CHINESE OPTICS LETTERS

Vol. 19, No. 8 | August 2021

Optical tensor core architecture for neural network training based on dual-layer waveguide topology and homodyne detection

Shaofu Xu (徐绍夫) and Weiwen Zou (邹卫文)*

State Key Laboratory of Advanced Optical Communication Systems and Networks, Intelligent Microwave Lightwave Integration Innovation Center (iMLic), Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

*Corresponding author: wzou@sjtu.edu.cn

Received September 29, 2020 | Accepted December 21, 2020 | Posted Online April 12, 2021

We propose an optical tensor core (OTC) architecture for neural network training. The key computational components of the OTC are the arrayed optical dot-product units (DPUs). The homodyne-detection-based DPUs can conduct the essential computational work of neural network training, i.e., matrix-matrix multiplication. Dual-layer waveguide topology is adopted to feed data into these DPUs with ultra-low insertion loss and cross talk. Therefore, the OTC architecture allows a large-scale dot-product array and can be integrated into a photonic chip. The feasibility of the OTC and its effectiveness on neural network training are verified with numerical simulations.

Keywords: optical tensor core; neural network training; matrix multiplication; homodyne detection; dual-layer waveguides.
DOI: 10.3788/COL202119.082501

1. Introduction

Deep learning becomes a milestone strategy of modern machine learning[1], performing with superior ability in many areas and applications[2–6]. One of the major driving forces of deep learning is the surge of computational power. Among the procedures of deep learning, neural network training consumes the most time and energy. This is because an inference only takes one forward propagation to complete. However, a complete training takes thousands of rounds of forward and backward propagations. For now, the computational power demanded by neural network training doubles every 3.4 months[7] due to dramatic neural network complexity expansion. Traditional digital processors are thereby faced with bottlenecks caused by neural network development.

Recently, optical neural networks (ONNs) were proposed as an alternative to break through electronic problems, such as the clock rate limit and energy dissipation of data movement[8]. By mapping the mathematical model of the neural network onto analog optical devices, ONNs obtain results on the fly of light with potential ultra-low energy consumption[9]. Various ONN architectures are proposed and demonstrated based on unitary optics[10,11], wavelength division multiplexing[12], free-space modulators[13], diffractive optics[14], free-space homodyne detection[15], etc. Especially, the free-space homodyne ONN[15] carries out dot-products by homodyne detection and electron accumulation (HDEA), enabling the matrix-matrix multiplications. Note that the matrix-matrix multiplications are the essential computing process of neural network training. Therefore, the free-space homodyne architecture can conduct neural network inference and training on the same hardware. However, free-space implementation is bulky and instable.

Here, we propose an optical tensor core (OTC) architecture that can be integrated into photonic chips for neural network training. In this architecture, the matrix-matrix multiplication is conducted by the dot-product units (DPUs) meshed on a two-dimensional (2D) plane. The principle of the DPUs is based on the HDEA process, i.e., multiplications are fulfilled by homodyne detection, and the summation is completed by electron accumulation. Here, the components of DPUs are optical waveguide devices so that the DPU array can be integrated. Besides, the input data are fed into the DPU array through dual-layer waveguides. Provided that the waveguide crossings are inevitable if the date-feeding waveguides and DPU array are deployed on a single 2D plane, the dual-layer waveguide topology of the data-feeding waveguides can mitigate the insertion loss and crosstalk of such crossings. The sub-millidecibel (mdB) insertion loss per crossing[16] guarantees a large-scale OTC. The proposed OTC succeeds the strengths of free-space homodyne ONN (high speed, high reconfigurability, and large scale) and potentially features aberration immunity and compactness by removing the third space dimension and lens structure of free-space architecture.

2. Principle

No matter how the neural network structure varies (fully connected, convolutional, and recurrent), the basic mathematical model of neural network training comprises matrix-matrix multiplications and nonlinear activation functions[17]. The OTC focuses on conducting matrix-matrix multiplications, which consume the most computational power during training. Figure 1 illustrates the OTC architecture. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two input matrices: $\boldsymbol{A}$ has dimensions of $M \times S$, and $\boldsymbol{B}$ has dimensions of $S \times N$. The multiplication between the two matrices comprises $M \times N$ dot-products. As illustrated in Fig. 1(a), the OTC works with a pulsed laser. The repetition rate of the pulse train is the system clock rate f_m. The generated pulse train splits into two equal branches for the data modulation of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. In each branch, the pulse train evenly splits multiple times to meet the scale of the rows in $\boldsymbol{A}$ and the columns in $\boldsymbol{B}$. In the modulation array, the data of matrices are modulated on the amplitudes of optical pulses. The modulation rate equals f_m. Each row of $\boldsymbol{A}$ ($\boldsymbol{A}_i$, in the plot) or each column of $\boldsymbol{B}$ ($\boldsymbol{B}_j$ in the plot) is serially modulated on the amplitude of the pulse train. The modulated pulse trains enter the DPU array through dual-layer waveguides. Inter-layer couplers (ILCs)[16,18] are adopted for the transition between layers. In the schematic of Fig. 1(a), two different colors are used to show the overpass and underpass waveguides. During transmission, the waveguide crossings of overpass and underpass waveguides impose ultra-low loss (below 1 mdB/crossing) and crosstalk (below 40 dB) between the signals of the upper and lower layers[14]. At each crossing of the bus waveguides, a DPU is deployed for the dot-product calculation. The structure of a DPU is illustrated in Fig. 1(b). Two splitters (directional couplers) are applied to drop a portion of light from the bus waveguides. A phase shifter on one arm is used to adjust the phase for homodyne detection. An ILC on the other arm is adopted to transit lower-layer optical pulses to the upper layer. An HDEA is employed for dot-product calculation (the principle is described below). The HDEA is set up by a 3 dB directional coupler, a balanced photo-detector (BPD), and an accumulation capacitor with a triggered switch. The DPU array contains $M \times N$ DPUs in total, and the optical intensity is averagely distributed on these DPUs. Note that optical pulses should pass through the same optical lengths before encountering at each DPU. Isosceles-shaped waveguides are designed to guarantee that the optical lengths between DPUs and the modulator array are always the same.

The principle of HDEA is described here. Suppose a pair of incident optical pulses have amplitudes of $A_{i,k}$ and $B_{k,j}$ (kth element of input vectors), respectively, and arrive at the 3 dB directional coupler at the same time. Because of optical interference, the upper and lower detectors of the BPD generate current pulses. The subtracted current pulse is accumulated on the capacitor in the form of electrons or charges. When the initial phase difference of incident optical pulses is $\pi/2$, the number of accumulated electrons on the anode panel is proportional to the amplitude multiplication of the incident optical pulses, i.e., $A_{i,k} \times B_{k,j}$. When the initial phase difference is $-\pi/2$, electrons on the anode drift away. The phase inversion takes place at the push–pull modulation rather than the phase shifter. All phase shifters stay static once the calibrations are completed. To calibrate the DPU, one should set all modulators to their maximal transmission rate and adjust the phase shifters to reach the maximal output current of every BPD. After multiple (k is from one to S) optical pulse pairs are fed and electrical pulses are accumulated on the capacitor, the dot-product of vectors is completed. Results are acquired by sampling the voltage on the anode panel. Once the voltage is sampled, the trigger switches on to discharge the electrons for the preparation of

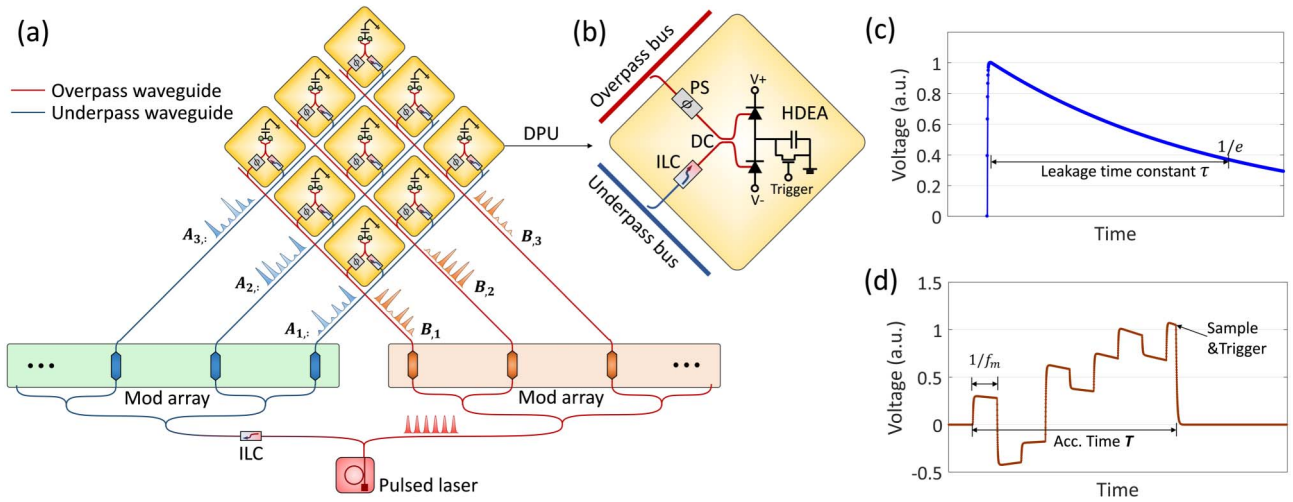

Fig. 1. (a) Schematic of the OTC. An example scale of 3 × 3 is depicted. ILC, inter-layer coupler; Mod array, modulator array. (b) Detailed schematic of a DPU. A portion of light is dropped from the bus waveguides. PS, phase shifter; DC, directional coupler. (c) Impulse response of the HDEA. Time constant τ of the circuit is defined as voltage decays to 1/e. (d) An example of electron accumulation. Optical pulses arrive at the HDEA with interval of $1/f_m$. The accumulation time is T.

the next dot-product. Note that the accumulated electrons are not permanent. If the input vectors have massive lengths, the initially accumulated electrons start to leak spontaneously. Figure 1(c) illustrates the impulse response of the HDEA. The electron leakage time constant τ is critical to the electron accumulation. As illustrated in Fig. 1(d), if the time constant of electron leakage is shorter than the accumulation time T, the accumulation result is distorted. Numerically, the final sampled voltage is described by

$$V_o \propto \sum_{k=1}^{S} e^{-\frac{S-k}{f_m \tau}} \cdot A_{i,k} \cdot B_{k,j}. \tag{1}$$

With fixed input vector length S and clock rate f_m, a larger time constant leads to better dot-product results. We implement a simulation program with integrated circuit emphasis (SPICE) to show the rise time and leakage time constants of HDEA. The adopted equivalent circuit model of the photodetector is referred to in Ref. [19]. With the accumulation capacitance of 10 pF, the rising time is 47.7 ps, and the leakage time constant is 109.1 ns. Results indicate that the sampling rate of electronic acquisition (the trigger) can be as slow as 10 MHz. Given that the clock rate of optical pulses is at the level of dozens of gigahertz (GHz), the HDEA process easily supports the dot-product calculations with vector length over 1000. Note that larger junction resistance ($> 100\,\mathrm{k\Omega}$) is often considered in photodetector models. Together with the progress on high-speed pulsed lasers and modulators, larger input lengths are expected. According to Ref. [15], large vector length indicates that the required optical power of each DPU is low. If the vector length surpasses 1000, a milliwatt-level pulsed laser has potential to support an OTC with 10^5 DPUs. Note that the laser efficiency, detector efficiency, modulator efficiency, coupling loss, and waveguide loss are highly relevant to the feasible DPU scale. These degrading factors should be carefully considered and designed in OTC fabrication.

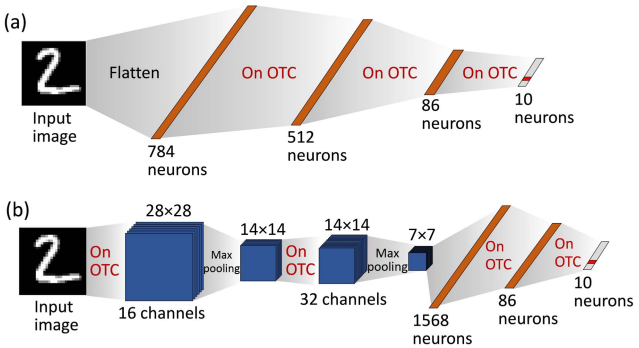

Fig. 2. (a) FC network. The matrix multiplications are implemented on OTC. ReLU after each layer is conducted in auxiliary electronics. The output is the one-hot classification vector given by the softmax function. (b) The convolutional network. Convolutions are conducted on OTC. Max pooling layers shrink the image size by half. All layers are ReLU-activated except for the pooling layers and the last layer.

3. Results

To validate the effectiveness of the OTC architecture, neural network training is simulated. In the simulation, we adopt two network models [fully-connected (FC) and convolutional] to conduct the image classification task of the modified National Institute of Standards and Technology (MNIST) handwritten digits. Figure 2 illustrates the network models in detail. The input images of the FC network and the convolutional network are from the MNIST dataset. In the four-layer FC network [Fig. 2(a)], the image is flattened to vectors in the first layer and propagates by matrix multiplications through the cascading layers. The numbers of neurons in the hidden layers are set to

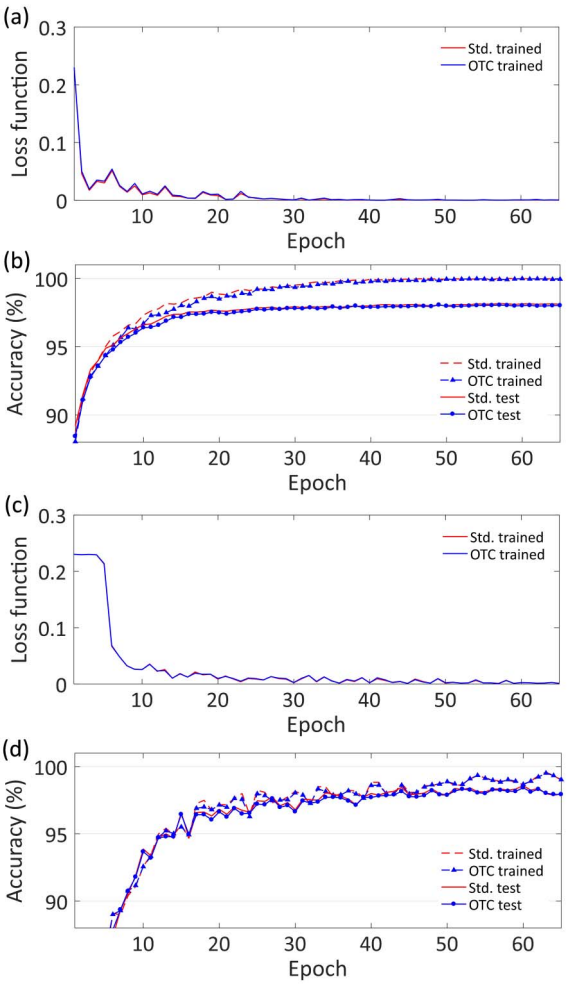

Fig. 3. (a) Loss functions of the FC network during training. Results of the standard MBGD algorithm (Std. train) and the on-OTC training are illustrated. (b) The prediction accuracy of the FC network during training. The training accuracy and the testing accuracy of the standard MBGD algorithm are depicted without marks. The on-OTC training is depicted with marks. (c) Loss functions of the convolutional network during training. (d) The prediction accuracy of the convolutional network during training.

784, 512, 86, and 10, respectively. The second and third layers adopt rectified linear units (ReLU) as the activation function, and the last layer uses softmax function to yield the one-hot classification vector. As illustrated in Fig. 2(b), the convolutional network comprises two convolutional layers, two max pooling layers, and two FC layers. The kernel size of the convolutional layers (first and third layers) is set to 3×3, and the output channel numbers are 16 and 32, respectively. The activation function used in the convolutional network is the ReLU except for the last layer. The last layer uses the softmax function.

The OTC is simulated to conduct all matrix multiplications of fully-connected layers and generalized matrix multiplication (GeMM) of convolutional layers. Auxiliary electronics, including analog-to-digital converters (ADCs) and digital processors, are utilized for the nonlinear operations. Specifically, max pooling, image flattening, nonlinear activation functions, and data rearrangement are executed by auxiliary electronics. Note that the temporal accumulation of the optical pulses significantly lowers the sampling speed by about 1000 times. Low-speed ADCs and digital processors can be utilized in the neural network training. Detailed discussions about the auxiliary electronics can be found in Ref. [15], where auxiliary electronics are similarly utilized. In the simulation, the optical pulses are assumed to be push–pull modulated with no phase shift. The clock rate (i.e., the repetition rate of optical pulses) is set at 50 GHz. The accumulation time T depends on the size of input vectors: for those larger than 100, T is set at 25 ns; otherwise, T is set at 2.5 ns to mitigate the impact from electron leakage. The leakage time constant ($\tau = 109.1$ ns) yielded in the SPICE simulation is used in the neural network training. We also consider the insertion loss of waveguide crossing as 1 mdB/crossing, while the crosstalk is neglected for its minor influence on the results.

We adopt mini-batch technology during training: the batch size of FC network training is 50, and the batch size of convolutional network training is 120. The mini-batch gradient descendent (MBGD) algorithm is applied to update the network parameters. Sixty-five epochs are executed in total. The learning rate is 0.02 during the initial 50 epochs and decreases to 0.004 from the 51st to 65th epochs.

Figure 3 shows the training procedure of the FC network and the convolutional network. As shown in Fig. 3(a), the loss function of the FC network drops with the growth of training epochs. For reference, we draw the loss function of the standard MBGD algorithm conducted by the 64 bit digital computer. The OTC-trained loss drops along with the standard MBGD algorithm, converging to a very small value. The corresponding prediction accuracy of the FC network is illustrated in Fig. 3(b). The training accuracy is calculated via 10,000 randomly picked inferences in the training set of MNIST, and the testing accuracy is calculated via 10,000 inferences in the test set. The initial parameters of the OTC training and standard training are the same. It can be found that the accuracies of the OTC training increase alongside with the standard MBGD algorithm. Finally, the training accuracy of the OTC reaches 100%, and the testing accuracy is around 98%, thus verifying the effectiveness of the OTC on the FC network training. Figure 3(c) shows the loss function of the convolutional network during training: the OTC-trained loss function almost overlaps with the standard-trained reference. From the prediction accuracy results in Fig. 3(d), we also observe that the training of the convolutional network on the OTC is effective. The training accuracy is around 99%, and the testing accuracy is around 98%. The results above validate the feasibility of OTC training on both the FC network and the convolutional network.

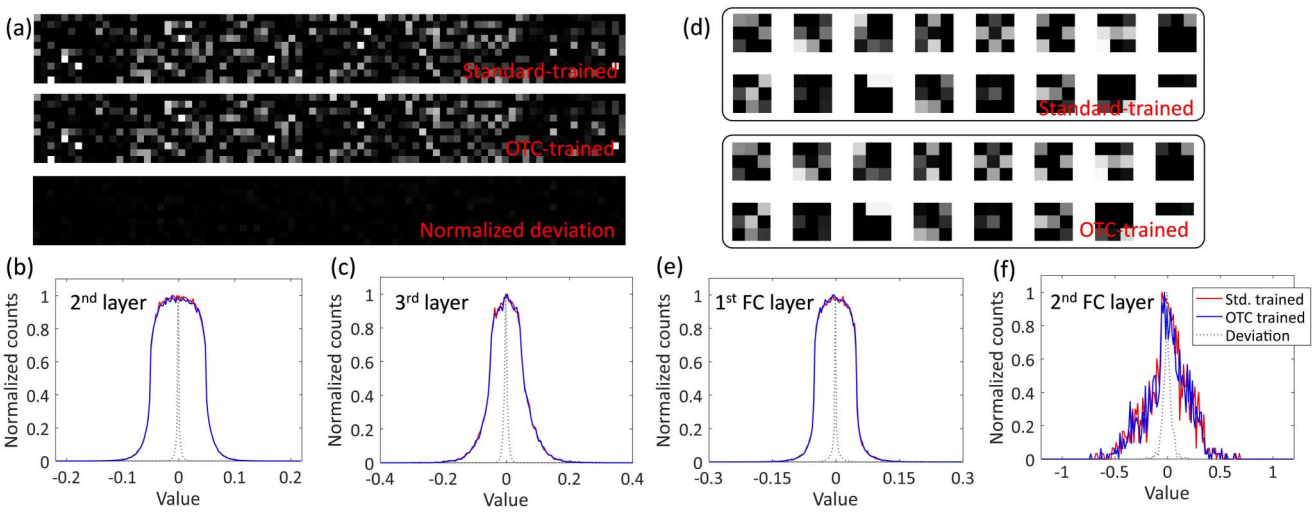

Fig. 4. Parameter visualization of the trained neural networks. (a) Trained parameters of the fourth layer in the FC network model. The standard-trained parameters are provided for reference, and the normalized deviation is depicted. (b) and (c) Distributions of trained parameters and deviations of the second and third layers of the FC network. The counts are normalized by the maximal counts. (d) Trained kernels of the first convolutional layer in the convolutional network. (e) and (f) Distributions of trained parameters and deviations of the first and second FC layers of the convolutional network. (b), (c), (e), and (f) share the same figure legends.

We visualize the trained parameters in Fig. 4 to study the impact of the OTC on the neural network training. The parameters of the OTC training and standard training are initialized with the same random seeds so that they converge to similar optimums. The standard-trained and the OTC-trained parameters of the fourth layer of the FC network are illustrated in Fig. 4(a). The parameters of the fourth FC layer form a 10×86 matrix. It is found that the OTC trained parameters have small deviations compared with the standard-trained parameters. The absolute deviations (magnified five times) are depicted. The stochastic distributions of the parameters and deviations in the second and third FC layers are shown in Figs. 4(b) and 4(c). We observe that the distribution of the OTC-trained parameters overlaps with that of the standard-trained ones. The deviations are small and concentrate at zero. It is inferred that the OTC has a fairly minor impact on FC network training. Figure 4(d) shows the convolutional kernels of the first convolutional layer, trained by the OTC and standard MBGD, respectively. The deviations between these two sets of kernels are unnoticeable. In Figs. 4(e) and 4(f), the parameter distributions and deviation distributions of the cascading FC layers are depicted. The deviations also concentrate at zero, implying that there is a minor impact from the OTC on convolutional network training. It is worth remembering that the OTC-trained inference accuracy is the same as the standard-trained one (as shown in Fig. 3). Therefore, the minor impact imposed by the OTC does not cause noticeable deterioration to the effectiveness of neural network training.

4. Conclusion

In summary, OTC architecture is proposed for neural network training. The linear operations of neural network training are conducted by a DPU array, where all optical components are waveguide-based for photonic integration. In view of the HDEA principle, the OTC architecture adopts high-speed optical components for linear operations and low-speed electronic devices for nonlinear operations of neural networks. According to the results of SPICE circuit simulation, large electronic leakage time constant (over 100 ns) allows the dot-product calculation of massive vectors (length over 1000) to be conducted by the HDEA. To solve the problems of insertion loss and crosstalk of the data-feeding waveguide crossings, dual-layer waveguide topology is applied for the data feeding. The ultra-low crossing loss and crosstalk enable a large-scale dot-product array. The 2D planar design of the OTC eradicates the demand for the third space dimension or the lens structures, potentially featuring high compactness and immunity to aberration. Simulation results show that neural network training with the OTC is effective, and the accuracies are equivalent to those of the standard training processes on digital computers. Through analyzing the trained parameters, we observe that the OTC training leaves minor deviations on the parameters compared with the standard processes without any apparent accuracy deterioration. In practice, the optical and electro-optic components including push–pull modulators, splitters, ILCs, waveguides, and photo-detectors suffer from fabrication deviations. These deviations affect the numerical accuracy of the OTC and may result in performance degradation of the trained neural networks. However, the OTC training is an *in-situ* training scheme, of which the training results are potentially robust to hardware imparities, as recently demonstrated in in-memories computing research[20]. In future study, investigation about the OTC's robustness to hardware imparity based on fabricated OTC chips is of great interest.

Acknowledgement

This work was supported by the National Key R&D Program of China (No. 2019YFB2203700) and the National Natural Science Foundation of China (No. 61822508).

References

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).
2. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," Science **362**, 1140 (2018).
3. J. Shi, F. Zhang, D. Ben, and S. Pan, "Photonic-assisted single system for microwave frequency and phase noise measurement," Chin. Opt. Lett. **18**, 092501 (2020).
4. R. Wang, S. Xu, J. Chen, and W. Zou, "Ultra-wideband signal acquisition by use of channel-interleaved photonic analog-to-digital converter under the assistance of dilated fully convolutional network," Chin. Opt. Lett. **18**, 123901 (2020).
5. S. Xu, X. Zou, B. Ma, J. Chen, L. Yu, and W. Zou, "Deep-learning-powered photonic analog-to-digital conversion," Light: Sci. Appl. **8**, 66 (2019).
6. L. Yu, W. Zou, X. Li, and J. Chen, "An X- and Ku-band multifunctional radar receiver based on photonic parametric sampling," Chin. Opt. Lett. **18**, 042501 (2020).
7. D. Amodei and D. Hernandez, "AI and compute," https://openai.com/blog/ai-and-compute/#addendum (2018).
8. M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE International Solid-state Circuits Conference* (2014), p. 10.
9. M. A. Nahmias, T. F. Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 7701518 (2020).
10. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photon. **11**, 441 (2017).
11. S. Xu, J. Wang, R. Wang, J. Chen, and W. Zou, "High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays," Opt. Express **27**, 19778 (2019).
12. V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. Lima, H. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," IEEE J. Sel. Top. Quantum Electron. **26**, 7701213 (2020).
13. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y. Chen, P. Chen, G. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**, 1132 (2019).
14. 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).
15. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X **9**, 021032 (2019).

16. J. Chiles, S. Buckley, N. Nader, S. Nam, R. P. Mirin, and J. M. Shainline, "Multi-planar amorphous silicon photonics with compact interplanar couplers, cross talk mitigation, and low crossing loss," APL Photon. **2**, 116101 (2017).
17. S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: efficient primitives for deep learning," arXiv:1410.0759 (2014).
18. J. Chiles, S. M. Buckley, S. Nam, R. P. Mirin, and J. M. Shainline, "Design, fabrication, and metrology of 10 × 100 multi-planar integrated photonic routing manifolds for neural networks," APL Photon. **3**, 106101 (2018).
19. J. Lee, S. Cho, and W. Choi, "An equivalent circuit model for a Ge waveguide photodetector on Si," IEEE Photon. Technol. Lett. **28**, 2435 (2016).
20. P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," Nature **577**, 641 (2020).