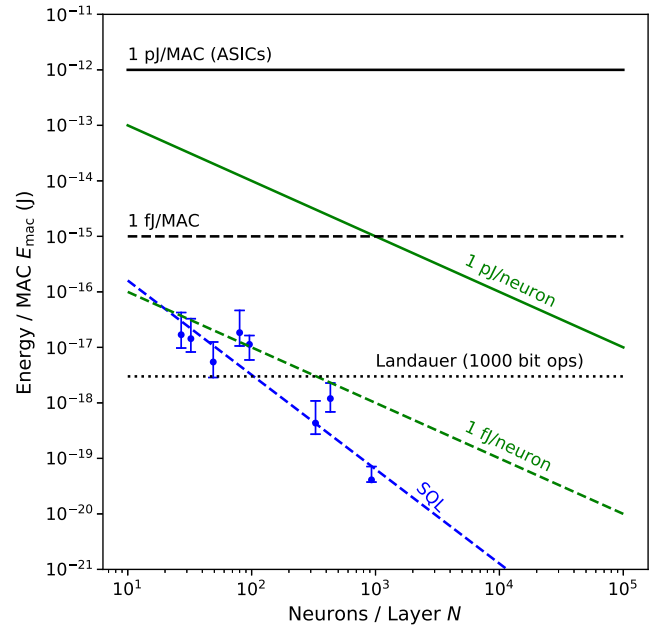


FIG. 5. Contributions to energy budget. SQL dots correspond to minimum E_{MAC} required to make the error rate $p_{\text{err}}(E_{\text{MAC}}) < 1.5 p_{\text{err}}(\infty)$ [error bars correspond to $p_{\text{err}}(E_{\text{MAC}}) = [1.2, 2.0] p_{\text{err}}(\infty)$]. $E_{\text{MAC}} = n_{\text{MAC}}(hc/\lambda)$, $\lambda = 1.55 \mu\text{m}$.

way to do this is to amplify and digitize the signal, perform the function $f(\cdot)$ in digital logic, serialize the outputs, convert back to analog, and send the analog signal into the modulator. Transimpedance amplifiers designed for optical interconnects operate at the ~ 100 fJ range [26,60], while analog to digital converters (ADCs) in the few-pJ/sample regime are available [61] and simple arithmetic (for the activation function) can be performed at the picojoule scale [15–17]. Modulators in this energy range are standard [57,58,60]. Thus a reasonable near-term estimate would be few-pJ/neuron; this figure is divided by the number of inputs per neuron to give the energy per MAC (solid green curve in Fig. 5). This few-pJ/neuron figure includes both optical and electrical energy: even though only a fraction of the energy is optical, the optical signal will be large compared to both shot noise [Eq. (3)] and amplifier Johnson noise $\langle \Delta n_e \rangle_{\text{rms}} \sim 10^3$ [62], so noise will not significantly degrade the network’s performance.

A much more aggressive goal is 1 fJ/neuron (dashed green curve). This figure is out of reach with current technology, but research into fJ/bit on-chip interconnects may enable it in the future [26,62]. A range of modulator designs supports few-fJ/bit operation [63–66]. On-chip interconnects also require photodetectors with ultralow (femtofarad) capacitance, so that a femtojoule of light produces a detectable signal without amplification [26,62]; such detectors have been realized with photonic crystals [67], plasmon antennas [68,69], and nanowires [70]. By eliminating the amplifier, ultrasmall “receiverless” detectors avoid its ~ 100 fJ energy cost as well as the Johnson noise associated with the amplifier. [Johnson noise still leads to fluctuations in the capacitor charge (kTC noise) that go as $\langle \Delta n_e \rangle_{\text{rms}} = \sqrt{kTC}/e \approx 12\sqrt{C/fF}$ [71], but for small detectors shot noise will dominate; see Sec. S4 of Supplemental Material [31].] Since 1 fJ/neuron is below the energy figures for ADCs, it would require well-designed analog electronics (for the nonlinear activation function) and very tight integration between detector, logic, and modulator [26]. At these energies, shot noise is also non-negligible and the SQL becomes relevant, but as mentioned above, due to optical inefficiencies the SQL will likely be relevant at higher energies as well.

For context, the ~ 1 pJ/MAC figure [15–17] for state-of-the-art ASICs is shown in Fig. 5. Energy consumption in nonreversible logic gates is bounded by the Landauer (thermodynamic) limit $E_{\text{op}} = kT \log(2) \approx 3$ zJ [72]. While multiply and accumulate is technically a reversible operation, all realistic computers implement it using nonreversible binary gates, so Landauer’s principle applies. A 32-bit multiplication [73,74] requires approximately 10^3 binary gates (see Sec. S4 of Supplemental Material [31]) and each bit operation consumes at least $kT \log(2)$, giving a limit $E_{\text{MAC}} \geq 3$ aJ (dotted line in Fig. 5). This is already higher than the SQL for the larger networks with $N \geq 100$. The optical neural network can achieve sub-Landauer

performance because (1) it operates in analog, avoiding the overhead of many bit operations per multiplication, and (2) the matrix product is performed through optical interference, which is reversible and not subject to the bound. To understand the second point, recall that homodyne detection computes the dot product via the polarization identity: $\vec{u} \cdot \vec{v} = \frac{1}{4}(\|\vec{u} + \vec{v}\|^2 - \|\vec{u} - \vec{v}\|^2)$. Optical interference, the reversible element that breaks Landauer’s assumption, is needed to convert the signals representing \vec{u} and \vec{v} to $\vec{u} \pm \vec{v}$ before squaring on the detectors and subtracting.

A final consideration is the electrical energy required to generate the weights. There is one weight pulse per MAC, so at the minimum this will be 1 fJ/MAC for the modulator, and may rise above 1 pJ/MAC once the driver electronics and memory access are included. However, once the optical signal is generated, it can be fanned out to many neural networks in parallel, reducing this cost by a factor of B , the batch size. Large batch sizes should enable this contribution to E_{MAC} to reach the few-femtojoule regime, and potentially much lower.

V. TRAINING AND CONVOLUTIONS WITH OPTICAL MATRIX-MATRIX MULTIPLIER

As discussed previously, the optical unit in Fig. 1(b) performs a matrix-vector product, and running multiple units in parallel with the same set of weights performs a general matrix-matrix product (GEMM), a key function in the basic linear algebra subprograms (BLAS) [75]. Figure 6 shows a schematic for an optical GEMM unit based on homodyne detection inspired by the neural-network concept. 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) integrated photonic transmitter arrays. Cylindrical lenses map these inputs 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 operation requires $m \cdot n \cdot k$ MACs, and the total energy consumption (and energy per MAC) are

$$E_{\text{tot}} = (mk + nk)E_{\text{in}} + (mn)E_{\text{out}},$$

$$E_{\text{MAC}} = \left(\frac{1}{n} + \frac{1}{m}\right)E_{\text{in}} + \frac{1}{k}E_{\text{out}}, \quad (4)$$

where E_{in} , E_{out} are the transmitter and receiver energy requirements, per symbol, which include all optical energy plus electronic driving, serialization, DAC or ADC, etc. If all matrix dimensions (m , n , k) are large, significant energy savings per MAC are possible if E_{in} , E_{out} can be kept reasonably small.

We saw above that the optical system could be used for neural-network inference. When running a batch of B instances $X = [x_1 \dots x_B]$, the output $Y = [y_1 \dots y_B]$ can be computed through the matrix-matrix product $Y = AX$.

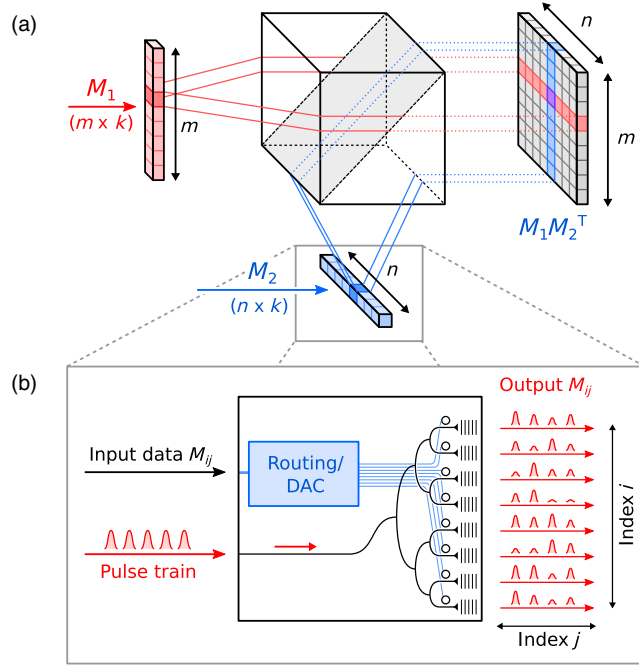


FIG. 6. (a) Matrix multiplication with a 2D detector array, two 1D transmitter arrays, and optical fan-out. Imaging lenses (including cylindrical lenses for row and column fan-out) not shown. (b) Schematic diagram of transmitter array.

In fully connected layers, training and backpropagation also rely heavily on GEMM. The goal of training is to find the set of weights $A^{(k)}$ that minimize the loss function L , which characterizes the inaccuracy of the model. Training typically proceeds by gradient-based methods. Since the loss depends on the network output, we start at the final layer and work backward, a process called backpropagation [7,8]. At each layer, we compute the gradient $(\nabla_A L)_{ij} = \partial L / \partial A_{ij}$ from the quantity $(\nabla_Y L)_{ij} = \partial L / \partial Y_{ij}$, and propagate the derivative back to the input $(\nabla_X L)_{ij} = \partial L / \partial X_{ij}$ [Fig. 7(a)]. These derivatives are computed from the chain rule and can be written as matrix-matrix multiplications:

$$\nabla_A L = (\nabla_Y L) X^T, \quad \nabla_X L = A^T (\nabla_Y L). \quad (5)$$

Once the derivative has been propagated to $\nabla_{X^{(k)}} L$ (for layer k), we use the chain rule to compute $\nabla_{Y^{(k-1)}} L = f'(\nabla_{X^{(k)}} L)$ and proceed to the previous layer. In this way, we sequentially compute the derivatives $\nabla_{A^{(k)}} L$ at each layer in the neural network.

In addition to fully connected layers, it is also possible to run convolutional layers on the optical GEMM unit by employing a ‘‘patching’’ technique [76]. In a convolutional layer, the input $x_{ij;k}$ is a $W \times H$ image with C channels. This is convolved to produce an output $y_{ij;k}$ of dimension $W' \times H'$ with C' channels [13]:

$$y_{ij;k} = \sum_{i',j',l} K_{i'j',kl} X_{(s_x i+i')(s_y j+j')l}. \quad (6)$$

Here $K_{i'j',kl}$ is the convolution kernel, a four-dimensional tensor of size $K_x \times K_y \times C' \times C$, and (s_x, s_y) are the strides of the convolution. Naively vectorizing Eq. (6) and running it as a fully connected matrix-vector multiply is very inefficient because the resulting matrix is sparse and contains many redundant entries. Patching expresses the image as a matrix X of size $K_x K_y C \times W' H'$, where each column corresponds to a vectorized $K_x \times K_y$ patch of the image [Fig. 7(b)]. The elements of the kernel are rearranged to form a (dense) matrix K of size $C' \times K_x K_y C$. Equation (6) can then be computed by taking the matrix-matrix product $Y = KX$, which has size $C' \times W' H'$. On virtually any microprocessor, GEMM is a highly optimized function with very regular patterns of memory access; the benefits of rewriting the convolution as a GEMM greatly outweigh the redundancy of data storage arising from overlapping patches [76]. The time required to rearrange the image as a patch matrix is typically very small compared to the time to compute the GEMM [77] (and can be further reduced if necessary with network-on-chip architectures [78] or optical buffering [79]); therefore, by accelerating the GEMM, the optical matrix multiplier will significantly improve the speed and energy efficiency of convolutional layers. Note also that, since we are performing the convolution as a matrix-matrix (rather than

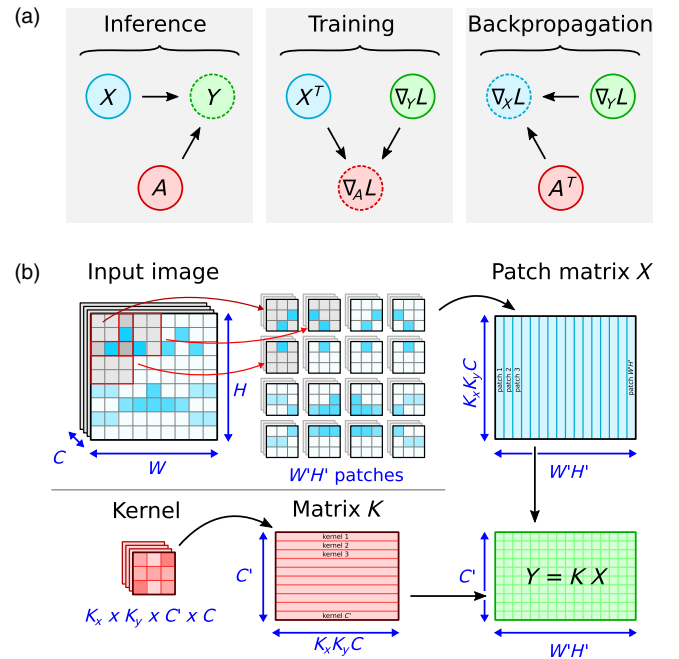


FIG. 7. Applications of optical GEMM. (a) Required matrix operations for inference, training, and backpropagation in a deep neural network. (b) Patching technique to recast a convolution ($K_x = K_y = 3$, $s_x = s_y = 2$ shown) as a matrix-matrix multiplication.

matrix-vector) operation, it is possible to obtain energy savings even without running the neural network on large batches of data. Computing the convolution requires $W'H'K_xK_yC'C$ MACs. Following Eq. (4), the energy per MAC (not including memory rearrangement for patching) is

$$E_{\text{MAC}} = \underbrace{\left(\frac{1}{C'} + \frac{1}{W'H'}\right)}_{1/c_{\text{in}}} E_{\text{in}} + \underbrace{\frac{1}{K_xK_yC}}_{1/c_{\text{out}}} E_{\text{out}}. \quad (7)$$

The coefficients $c_{\text{in}} = (1/C + 1/W'H')^{-1}$ and $c_{\text{out}} = K_xK_yC$ govern the energy efficiency when we are limited by input or output energies (transmitter or receiver and associated electronics). Since reading a 32-bit register takes \sim pJ of energy [13], a reasonable lower bound for near-term systems is $E_{\text{in}}, E_{\text{out}} \gtrsim$ pJ. Thus it is essential that $c_{\text{in}}, c_{\text{out}} \gg 1$ for the energy performance of the optical system to beat an ASIC (\sim pJ/MAC).

As a benchmark problem, we consider ALEXNET [1], the first convolutional neural network to perform competitively at the ImageNet Large-Scale Visual Recognition Challenge [9]. ALEXNET consists of five convolutional (CONV) layers and three fully connected (FC) layers, and consistent with deep neural networks generally, the majority of the energy consumption comes from the CONV layers [13]. Table I gives the layer dimensions and the values of $c_{\text{in}}, c_{\text{out}}$ for the CONV layers in ALEXNET [1]. The MAC-weighted averages for all layers are $\langle c_{\text{in}} \rangle > 100$ and $\langle c_{\text{out}} \rangle > 1000$. Thus, even under extremely conservative assumptions of $E_{\text{in}}, E_{\text{out}} \gtrsim 100$ pJ (comparable to DRAM read energies [13,14]), it is still possible to achieve sub-pJ/MAC performance.

TABLE I. Layers in ALEXNET [1]. Input dimension is $227 \times 227 \times 3$. Values of $c_{\text{in}}, c_{\text{out}}$ are calculated from Eq. (7). Max pooling layers after CONV1, CONV2, and CONV5 are used to reduce the image size, but the relative computational cost for these layers is negligible.

Layer	Output	Kernel	Stride	MACs	c_{in}	c_{out}
CONV1	$55 \times 55 \times 96$	11×11	4	105M	93	363
(Pool)	$27 \times 27 \times 96$...	2
CONV2	$27 \times 27 \times 256$	5×5	1	448M	189	2400
(Pool)	$13 \times 13 \times 256$...	2
CONV3	$13 \times 13 \times 384$	3×3	1	150M	117	2304
CONV4	$13 \times 13 \times 384$	3×3	1	224M	117	3456
CONV5	$13 \times 13 \times 256$	3×3	1	150M	102	3456
(Pool)	$6 \times 6 \times 256$...	2
FC1	4096	38M
FC2	4096	17M
FC3	1000	4M
Total CONV layers				1.08G	132	1656
Total FC layers				59M

$M = 10^6, G = 10^9$.

More advanced technology, such as few-femtojoule optical interconnects [26], may significantly reduce E_{in} and E_{out} , and therefore the energy per MAC. However, the performance is still fundamentally limited by detector shot noise [see, e.g., Eq. (3) for FC layers]. Section S3 of Supplemental Material [31] extends the shot-noise analysis to the case of matrix-matrix products needed for the convolutional case. Using a pretrained ALEXNET model (see Sec. VII for details), Figure 8(b) shows the top-ten accuracy on the ImageNet validation set as a function of the number of photons per MAC n_{MAC} . Consistent with Fig. 3, there are two limits: $n_{\text{MAC}} \ll 1$ corresponds to the random guess regime with 99% error rate (for top-ten accuracy with 1000 classes), while $n_{\text{MAC}} \gg 1$ recovers the accuracy of the noiseless model.

The dashed lines in Fig. 8(b) show the fictitious case where noise is present in only a single layer, while the solid green line corresponds to the case where all layers have noise and n_{MAC} is the same for each layer. Not all layers contribute equally to the noise: CONV1 is the most sensitive, requiring $n_{\text{MAC}} \gtrsim 20$, while the deeper layers (particularly the fully connected layers) can tolerate much

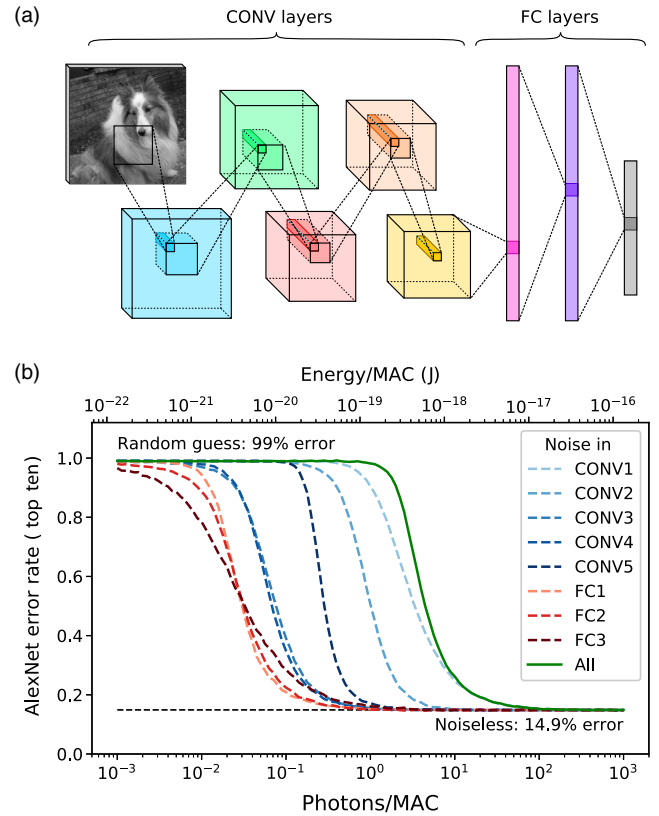


FIG. 8. (a) Schematic drawing of ALEXNET, which consists of five convolutional layers and 3 fully connected layers. Pooling and normalization steps not shown. (b) Error rate for pretrained ALEXNET as a function of n_{MAC} . Dashed lines show the effect of noise in a single layer, while solid green line shows the performance of the actual machine where all layers have noise.

lower energies $n_{\text{MAC}} \gtrsim 1$. Since the SNR is related to the total power received, which scales as $c_{\text{out}} n_{\text{MAC}}$ for the convolutional layers (c_{out} pulses per detector), it is not surprising that the deeper layers, which have a larger c_{out} , are less sensitive to quantum noise. The SQL obtained for ALEXNET ($n_{\text{MAC}} \gtrsim 20$ or $E_{\text{MAC}} \gtrsim 3$ aJ) is slightly larger than that from the MNIST networks in Fig. 3, but of the same order of magnitude, suggesting that the SQL is somewhat problem dependent.

It is worth contrasting the optical GEMM to more familiar optical convolvers. It has long been known that 2D convolutions can be performed with optical Fourier transforms [80–82]. However, this technique suffers from two significant drawbacks. First, it employs spatial light modulators, which limits the speed at which the kernel can be reprogrammed. In addition, optics performs a single-channel ($C = C' = 1$) convolution, and while extending to multiple output channels is possible by tiling kernels [83], multiple input *and* output channels may be difficult.

In contrast to free-space and fully integrated approaches, the optical GEMM leverages the complementary strengths of both free-space and integrated photonics. Integrated photonics is an ideal platform for realizing the transmitters, as these employ a large number of fast (gigahertz) modulators on chip. On-chip integration allows scaling to large arrays with control over the relative phase of each output beam (a capability exploited in recent chip-based phased arrays for beam steering [84–86]). Free-space propagation provides an essential third dimension, which enables high bandwidths at moderate clock frequencies [26] and data fan-out patterns that are difficult to implement on a 2D photonic chip. However, having a free-space element leads to issues with phase stability and aberrations. Since the transmitters are integrated, it is the relative phase between the beam paths that drifts (on timescales long compared to a computation), and this can be stabilized with a single feedback loop to the overall transmitter phase, a small constant overhead that does not scale with matrix size. To correct for geometric aberrations and minimize cross talk between detectors, multilens arrangements can be used, a standard practice in high-resolution imaging systems [87]. Section S2 of Supplemental Material [31] presents an optical design and analysis using Zemax® simulation software supporting the hypothesis that a $10^3 \times 10^3$ optical GEMM is achievable.

VI. DISCUSSION

This paper has 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 photoelectric effect, via the relation $I \propto |E|^2$, to compute the required matrix products optoelectronically without need for an all-optical nonlinearity, a key difficulty that has hobbled conventional approaches to optical computing [32]. Since the

device can be constructed with free-space optical components, it can scale to much larger sizes than purely nanophotonic implementations [23], being ultimately limited by the size of the detector array ($N \gtrsim 10^6$).

A key advantage to this scheme is that the multiplication itself is performed passively by optical interference, so the main speed and energy costs are associated with routing data into and out of the device. For a matrix multiplication $C_{m \times n} = A_{m \times k} B_{k \times n}$, the input-output (IO) energy scales as $O(mk) + O(nk) + O(mn)$, while the number of MACs scales as $O(mnk)$. For moderately large problems found in convolutional neural-network layers ($m, n, k \geq 100$) with moderate IO energies (\sim pJ), performance in the ~ 10 fJ/MAC range should be feasible, which is 2–3 orders of magnitude smaller than for state-of-the-art CMOS circuits [15–17]. Advances in optical interconnects [57,58,63] may reduce the IO energies by large factors [26], translating to further improvements in energy per MAC.

The fundamental limits to a technology are important to its long-term scaling. For the optical neural network presented here, detector shot noise presents a standard quantum limit to neural-network energy efficiency [44]. Because this limit is physics based, it cannot be engineered away unless nonclassical states of light are employed [45,46]. To study the SQL in neural networks, we performed Monte Carlo simulations on pretrained models for MNIST digit recognition (fully connected) and ImageNet image classification (convolutional). In both cases, network performance is a function of the number of photons used, which sets a lower bound on the energy per MAC. This bound is problem and network dependent, and for the problems tested in this paper, lies in the range 50 zJ–5 aJ/MAC. By contrast, the Landauer (thermodynamic) limit for a digital processor is 3 aJ/MAC (assuming 1000 bit operations per MAC [73,74]); sub-Landauer performance is possible because the multiplication is performed through optical interference, which is reversible and not bounded by Landauer’s principle.

Historically, the exponential growth in computing power has driven advances in machine learning by enabling the development of larger, deeper, and more complex models [11,13,16,17]. As Moore’s law runs out of steam, photonics may become necessary for continued growth in processing power—not only for interconnects [26], but also for logic. The architecture sketched in this paper promises significant short-term performance gains over state-of-the-art electronics, with a long-term potential, bounded by the standard quantum limit, of many orders of magnitude of improvement.

VII. METHODS

Neural-network performance was computed using Monte Carlo simulations. For fully connected layers, Eq. (3) was used, while for convolutional layers, the

convolution was performed by first forming the patch matrix [Fig. 7(b)] and performing the matrix-matrix multiplication (noise model discussed in Sec. S3 of Supplemental Material [31]). The weights for the fully connected MNIST neural networks were trained on a GPU using TENSORFLOW. A pretrained TENSORFLOW version of ALEXNET (available online at Ref. [88]) was modified to implement the quantum noise model and used for ImageNet classification. Simulations were performed on an NVIDIA Tesla K40 GPU.

ACKNOWLEDGMENTS

R.H. is supported by an IC Postdoctoral Research Fellowship at MIT, administered by ORISE through U.S. DOE and Office of the Director of National Intelligence (ODNI). L.B. is supported by a Doctoral Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.E. and M.S. acknowledge support from the U.S. ARO through the ISN at MIT (No. W911NF-18-2-0048). The authors acknowledge John Peurifoy (MIT) for training a number of the MNIST neural-network models, and NVIDIA Corporation for the donation of the Tesla K40 GPU used in this research. We are grateful to Joel Emer (NVIDIA and MIT) and Vivienne Sze (MIT) for helpful discussions.

-
- [1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, in *Advances in Neural Information Processing Systems* (NeurIPS, Tahoe, Nevada, 2012), pp. 1097–1105.
- [2] T. Young, D. Hazarika, S. Poria, and E. Cambria, *Recent Trends in Deep Learning Based Natural Language Processing*, *IEEE Comput. Intell. Mag.* **13**, 55 (2018).
- [3] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, *A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play*, *Science* **362**, 1140 (2018).
- [4] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, and G.E. Dahl, *Neural Message Passing for Quantum Chemistry*, in *Proceedings of the 34th International Conference on Machine Learning (ICML)* (ACM, Sydney, 2017), pp. 1263–1272.
- [5] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, *Deep Learning for Identifying Metastatic Breast Cancer*, [arXiv:1606.05718](https://arxiv.org/abs/1606.05718).
- [6] F. Rosenblatt, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*, *Psychol. Rev.* **65**, 386 (1958).
- [7] P. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, Ph.D. thesis, Harvard University, 1974.
- [8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-Based Learning Applied to Document Recognition*, *Proc. IEEE* **86**, 2278 (1998).
- [9] 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*, *Int. J. Comput. Vis.* **115**, 211 (2015).
- [10] G. E. Moore, *Cramming More Components onto Integrated Circuits*, *Electronics* **38**, 114 (1965).
- [11] D. Steinkraus, I. Buck, and P. Y. Simard, *Using GPUs for Machine Learning Algorithms*, in *Proceedings of the Eighth International Conference on Document Analysis and Recognition, 2005* (IEEE, New York, 2005), pp. 1115–1120, <https://doi.org/10.1109/ICDAR.2005.251>.
- [12] A. Canziani, E. Culurciello, and A. Paszke, *Evaluation of Neural Network Architectures for Embedded Systems*, in *IEEE International Symposium on Circuits and Systems (ISCAS), 2017* (IEEE, New York, 2017), pp. 1–4, <https://doi.org/10.1109/ISCAS.2017.8050276>.
- [13] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, *Efficient Processing of Deep Neural Networks: A Tutorial and Survey*, *Proc. IEEE* **105**, 2295 (2017).
- [14] M. Horowitz, *Computing's Energy Problem (and What We Can Do About It)*, in *IEEE International Solid-State Circuits Conference (ISSCC), Digest of Technical Papers* (IEEE, New York, 2014), pp. 10–14, <https://doi.org/10.1109/ISSCC.2014.6757323>.
- [15] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, *GPUs and the Future of Parallel Computing*, *IEEE Micro* **31**, 7 (2011).
- [16] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, *DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning*, *ACM Sigplan Not.* **49**, 269 (2014).
- [17] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, *In-Datcenter Performance Analysis of a Tensor Processing Unit*, in *Proceedings of the ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), 2017* (IEEE, New York, 2017), pp. 1–12, <https://doi.org/10.1145/3079856.3080246>.
- [18] S. George, S. Kim, S. Shah, J. Hasler, M. Collins, F. Adil, R. Wunderlich, S. Nease, and S. Ramakrishnan, *A Programmable and Configurable Mixed-Mode FPAA SoC*, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **24**, 2253 (2016).
- [19] K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, *A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications*, *Nano Lett.* **12**, 389 (2012).
- [20] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang *et al.*, *Efficient and Self-Adaptive In-Situ Learning in Multilayer Memristor Neural Networks*, *Nat. Commun.* **9**, 2385 (2018).
- [21] B. Feinberg, S. Wang, and E. Ipek, *Making Memristive Neural Network Accelerators Reliable*, in *Proceedings of the 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (IEEE, New York, 2018), pp. 52–65, <https://doi.org/10.1109/HPCA.2018.00015>.

- [22] 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**, 1004 (2018).
- [23] 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*, *Nat. Photonics* **11**, 441 (2017).
- [24] 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*, *Sci. Rep.* **7**, 7430 (2017).
- [25] J. M. Kahn and D. A. B. Miller, *Communications Expands Its Space*, *Nat. Photonics* **11**, 5 (2017).
- [26] D. A. B. Miller, *Attojoule Optoelectronics for Low-Energy Information Processing and Communications*, *J. Lightwave Technol.* **35**, 346 (2017).
- [27] A. Rogalski, *Progress in Focal Plane Array Technologies*, *Prog. Quantum Electron.* **36**, 342 (2012).
- [28] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, *Network of Time-Multiplexed Optical Parametric Oscillators as a Coherent Ising Machine*, *Nat. Photonics* **8**, 937 (2014).
- [29] T. Inagaki, K. Inaba, R. Hamerly, K. Inoue, Y. Yamamoto, and H. Takesue, *Large-Scale Ising Spin Network Based on Degenerate Optical Parametric Oscillators*, *Nat. Photonics* **10**, 415 (2016).
- [30] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara *et al.*, *A Fully Programmable 100-Spin Coherent Ising Machine with All-to-All Connections*, *Science* **354**, 614 (2016).
- [31] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevX.9.021032> for homodyne product implementation details, aberration analysis, and detailed derivations of shot noise, Johnson noise, and Landauer limit formulas.
- [32] D. A. B. Miller, *Are Optical Transistors the Logical Next Step?*, *Nat. Photonics* **4**, 3 (2010).
- [33] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, *Broadcast and Weight: An Integrated Network for Scalable Photonic Spike Processing*, *J. Lightwave Technol.* **32**, 4029 (2014).
- [34] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. Van Campenhout, *Toward Optical Signal Processing Using Photonic Reservoir Computing*, *Opt. Express* **16**, 11182 (2008).
- [35] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, *Optoelectronic Reservoir Computing*, *Sci. Rep.* **2**, 287 (2012).
- [36] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutiérrez, L. Pesquera, C. R. Mirasso, and I. Fischer, *Photonic Information Processing Beyond Turing: An Optoelectronic Implementation of Reservoir Computing*, *Opt. Express* **20**, 3241 (2012).
- [37] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, *A Leaky Integrate-and-Fire Laser Neuron for Ultrafast Cognitive Computing*, *IEEE J. Sel. Topics Quantum Electron.* **19**, 1 (2013).
- [38] D. Brunner, S. Reitzenstein, and I. Fischer, *All-Optical Neuromorphic Computing in Optical Networks of Semiconductor Lasers*, in *Proceedings of the 2016 IEEE International Conference on Rebooting Computing (ICRC)* (IEEE, New York, 2016), pp. 1–2, <https://doi.org/10.1109/ICRC.2016.7738705>.
- [39] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, *Experimental Realization of Any Discrete Unitary Operator*, *Phys. Rev. Lett.* **73**, 58 (1994).
- [40] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, *Optimal Design for Universal Multiport Interferometers*, *Optica* **3**, 1460 (2016).
- [41] A. Einstein, *Über Einen die Erzeugung und Verwandlung des Lichtes Betreffenden Heuristischen Gesichtspunkt*, *Ann. Phys. (Berlin)* **322**, 132 (1905).
- [42] D. F. Walls and G. J. Milburn, *Quantum Optics* (Springer Science & Business Media, Berlin, 2007).
- [43] M. M. Hayat, B. E. A. Saleh, and J. A. Gubner, *Shot-Noise-Limited Performance of Optical Neural Networks*, *IEEE Trans. Neural Networks* **7**, 700 (1996).
- [44] C. M. Caves, *Quantum-Mechanical Noise in an Interferometer*, *Phys. Rev. D* **23**, 1693 (1981).
- [45] M. T. Jaekel and S. Reynaud, *Quantum Limits in Interferometric Measurements*, *Europhys. Lett.* **13**, 301 (1990).
- [46] H. Grote, K. Danzmann, K. L. Dooley, R. Schnabel, J. Slutsky, and H. Vahlbruch, *First Long-Term Application of Squeezed States of Light in a Gravitational-Wave Observatory*, *Phys. Rev. Lett.* **110**, 181101 (2013).
- [47] L. Holmstrom and P. Koistinen, *Using Additive Noise in Back-Propagation Training*, *IEEE Trans. Neural Networks* **3**, 24 (1992).
- [48] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors*, arXiv:1207.0580.
- [49] H. Gibbs, *Optical Bistability: Controlling Light with Light* (Elsevier, New York, 2012).
- [50] C. M. Savage and H. J. Carmichael, *Single Atom Optical Bistability*, *IEEE J. Quantum Electron.* **24**, 1495 (1988).
- [51] C. Santori, J. S. Pelc, R. G. Beausoleil, N. Tezak, R. Hamerly, and H. Mabuchi, *Quantum Noise in Large-Scale Coherent Nonlinear Photonic Circuits*, *Phys. Rev. Applied* **1**, 054005 (2014).
- [52] J. Kerckhoff, M. A. Armen, and H. Mabuchi, *Remnants of Semiclassical Bistability in the Few-Photon Regime of Cavity QED*, *Opt. Express* **19**, 24468 (2011).
- [53] X. Ji, F. A. S. Barbosa, S. P. Roberts, A. Dutt, J. Cardenas, Y. Okawachi, A. Bryant, A. L. Gaeta, and M. Lipson, *Ultra-Low-Loss On-Chip Resonators with Sub-Milliwatt Parametric Oscillation Threshold*, *Optica* **4**, 619 (2017).
- [54] S. Hu, M. Khater, R. Salas-Montiel, E. Kratschmer, S. Engelmann, W. M. J. Green, and S. M. Weiss, *Experimental Realization of Deep-Subwavelength Confinement in Dielectric Optical Resonators*, *Sci. Adv.* **4**, eaat2355 (2018).
- [55] C. Wang, C. Langrock, A. Marandi, M. Jankowski, M. Zhang, B. Desiatov, M. M. Fejer, and M. Loncar, *Ultrahigh-Efficiency Second-Harmonic Generation in Nanophotonic PPLN Waveguides*, *Optica* **5**, 1438 (2018).
- [56] D. A. B. Miller, *Energy Consumption in Optical Modulators for Interconnects*, *Opt. Express* **20**, A293 (2012).

- [57] C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin *et al.*, *Single-Chip Microprocessor that Communicates Directly Using Light*, *Nature (London)* **528**, 534 (2015).
- [58] 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 (London)* **556**, 349 (2018).
- [59] A. Michaels and E. Yablonovitch, *Inverse Design of Near Unity Efficiency Perfectly Vertical Grating Couplers*, *Opt. Express* **26**, 4766 (2018).
- [60] 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*, *J. Lightwave Technol.* **34**, 2924 (2016).
- [61] B. E. Jonsson, *An Empirical Approach to Finding Energy Efficient ADC Architectures*, in *Proceedings of 2011 International Workshop on ADC Modeling, Testing and Data Converter Analysis and Design (IMEKO, Orvieto, Italy, 2011)*, pp. 1–6, <https://www.imeko.org/publications/iwadc-2011/IMEKO-IWADC-2011-28.pdf>.
- [62] M. Notomi, K. Nozaki, A. Shinya, S. Matsuo, and E. Kuramochi, *Toward f/Bit Optical Communication in a Chip*, *Opt. Commun.* **314**, 3 (2014).
- [63] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. S. Hosseini, A. Biberman, and M. R. Watts, *An Ultralow Power Athermal Silicon Modulator*, *Nat. Commun.* **5**, 4008 (2014).
- [64] C. Koos, J. Leuthold, W. Freude, M. Kohl, L. Dalton, W. Bogaerts, A. L. Giesecke, M. Lauermann, A. Melikyan, S. Koeber *et al.*, *Silicon-Organic Hybrid (SOH) and Plasmonic-Organic Hybrid (POH) Integration*, *J. Lightwave Technol.* **34**, 256 (2016).
- [65] C. Haffner, D. Chelladurai, Y. Fedoryshyn, A. Josten, B. Baeuerle, W. Heni, T. Watanabe, T. Cui, B. Cheng, S. Saha *et al.*, *Low-Loss Plasmon-Assisted Electro-optic Modulator*, *Nature (London)* **556**, 483 (2018).
- [66] S. A. Srinivasan, M. Pantouvaki, S. Gupta, H. T. Chen, P. Verheyen, G. Lepage, G. Roelkens, K. Saraswat, D. Van Thourhout, P. Absil *et al.*, *56 Gb/s Germanium Waveguide Electro-Absorption Modulator*, *J. Lightwave Technol.* **34**, 419 (2016).
- [67] K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, M. Ono, A. Shakoor, E. Kuramochi, and M. Notomi, *Photonic-Crystal Nano-Photodetector with Ultrasmall Capacitance for On-Chip Light-to-Voltage Conversion without an Amplifier*, *Optica* **3**, 483 (2016).
- [68] T. Ishi, J. Fujikata, K. Makita, T. Baba, and K. Ohashi, *Si Nano-Photodiode with a Surface Plasmon Antenna*, *Jpn. J. Appl. Phys.* **44**, L364 (2005).
- [69] L. Tang, S. E. Kocabas, S. Latif, A. K. Okyay, D.-S. Ly-Gagnon, K. C. Saraswat, and D. A. B. Miller, *Nanometre-Scale Germanium Photodetector Enhanced by a Near-Infrared Dipole Antenna*, *Nat. Photonics* **2**, 226 (2008).
- [70] L. Cao, J.-S. Park, P. Fan, B. Clemens, and M. L. Brongersma, *Resonant Germanium Nanoantenna Photodetectors*, *Nano Lett.* **10**, 1229 (2010).
- [71] J. R. Pierce, *Physical Sources of Noise*, *Proc. IRE* **44**, 601 (1956).
- [72] R. Landauer, *Irreversibility and Heat Generation in the Computing Process*, *IBM J. Res. Dev.* **5**, 183 (1961).
- [73] M. Nagamatsu, S. Tanaka, J. Mori, T. Noguchi, and K. Hatanaka, *A 15-ns 32 × 32-Bit CMOS Multiplier with an Improved Parallel Structure*, *IEEE J. Solid State Circuits* **25**, 494 (1990).
- [74] H. H. Yao and E. E. Swartzlander, *Serial-Parallel Multipliers*, in *Proceedings of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, 1993* (IEEE, New York, 1993), pp. 359–363, <https://doi.org/10.1109/ACSSC.1993.342534>.
- [75] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh, *Basic Linear Algebra Subprograms for FORTRAN Usage*, *ACM Trans. Math. Softw.* **5**, 308 (1979).
- [76] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, *cuDNN: Efficient Primitives for Deep Learning*, [arXiv:1410.0759](https://arxiv.org/abs/1410.0759).
- [77] X. Li, G. Zhang, H. H. Huang, Z. Wang, and W. Zheng, *Performance Analysis of GPU-Based Convolutional Neural Networks*, in *Proceedings of the 45th International Conference on Parallel Processing (ICPP), 2016* (IEEE, New York, 2016), pp. 67–76, <https://doi.org/10.1109/ICPP.2016.15>.
- [78] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, *Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks*, *IEEE J. Solid-State Circuits* **52**, 127 (2017).
- [79] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, *On-Chip Optical Convolutional Neural Networks*, [arXiv:1808.03303](https://arxiv.org/abs/1808.03303).
- [80] A. V. Lugt, *Signal Detection by Complex Spatial Filtering*, *IEEE Trans. Inf. Theory* **10**, 139 (1964).
- [81] E. G. Paek and D. Psaltis, *Optical Associative Memory Using Fourier Transform Holograms*, *Opt. Eng.* **26**, 265428 (1987).
- [82] N. J. New, *Reconfigurable Optical Processing System*, U.S. Patent No. 9,594,394 (2017).
- [83] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, *Hybrid Optical-Electronic Convolutional Neural Networks with Optimized Diffractive Optics for Image Classification*, *Sci. Rep.* **8**, 12324 (2018).
- [84] J. Sun, E. Timurdogan, A. Yaacobi, E. S. Hosseini, and M. R. Watts, *Large-Scale Nanophotonic Phased Array*, *Nature (London)* **493**, 195 (2013).
- [85] S. W. Chung, H. Abediasl, and H. Hashemi, *A Monolithically Integrated Large-Scale Optical Phased Array in Silicon-on-Insulator CMOS*, *IEEE J. Solid-State Circuits* **53**, 275 (2018).
- [86] C. T. Phare, M. C. Shin, S. A. Miller, B. Stern, and M. Lipson, *Silicon Optical Phased Array with High-Efficiency Beam Formation over 180 Degree Field of View*, [arXiv:1802.04624](https://arxiv.org/abs/1802.04624).
- [87] W. J. Smith, *Modern Optical Engineering* (Tata McGraw-Hill Education, New York, 1966).
- [88] Y. Peng, *Implementation of ALEXNET with Tensorflow*, <https://github.com/ykpengba/AlexNet-A-Practical-Implementation>.