Supplementary Materials for

Single-shot optical neural network

Liane Bernstein et al.

Corresponding author: Liane Bernstein, lbern@mit.edu

Sci. Adv. 9, eadg7904 (2023)
DOI: 10.1126/sciadv.adg7904

This PDF file includes:

Experiment: Detailed description
Maximum DNN layer size
Latency, throughput, energy, and chip area of a near-term system
Negative weights
Calibration and image processing
Additional confusion matrices
Measurement of filtered supercontinuum laser spectra
Scaling up number of neurons
Figs. S1 to S3
Tables S1 and S2
References
Experiment: detailed description

Figure S1: Complete system elaborating on Fig. 3a from main text, including isolator (Isol.), single-mode polarization-maintaining fiber (PMF), iris and polarizers (Pol.) to reduce stray light. Focusing lens $f_2$ is in fact two achromatic lenses, with individual focal lengths of 750 mm and 180 mm.

This section provides more details on our experimental implementation, including added components to reduce stray light (iris, polarizers) — see Fig. S1. The first spatial light modulator (SLM, Meadowlark, AVR Optics P1920-400-800-HDMI-T, pixel width 9.2 µm) displays an input image $x$ by modulating the phase of the polarization component of the incident collimated beam that is parallel to the extraordinary axis. (The polarization component along the ordinary axis is unchanged.) Because the light incident on this SLM is polarized at 45° after rotation by the half-wave plates, the SLM effectively performs pixel-wise polarization rotation of the beam. A polarizing beamsplitter (PBS) then rejects the unrotated polarization from the SLM. The input activations are thus encoded into optical intensities through this ‘amplitude-mode’ operation of a ‘phase-only’ SLM. To display the input activations in the first layer of the DNN, we use every second pixel of the first SLM and turn every other
pixel ‘off’ such that the input pattern is effectively multiplied by a grating. We then block the zero order in the Fourier plane to reduce background, such as reflection from the backplane. In subsequent layers with fewer input activations (shorter $x$), we use every eighth pixel of the SLM to further reduce crosstalk. The incident light on the second SLM (Hamamatsu X10468-04, pixel width 20 $\mu$m) is horizontally polarized (along the extraordinary axis); this SLM is in the Fourier plane and adds a variable phase delay to each pixel to impart a fan-out phase pattern. We used the weighted Gerchberg-Saxton algorithm (without camera feedback) to determine the phase patterns ($39, 52$), which only needed to be calculated once per network size, as they are independent of the weight and input activation values. The third SLM (Meadowlark, AVR Optics P1920-400-800-HDMI-T, pixel width 9.2 $\mu$m), in the image plane, is used in ‘amplitude mode’, similarly to the first SLM. Telescopes of achromatic lenses (Thorlabs ACT508-250-A, AC508-180-A, and ACT508-750-A) transmit the replicated input activations to the image planes for 1:1 mapping from SLM #1 to SLM #3 to the camera (Thorlabs DCC3240M, pixel width 5.3 $\mu$m). The lens positions along the optical axis are fine-tuned with linear stages. A MATLAB program performs all the control and processing on a digital electronic computer. The supercontinuum source in the throughput experiments is the SuperK EXW-12 from NKT with a VARIA tunable filter.

**Maximum DNN layer size**

The number of pixels in the weighting SLM and camera limit the maximum weight matrix size ($N \times K$) that can be applied in one pass through our system. The number of pixels in the fan-out SLM (DOE) in the Fourier plane dictates the number of spots that can be generated in the image plane, which in turn, sets the maximum output vector length $N$. In Ref. ($52$), with a 1.3-megapixel SLM, 1,500 uniform spots were experimentally generated (which translates to $N = 1,500$ in our scheme). Considering these constraints, with our megapixel spatial light modulators, our setup can theoretically perform million-element matrix-vector multiplication, with $N = K = 1,000$, which we could implement with reduced stray light on the detector for higher data transmission accuracy,
as mentioned in the Discussion. We can achieve additional scaling with higher-resolution SLMs or tiling (see Scaling section below).

**Latency, throughput, energy and chip area of a near-term system**

![Diagram](image)

Figure S2: Path of one input activation through the single-shot ONN in an optimized setup. A VCSEL or µLED converts the signal from the electronic to optical domain (EO conversion) and the optical copy is a reconfigurable diffractive optical element. Each photodetector (PD) includes a weighting element and is electrically connected to $K$ other PDs for analog electronic summation. A transimpedance amplifier (TIA) reads out the analog output signal from each block, which the DAC then converts to the digital domain. The nonlinearity is a simple comparator. Each of the $K$ input activations goes through these processing steps simultaneously such that after one pass, the matrix-vector product is complete.

In this section, we describe the latency, throughput, energy and chip area of a near-term, CMOS-compatible implementation of our single-shot ONN. This proposed system, as shown in Fig. S2 and described in the main text, includes digital-to-analog converters (DAC) and a high-speed source array with $K$ elements, LCoS SLMs for fan-out and weighting, photodetectors (PDs) for optical-to-electrical (OE) conversion, passive, analog electronic summation, and for each block, amplification by a transimpedance amplifier (TIA), analog-to-digital conversion (ADC) and a nonlinearity (NL). We assume a matrix size of 1 million elements ($N = K = 1,000$), as described in the Maximum DNN layer size section above, since megapixel cameras and SLMs are readily available. The latency of the system is defined by the sum of the latencies of each of the components encountered by one input activation. The DAC, light source, TIA and ADC take $\sim 1$ ns each (a standard computer clock
operates at $\sim$GHz). In terms of the NL, Ref. (54) demonstrates an optical ReLU function that operates in sub-ps time scales and also surveys digital electronic, analog electronic and optoelectronic implementations of ReLU with latencies $\leq$ ns. With a photon time of flight of $<10$ ns, the system thus operates with a latency on the order of $\sim10$ ns for a full matrix-vector computation.

As mentioned in the main text, computation is pipelined such that the throughput is defined by the slowest component, yielding $10^6$ MACs per $\sim1$ ns, i.e., petaMAC/s for $N = K = 1,000$.

The energy consumption for a complete matrix-vector multiplication that comprises $N \times K$ MACs is the sum of the DAC, SLM, TIA, ADC and nonlinearity energies, plus the photon energy required to discriminate 256 levels on the TIA:

$$E_{total} = N \cdot \frac{1}{\eta_s \eta_d} \cdot \eta_{PD} \cdot 2^{n_b} \cdot \xi \cdot t + K \cdot E_{DAC} + 2 \cdot E_{SLM} + N \cdot (E_{TIA} + E_{ADC} + E_{NL})$$

where $\eta_s \approx 10\%$ is the source wall-plug efficiency (37, 59), $\eta_d \approx 80\%$ is the optical efficiency of the DOE, $\eta_{PD} \approx 0.2$ A/W is the PD responsivity (40), $n_b = 8$ bits, $\xi \approx 1$ $\mu$A is the TIA sensitivity at 1 GHz (62), $t = 1$ ns is the integration time, i.e., computer clock cycle time, $E_{SLM}$ is the energy consumed by the SLM in one clock cycle ($<10$ nJ) and each of the remaining component energies (TIA (62), ADC (63, 64), DAC (64, 65), NL (7, 11, 54)) are 1 pJ per operation. The total optical efficiency of the DOE, $\eta_d$, depends on the light utilization and diffraction efficiency of the SLM (both $\sim95\%$ in commercial LCoS SLMs such as our Hamamatsu X10468 (61)), as well as the fraction of optical energy in the desired spots generated by the fan-out pattern (determined by Fourier transform, can be $>90\%$ for $\sim100$-$1,000 \times$ fan-out (52, 60)). The product of these factors yields $>80\%$. For $N = K = 1,000$, the energy per MAC ($E_{total}/(N \cdot K)$) is therefore on the order of $\sim10$ fJ/MAC.

These parameters and calculations are summarized in Table S1.

The total chip area of the system is calculated in Table S2. The component areas have been demonstrated experimentally in the literature, in CMOS technology nodes larger than the state of the art – therefore, the TIA, ADC, NL, DAC and VCSEL areas could likely be further miniaturized. The weighting elements, on the other hand, will also be limited by optical spot size. The diffraction limit is sub-$\mu$m for green light, but a pixel size of a few $\mu$m is more reasonable to maintain low crosstalk.
Table S1: Parameters in energy calculation

| Symbol | Parameter | Value \(^{c}\) | Fan-out | Energy/MAC \(^{d}\) |
|--------|-----------|----------------|---------|------------------|
| \(\eta_s\) | Laser wall-plug efficiency | 10\% \((37, 59)\) | | |
| \(\eta_d\) | Optical efficiency of DOE | \(>80\% \((52, 60, 61)\)\) | | |
| \(\eta_{PD}\) | PD responsivity | 0.2 A/W \((40)\) | | |
| \(n_b\) | Effective number of bits | 8 | | |
| \(\xi\) | TIA sensitivity | 1 \(\mu\)A \((62)\) | | |
| \(t\) | Clock cycle time | 1 ns | | |
| \(E_{DAC}\) | Energy per DAC conversion | \(1 \text{ pJ} \((65), (64)\) \) \(N\) | 1 \(\text{pJ}/N \rightarrow 1 \text{fJ/MAC}\) |
| \(E_{SLM}\) | Energy of LCoS SLM | \((<10 \text{ W})t\) \(N \times K\) | <10 nJ/(\(N \times K\)) \(\rightarrow<10 \text{ fJ/MAC}\) |
| \(E_{optical}\) | Optical energy per block \(^{1}\) | \(\frac{1}{\eta_s \cdot \eta_d \cdot \eta_{PD}} \cdot 2^{n_b} \cdot \xi \cdot t \approx 10 \text{ pJ}\) \(K\) | 10 \(\text{pJ}/K \rightarrow 10 \text{ fJ/MAC}\) |
| \(E_{TIA}\) | Energy of TIA | 1 \(\text{pJ} \(62)\) \(K\) | 1 \(\text{pJ}/K \rightarrow 1 \text{ fJ/MAC}\) |
| \(E_{ADC}\) | Energy per ADC conversion | 2 \(\text{pJ} \((63, 64)\)\) \(K\) | 2 \(\text{pJ}/K \rightarrow 2 \text{ fJ/MAC}\) |
| \(E_{NL}\) | Energy of nonlinearity | <1 \(\text{pJ} \((7, 11, 54)\)\) \(K\) | <1 \(\text{pJ}/K \rightarrow <1 \text{ fJ/MAC}\) |
| \(E_{total}\) | Energy of full system | | \(\sim 10 \text{ fJ/MAC}\) |

\(^{1}\) Optical energy for \(2^{n_b}\) distinguishable levels by the TIA.

\(^{2}\) Demonstrated in the literature.

\(^{3}\) DAC within ADC.

\(^{4}\) Assuming \(N = K = 1,000\).

Table S2: Projected chip area

| Element | Area (mm\(^2\)) | Number of elements | Total area (mm\(^2\)) |
|---------|-----------------|--------------------|----------------------|
| Weighting element \(^{1}\) | \(1.4 \cdot 10^{-2} \ (69)\) | \(10^4\) | 14 |
| TIA | 0.0022 \((62)\) | \(10^4\) | 2.2 |
| ADC | 0.0016 \((70)\) | \(10^3\) | 1.6 |
| NL | 0.001 \((71)\) | \(10^3\) | 1 |
| DAC | <0.0016 \((70)\) \(^3\) | \(10^3\) | <1.6 |
| VCSEL | .01 \((72)\) \(^3\) | \(10^3\) | 10 |

| Area of full system | | | 30 |

\(^{1}\) Includes PD (e.g., PDs in Refs. \((18, 43)\)), since weighting element is placed on top of each PD.

\(^{2}\) DAC within ADC.

\(^{3}\) Aperture of 6 \(\mu\)m, pitch of 100 \(\mu\)m in demonstrated array.

in a real system that includes imperfections (i.e., aberrations and misalignment).

**Negative weights**

A number of solutions can implement negative weighting. For example, a second wire can connect all the PDs in a block; charge from the negatively weighted pixels would be directed into this second wire with an analog switch. The output from the ‘negative’ wire can then be subtracted from the output of the ‘positive’ wire. Because this subtraction only occurs once per block, its cost is amortized by a
factor of $K$ and is therefore small with respect to the other costs of the system. Another possibility for negative weighting is to use two PDs per receiver pixel, where one PD pushes charge into the block’s wire in the case of a positive weight, and another pulls charge in the case of a negative weight. The weight value of the unused photodetector is set to maximum extinction. Lastly, the weights can be shifted to all positive values, as described by Wang et al. (26).

**Calibration and image processing**

The LCoS SLM calibrations from the manufacturers are set to map the input values of 0 to 255 linearly to a 0 to $2\pi$ phase shift for normally incident 532 nm light. Since we use the SLMs from Meadowlark/AVR Optics (i.e., SLMs #1 and #3) to modulate amplitude and not phase, the output light intensity then varies sinusoidally with input value and requires recalibration. This SLM model has 2048 voltage settings in hardware with $>2\pi$ phase modulation. In the calibration step, we display a uniform array at one voltage value, then average the outputs to obtain one output value per input voltage. We then step through the input voltage values linearly through time, and fit the inputs versus averaged outputs with a 9th-order polynomial. We use the resulting lookup table as a global calibration to map desired output intensities to SLM input values. The latter are then restricted to 7-8 bits of precision depending on the intensity region since there are multiple maximum to minimum output intensity cycles within the 2048 available input voltage values. Furthermore, because a wide region of SLM #3 is illuminated, local nonuniformities will cause different blocks (subimages) to require slightly different calibrations to achieve a linear weight display. Therefore, we calculate a refined fit per subimage, where we once again display a uniform array at each frame, but here, step through the display values from the global lookup table. We then fit 8th-order polynomials to the inputs versus averaged outputs per subimage and adjust the displayed weight values accordingly.

We also perform simple processing of the output images from the camera. To reduce the impact of stray light in the system, we acquire a background with SLMs #1 and #3 set to all zeros, and subtract this background from every output frame. We also perform $2 \times 2$ pixel binning. Furthermore,
the fan-out pattern displayed on SLM #2 does not yield subimages of equal average intensities on the camera. Therefore, we also acquire a smoothed and background-subtracted calibration map by setting all activations and weights to a constant value, used for normalization in our experiments. In layer 1, the outputs are also divided by the mean intensity per subimage of 100 randomly selected images from the MNIST validation set. The normalization process reduces the effective number of bits for some subimages since we do not make use of the full dynamic range of the camera. We also fan out an extra row to replace the subimage that overlaps with the zero order (as stated in Fig. 4b-d) and the dimmest subimages. The spot uniformity could be improved by adjusting the pattern on SLM #2 with a feedback algorithm \((39, 52)\), eliminating the need for normalization and replacement. In an optimized implementation, the background subtraction can be implemented with a bias voltage.

**Additional confusion matrices**

Figure S3 shows the confusion matrices that correspond to the classification accuracies reported in Table 1.

**Measurement of filtered supercontinuum laser spectra**

We corrected the measured supercontinuum spectra for the wavelength response of our experimental setup, which has low transmission for wavelengths \(<460\) nm or \(>560\) nm, primarily due to the dielectric mirror of SLM #2. To determine our system’s response function, we first set the VARIA filter attachment on the SuperK to the narrowest bandwidth. We then swept its center wavelength (every 10 nm from 450 nm to 630 nm) and measured the summed intensity on the camera with SLMs #1 and #3 set to a uniform display. We performed a similar measurement when we connected the supercontinuum laser to the spectrometer, where we measured and integrated the filtered spectrum at the narrowest bandwidth setting for different center wavelengths. We then estimated our system’s wavelength response as the summed camera intensity per wavelength divided by the spectrometer’s summed intensity per wavelength. We could then calculate the corrected spectra by multiplying the
Figure S3: Confusion matrices for the classification of datasets reported in Table 1.
broad-bandwidth spectra (measured on the spectrometer) by this system response (interpolated).

**Scaling up number of neurons**

With $N = K = 1,000$, our system will be able to accommodate some of the largest DNN layer sizes currently in use, such as those found in Transformers (2). However, as networks are becoming steadily larger (e.g., GPT-3, with fully connected layers with input vectors up to a length of 10,000 elements (73)), the system may need to be further scaled up. One way to increase the number of matrix elements and output neurons is to use SLMs with more pixels, such as the commercially available 10-megapixel SLM in Ref. (69). Another option could be to use multiple SLMs and combine their outputs with a series of beamsplitters. Lastly, at the cost of decreasing energy efficiency and increasing latency, the inputs can be tiled and the weights updated over different time steps with a small local electronic memory at each weight pixel. In state-of-the-art nodes, an SRAM cell’s area is 0.021 $\mu$m$^2$ (74). Therefore, $\sim$80 digital values of 8 bits each can fit into a CMOS chip area of (3.74 $\mu$m)$^2$, which is the area of a pixel in high-resolution SLMs (69). This trade-off space can be evaluated and optimized in future work, e.g., using the DNN mapping tool Timeloop (75).
REFERENCES AND NOTES

1. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in Advances in Neural Information Processing Systems (2012), vol. 25, pp. 1097–1105.

2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in Advances in Neural Information Processing Systems (2017), vol. 30, pp. 5998–6008.

3. A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare. Nat. Med. 25, 24–29 (2019).

4. X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, Y. Shi, Scaling for edge inference of deep neural networks. Nat. Electron. 1, 216–222 (2018).

5. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models. arXiv:2001.08361 [cs.LG] (23 January 2020).

6. E. Nurvitadhi, G. Venkatesh, J. Sim, D. Marr, R. Huang, J. Ong Gee Hock, Y. T. Liew, K. Srivatsan, D. Moss, S. Subhaschandra, G. Boudoukh, Can FPGAs beat GPUs in accelerating next-generation deep neural networks?, in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (Association for Computing Machinery, 2017), pp. 5–14.

7. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-I. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, D. H. Yoon, In-datacenter performance analysis of a tensor processing unit, in Proceedings of the 44th Annual International Symposium on Computer Architecture (2017), pp. 1–12.

8. D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Mungua, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training. arXiv:2104.10350 [cs.LG] (23 April 2021).
9. E. A. Huerta, G. Allen, I. Andreoni, J. M. Antelis, E. Bachelet, G. B. Berriman, F. B. Bianco, R. Biswas, M. Carrasco Kind, K. Chard, M. Cho, P. S. Cowperthwaite, Z. B. Etienne, M. Fishbach, F. Forster, D. George, T. Gibbs, M. Graham, W. Gropp, R. Gruendl, A. Gupta, R. Haas, S. Habib, E. Jennings, M. W. G. Johnson, E. Katsavounidis, D. S. Katz, A. Khan, V. Kindratenko, W. T. C. Kramer, X. Liu, A. Mahabal, Z. Marka, K. McHenry, J. M. Miller, C. Moreno, M. S. Neubauer, S. Oberlin, A. R. Olivas, D. Petrivick, A. Rebei, S. Rosofsky, M. Ruiz, A. Saxton, B. F. Schutz, A. Schwing, E. Seidel, S. L. Shapiro, H. Shen, Y. Shen, L. P. Singer, B. M. Sipocz, L. Sun, J. Towns, A. Tsokaros, W. Wei, J. Wells, T. J. Williams, J. Xiong, Z. Zhao, Enabling real-time multi-messenger astrophysics discoveries with deep learning. *Nat. Rev. Phys.* **1**, 600–608 (2019).

10. J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. D. de las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, M. Riedmiller, Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602**, 414–419 (2022).

11. V. Sze, Y. Chen, T. Yang, J. S. Emer, Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **105**, 2295–2329 (2017).

12. A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).

13. P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).

14. T. P. Xiao, C. H. Bennett, B. Feinberg, S. Agarwal, M. J. Marinella, Analog architectures for neural network acceleration based on non-volatile memory. *Appl. Phys. Rev.* **7**, 031301 (2020).

15. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).

16. A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).

17. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, H. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, H.
Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).

18. D. A. B. Miller, Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).

19. S. Maktoobi, L. Froehly, L. Andreoli, X. Porte, M. Jacquot, L. Larger, D. Brunner, Diffractive coupling for photonic networks: How big can we go? *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–8 (2019).

20. J. W. Goodman, A. R. Dias, L. M. Woody, Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Lett.* **2**, 1–3 (1978).

21. R. A. Athale, W. C. Collins, Optical matrix–matrix multiplier based on outer product decomposition. *Appl. Optics* **21**, 2089–2090 (1982).

22. D. Psaltis, N. Farhat, Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. *Opt. Lett.* **10**, 98–100 (1985).

23. K. Wagner, D. Psaltis, Multilayer optical learning networks. *Appl. Optics* **26**, 5061–5076 (1987).

24. J. W. Goodman, 4 Decades of optical information processing. *Opt. Photonics News* **2**, 11–15 (1991).

25. R. Hamerly, L. Bernstein, A. Sluidds, M. Soljačić, D. Englund, Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).

26. T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, P. L. McMahon, An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).

27. J. Spall, X. Guo, A. I. Lvovsky, Hybrid training of optical neural networks. *Optica* **9**, 803–811 (2022).

28. H. Zheng, Q. Liu, Y. Zhou, I. I. Kravchenko, Y. Huo, J. Valentine, Meta-optic accelerators for object classifiers. *Sci. Adv.* **8**, eabo6410 (2022).

29. L. Bottou, C. Cortes, J. Denker, I. G. H. Drucker, L. Jackel, Y. le Cun, U. Muller, E. Sackinger, P. Simard, V. Vapnik, Comparison of classifier methods: A case study in handwritten digit recognition, in *Proceedings of the International Conference on Pattern Recognition* (1994), pp. 77–82.

30. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, Q. Dai, Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photonics* **15**, 367–373 (2021).

31. M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, V. J. Sorger, Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812–1819 (2020).

32. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
33. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, G. Wetzstein, Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 1–10 (2018).

34. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

35. H. Xiao, K. Rasul, R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747 [cs.LG] (15 September 2017).

36. J. Jongejan, H. Rowley, T. Kawashima, J. Kim, N. Fox-Gieg, The Quick, Draw! AI experiment (2016); https://quickdraw.withgoogle.com/.

37. A. Skalli, J. Robertson, D. Owen-Newns, M. Hejda, X. Porte, S. Reitzenstein, A. Hurtado, D. Brunner, Photonic neuromorphic computing using vertical cavity semiconductor lasers. *Opt. Mater. Express* **12**, 2395–2414 (2022).

38. M. S. Wong, S. Nakamura, S. P. DenBaars, Progress in high performance III-nitride micro-light-emitting diodes. *ECS J. Solid State Sci. Technol.* **9**, 015012 (2019).

39. C. L. Panuski, I. Christen, M. Minkov, C. J. Brabec, S. Trajtenberg-Mills, A. D. Griffi ths, J. J. D. McKendry, G. L. Leake, D. J. Coleman, C. Tran, J. St Louis, J. Mucci, C. Horvath, J. N. Westwood-Bachman, S. F. Preble, M. D. Dawson, M. J. Strain, M. L. Fanto, D. R. Englund, A full degree-of-freedom spatiotemporal light modulator. *Nat. Photonics* **16**, 834–842 (2022).

40. B. F ah s, A. J. Chowdhury, Y. Zhang, J. Ghasemi, C. Hitchcock, P. Zarkesh-Ha, M. M. Hella, Design and modeling of blue-enhanced and bandwidth-extended PN photodiode in standard CMOS technology. *IEEE Trans. Electron Devices* **64**, 2859–2866 (2017).

41. H. Zimmermann, A. Marchlewski, W. Gaberl, I. Jonak-Auer, G. Meinhardt, E. Wachmann, Blue-enhanced PIN finger photodiodes in a 0.35-μm SiGe BiCMOS technology. *IEEE Photonics Technol. Lett.* **21**, 1656–1658 (2009).

42. W. T. Chen, A. Y. Zhu, F. Capasso, Flat optics with dispersion-engineered metasurfaces. *Nat. Rev. Mater.* **5**, 604–620 (2020).

43. S. Latif, S. Kocabas, L. Tang, C. Debaes, D. Miller, Low capacitance CMOS silicon photodetectors for optical clock injection. *Appl. Phys. A* **95**, 1129–1135 (2009).

44. Y. Gao, H. Cansizoglu, K. G. Polat, S. Ghandiparsi, A. Kaya, H. H. Mamtaz, A. S. Mayet, Y. Wang, X. Zhang, T. Yamada, E. P. Devine, A. F. Elrefaie, S.-Y. Wang, M. S. Islam, Photon-trapping microstructures enable high-speed high-efficiency silicon photodiodes. *Nat. Photonics* **11**, 301–308 (2017).
45. M. Burla, C. Hoessbacher, W. Heni, C. Haffner, Y. Fedoryshyn, D. Werner, T. Watanabe, H. Massler, D. L. Elder, L. R. Dalton, J. Leuthold, 500 GHz plasmonic Mach-Zehnder modulator enabling sub-THz microwave photonics. *APL Photonics* **4**, 056106 (2019).

46. C. Han, Z. Zheng, H. Shu, M. Jin, J. Qin, R. Chen, Y. Tao, B. Shen, B. Bai, F. Yang, Y. Wang, H. Wang, F. Wang, Z. Zhang, S. Yu, C. Peng, X. Wang, Slow light silicon modulator beyond 110 GHz bandwidth. arXiv:2302.03652 [physics.optics] (7 February 2023).

47. P. Kharel, C. Reimer, K. Luke, L. He, M. Zhang, Breaking voltage-bandwidth limits in integrated lithium niobate modulators using micro-structured electrodes. *Optica* **8**, 357–363 (2021).

48. W. Dong, H. Liu, J. K. Behera, L. Lu, R. J. H. Ng, K. V. Sreekanth, X. Zhou, J. K. W. Yang, R. E. Simpson, Wide bandgap phase change material tuned visible photonics. *Adv. Funct. Mater.* **29**, 1806181 (2019).

49. Y. Jung, H. Han, A. Sharma, J. Jeong, S. S. Parkin, J. K. Poon, Integrated hybrid VO$_2$–silicon optical memory. *ACS Photonics* **9**, 217–223 (2022).

50. Y. Zhang, C. Fowler, J. Liang, B. Azhar, M. Y. Shalaginov, S. Deckoff-Jones, S. An, J. B. Chou, C. M. Roberts, V. Liberman, M. Kang, C. Rios, K. A. Richardson, C. Rivero-Baleine, T. Gu, H. Zhang, J. Hu, Electrically reconfigurable non-volatile metasurface using low-loss optical phase-change material. *Nat. Nanotechnol.* **16**, 661–666 (2021).

51. N. Quack, H. Sattari, A. Y. Takabayashi, Y. Zhang, P. Verheyen, W. Bogaerts, P. Edinger, C. Errando-Herranz, K. B. Gyffason, MEMS-enabled silicon photonic integrated devices and circuits. *IEEE J. Quantum Electron.* **56**, 1–10 (2019).

52. D. Kim, A. Keesling, A. Omran, H. Levine, H. Bernien, M. Greiner, M. D. Lukin, D. R. Englund, Large-scale uniform optical focus array generation with a phase spatial light modulator. *Opt. Lett.* **44**, 3178–3181 (2019).

53. C. Rosales-Guzmán, A. Forbes, *How to Shape Light with Spatial Light Modulators* (SPIE Press, 2017).

54. G. H. Y. Li, R. Sekine, R. Nehra, R. M. Gray, L. Ledezma, Q. Guo, A. Marandi, All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning. *Nanophotonics* **12**, 9 (2022).

55. L. Bernstein, A. Sluuds, R. Hamerly, V. Sze, J. Emer, D. Englund, Freely scalable and reconfigurable optical hardware for deep learning. *Sci. Rep.* **11**, 3144 (2021).

56. Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina, S. G. Tell, Y. Zhang, W. J. Dally, J. Emer, C. T. Gray, B. Khailany, S. W. Keckler, Simba: Scaling deep-learning inference with multi-chip-module-based architecture, in *Proceedings of*
the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO ’52 (Association for Computing Machinery, 2019), pp. 14–27.
57. M. Horowitz, Computing’s energy problem (and what we can do about it), in 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (IEEE, 2014), pp. 10–14.
58. A. Stillmaker, B. Baas, Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm. Dermatol. Int. 58, 74–81 (2017).
59. M. Kuramoto, S. Kobayashi, T. Akagi, K. Tazawa, K. Tanaka, T. Saito, T. Takeuchi, High-output-power and high-temperature operation of blue GaN-based vertical-cavity surface-emitting laser. App. Phys. Express 11, 112101 (2018).
60. R. Di Leonardo, F. Ianni, G. Ruocco, Computer generation of optimal holograms for optical trap arrays. Opt. Express 15, 1913–1922 (2007).
61. Hamamatsu, LCoS-SLM (Liquid Crystal on Silicon - Spatial Light Modulator) X10468/X13267/X13138 series, https://hamamatsu.su/media/index/?type=catalog&id=115 (2022) [accessed 21 December 2022].
62. N. Mehta, C. Sun, M. Wade, S. Lin, M. Popovic, V. Stojanović, A 12Gb/s, 8.6 μApp input sensitivity, monolithic-integrated fully differential optical receiver in CMOS 45nm SOI process, in ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference (IEEE, 2016), pp. 491–494.
63. B. E. Jonsson, An empirical approach to finding energy efficient ADC architectures, in Proc. of 2011 IMEKO IWADC & IEEE ADC Forum (2011), pp. 1–6.
64. V. Tripathi, B. Murmann, An 8-bit 450-MS/s single-bit/cycle SAR ADC in 65-nm CMOS, in 2013 Proceedings of the ESSCIRC (ESSCIRC) (IEEE, 2013), pp. 117–120.
65. O. Morales Chacón, J. J. Wikner, C. Svensson, L. Siek, A. Alvandpour, Analysis of energy consumption bounds in CMOS current-steering digital-to-analog converters. Analog Integr. Circuits Signal Process. 111, 339–351 (2022).
66. H. Saadat, H. Bokhari, S. Parameswaran, Minimally biased multipliers for approximate integer and floating-point multiplication. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 37, 2623–2635 (2018).
67. M. Shoba, R. Nakkeeran, Energy and area efficient hierarchy multiplier architecture based on Vedic mathematics and GDI logic. Eng. Sci. Technol. Int. J. 20, 321–331 (2017).
68. J. Johnson, Rethinking floating point for deep learning. arXiv:1811.01721 [cs.NA] (1 November 2018).
69. HOLOEYE Photonics AG, GAEA-2 10 megapixel phase only LCOS-SLM (reflective), holoeye.com/gaea-4k-phase-only-spatial-light-modulator/ (2022) [accessed 25 May 2022].

70. L. Kull, D. Luu, C. Menolfi, M. Braendli, P. A. Francese, T. Morf, M. Kossel, H. Yueksel, A. Cevrero, I. Ozkaya, T. Toifl, A 10b 1.5GS/s pipelined-SAR ADC with background second-stage common-mode regulation and offset calibration in 14nm CMOS FinFET, in 2017 IEEE International Solid-State Circuits Conference (ISSCC) (IEEE, 2017), pp. 474–475.

71. N. V. K. Boppana, S. Ren, A low-power and area-efficient 64-bit digital comparator. J. Circuits Syst. Comput. 25, 1650148 (2016).

72. M. Kuramoto, S. Kobayashi, T. Akagi, K. Tazawa, K. Tanaka, K. Nakata, T. Saito, Watt-class blue vertical-cavity surface-emitting laser arrays. Appl. Phys. Express 12, 091004 (2019).

73. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. arXiv:2005.14165 [cs.CL] (22 July 2020).

74. G. Yeap, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, Y. C. Peng, J. Y. Sheu, M. Wang, X. Chen, B. R. Yang, C. P. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. H. Chen, C. Y. Chang, H. K. Chen, P. Hung, C. S. Hou, Y. K. Cheng, J. Chang, L. Yuan, C. K. Lin, C. C. Chen, Y. C. Yeo, M. H. Tsai, H. T. Lin, C. O. Chui, K. B. Huang, W. Chang, H. J. Lin, K. W. Chen, R. Chen, S. H. Sun, Q. Fu, H. T. Yang, H. T. Chiang, C. C. Yeh, T. L. Lee, C. H. Wang, S. L. Shue, C. W. Wu, R. Lu, W. R. Lin, J. Wu, F. Lai, Y. H. Wu, B. Z. Tien, Y. C. Huang, L. C. Lu, J. He, Y. Ku, J. Lin, M. Cao, T. S. Chang, S. M. Jang, 5nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 $\mu$m$^2$ SRAM cells for mobile SoC and high performance computing applications, in 2019 IEEE International Electron Devices Meeting (IEDM) (IEEE, 2019), pp. 36.7.1–36.7.4.

75. A. Parashar, P. Raina, Y. S. Shao, Y.-H. Chen, V. A. Ying, A. Mukkara, R. Venkatesan, B. Khailany, S. W. Keckler, J. Emer, Timeloop: A systematic approach to DNN accelerator evaluation, in 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) (IEEE, 2019), pp. 304–315.