Advances and marine applications of optical neural network

Xiong, Jianmin, Zhang, Zejun, Xu, Jing

Jianmin Xiong, Zejun Zhang, Jing Xu, "Advances and marine applications of optical neural network," Proc. SPIE 11763, Seventh Symposium on Novel Photoelectronic Detection Technology and Applications, 117631O (12 March 2021); doi: 10.1117/12.2586296

Event: Seventh Symposium on Novel Photoelectronic Detection Technology and Application 2020, 2020, Kunming, China

Advances and marine applications of optical neural network

Jianmin Xiong[a,b], Zejun Zhang[a,b,c], Jing Xu[a,b,c,*]

[a] Optical Communications Laboratory, Ocean College, Zhejiang University, Zheda Road 1, Zhoushan, Zhejiang, 316021 P.R. China;

[b] Key Laboratory of Ocean Observation-Imaging Testbed of Zhejiang Province, Ocean College, Zhejiang University, Zheda Road 1, Zhoushan, Zhejiang, 316021, China;

[c] The Engineering Research Center of Oceanic Sensing Technology and Equipment, Ministry of Education, China

ABSTRACT

With the development of artificial intelligence technology, such as artificial neural networks, the increasing demand for computing drives the upgrading of computing accelerators. It's known that the semiconductor process is approaching physical limits and the Von Neumann architecture of storage-computing separation affects the computing efficiency, which both lead to the gradual failure of electronic devices to meet application requirements. Optical neural networks (ONNs) can take full advantage of high speed, high bandwidth, high parallelism, and low power consumption of optical transmission to overcome the deficiencies of electronic devices. In this paper, we summarize and analyze previous researches on optical neural networks according to different physical implementations. And we conclude that most studies apply the characteristics of special materials to realize the dense matrix multiplication and nonlinear activation function of ONNs. Less research focuses on the nonlinear characteristics inherent in the optical signal transmission to realize important components of traditional neural networks. ONNs show great potentials in analog computing and information processing, such as marine in-situ imaging and optical receiver of underwater optical communication. And ONN is possible to be a new generation of neural network accelerator. But the large-scale application of ONNs requires more studies in optical implementation of nonlinear activation function and loss function, and accuracy improvement of optical computing.

Keywords: Optical neural network, optoelectronic technology, optical computing

1. INTRODUCTION

As artificial intelligence technologies continue to evolve, the rapidly increasing demand for computing leads to the upgrading of computing hardware accelerators. Nowadays, the semiconductor process is constantly approaching its limit. The traditional von Neumann computing architecture of memory-computing separation results in a large number of tidal data loads between memory and processor during computation [1], which both reduces the computation rate and increases the power consumption. Therefore, it is necessary to explore new artificial neural network implementations that are different from electronical artificial neural networks.

Artificial neural networks based on dielectric devices are subject to many limitations in practical applications. Firstly, the parameters training of neural network takes a long time, which consumes a lot of energy, and requires large computing and storage resources; secondly, the trained model also requires a certain amount of computing and storage resources for its each operation. Considering the equipment cost and power consumption in the actual environment, it is always difficult to provide sufficient computing and storage resources for traditional artificial neural networks. However, artificial neural networks based on optical devices can overcome the above existing deficiencies. Compared with electrical transmission, optical transmission has obvious advantages such as high speed, high bandwidth, high parallelism, low latency, and low energy consumption. Optical transmission can be used in artificial neural networks to alleviate the need for storage and the integration of storage. Computing can be realized in model training and actual operation by taking advantage of the high-speed characteristics of optical transmission, which can greatly improve the computation rate without von Neumann computing architecture.

*jxu-optics@zju.edu.cn

Seventh Symposium on Novel Photoelectronic Detection Technology and Applications, edited by
Junhong Su, Junhao Chu, Qifeng Yu, Huilin Jiang, Proc. of SPIE Vol. 11763, 117631O
© 2021 SPIE · CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2586296

Parallel direct processing of optical signals can avoid analog-to-digital conversion and series-parallel conversion so that it can further reduce the overall system power loss and operating delay.

The implementation of ONN benefits from the optical decomposition of matrix operations. The matrix computing in ONN is mainly based on the implementation of the triangular decomposition algorithm proposed by Reck et al. [2]. The triangular decomposition algorithm takes advantage of the property that passive and dissipative devices can transmit matrices, and it was proved that the triangular network built by phase shifter and beam splitter can be composed of any size of unitary matrix [2]. In 2016, Clements et al [3] further extended the triangular decomposition algorithm in combination with singular value decomposition to implement an arbitrary matrix, which using Mach-Zehnder interferometer (MZI) and variable optical attenuator (VOA) to implement the unitary and diagonal matrices respectively.

In 1985, Farhat et al. [4] first proposed an optical interconnection method to implement artificial neural networks and verified that the optical implementation method has significant advantages over traditional artificial neural networks in terms of energy efficiency, parallel computation, and implementation of multiple interconnections. Later studies [5][6][7] had explored optical methods to implement artificial neural networks, but nonlinear activation functions have been commonly realized with the assistance of electronic devices for a long time due to the difficulty in implementing nonlinear activation functions through integrated high-power laser and optical devices.

Optical Neural Network (ONN) is a kind of artificial neural networks implemented by physical methods using optical (electrical) devices [8]. ONN can be classified according to the physical implementation method, as shown in Figure 1.

- ONN
  - Hybrid Optoelectronic NN
  - All-optical NN
    - Passive Diffraction NN
    - Nanophotonic NN
    - Optical Scattering NN

Figure 1 The classification of ONN

ONN can be divided into Hybrid Optic-electronic Neural Network (HONN) and All-optical Neural Network (AONN) according to different physical implementation methods.

2. HYBRID OPTIC-ELECTRONIC NEURAL NETWORK (HONN)

HONN is a kind of ONNs that applies photoelectric conversion devices to transmit data during the computation. Its main feature is the usage of optical devices for partial layer manipulation, and then the results of the optical computation are fed into the traditional neural network for the remaining layers via photoelectric conversion.

In 2016, Chen et al. [9] used a bionic angle-sensitive sensor (Angle Sensitive Pixel, ASP) for the first time to replace the first convolutional layer of a convolutional neural network. This new hardware and algorithm combined with an optoelectronic hybrid network structure are called ASP Vision. ASP is a diffractive sensor that performs edge filtering of the image using optical convolution. With the ability to perform both image acquisition and image filtering simultaneously, the sensor can significantly save system power, reduce data bandwidth, and decrease floating-point operations in a processor.

In 2018, Chang et al. [10] proposed an optoelectronic hybrid neural network based on diffractive optics devices. The network implements optical convolution operations by adding a layer of phase plane before electronic computation, which can greatly reduce the computational load of the entire network. The "4f system" is composed of two convex lenses with focal lengths of f. The two Fourier transforms can be cascaded. In this system, there are two convex lenses with a focal length of f. The phase plane is divided into several flat convolution kernels, which can modulate the amplitude and phase of the incident light, thus realizing a multi-convolution process. The amplitude can be modulated by the transmittance of the phase plate, while the phase information can be modulated by the thickness of the phase plane.

In 2020, Miscuglio et al. [11] proposed a 4-bit 4×4 photonic tensor core (PTC) that implements passive parallel computation of tensors by optical methods. The core principle of the PTC is the Wavelength Division Multiplexing (WDM) technique of optical signals and Phase Change Memory (PCM) based on $Ge_2Sb_2Se_5$. The PTC structure is shown in Figure 2 below. Four bits of data are loaded on optical signals of different wavelengths, which are modulated by WDM. Then signals transmit in the waveguide and are divided into signals of four wavelengths by Micro-ring Resonators (MRR). And the four signals pass through the PCM respectively. Eventually, the photodiode collects all light signals. The output electrical signal is the result of tensor multiplication and accumulation. The light transmittance of PCM can be selectively changed by electrothermal conversion of tungsten electrodes at both ends of PCM, which affects the crystalline-amorphous state of $Ge_2Sb_2Se_5$. The parameters of the neural network are stored in the light transmittance of PCM.

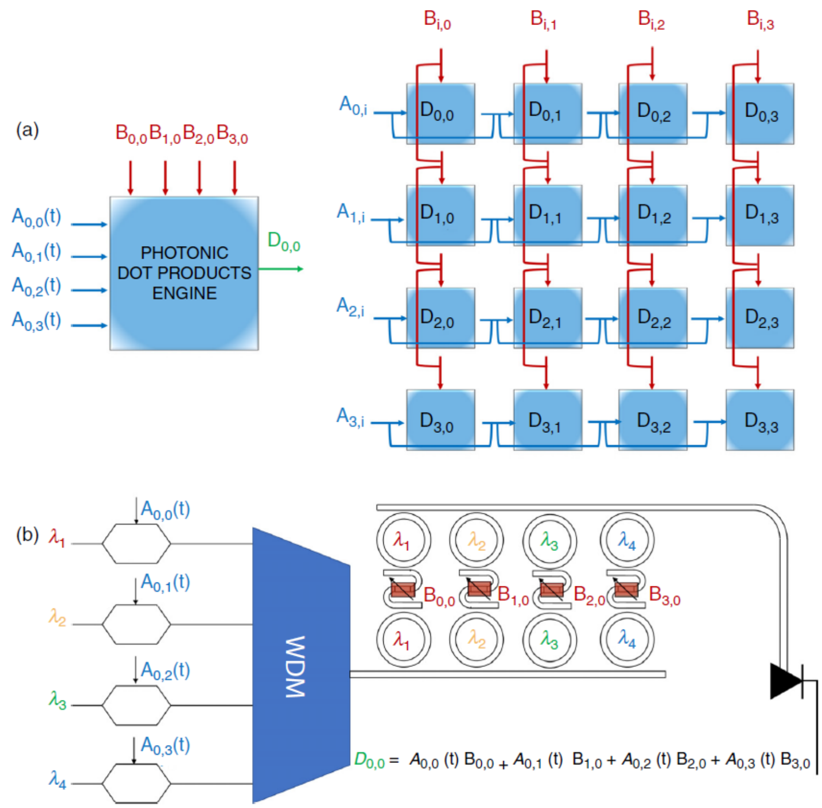

Figure 2 The Structure of PTC [11]. (a) The structure of the photonic dot product engine. (b) The structure of one unit of PTC

WDM modulation enables parallel processing of the calculations, and matrix multiplication is implemented at PCM. The electro-thermal conversion of the tungsten electrodes adjusts the light transmittance of the phase change material. And the output of the optical signal after interacting with the PCM can be used as the result of two-matrix multiplication. The parameters are stored at the PCM and the calculations are also performed at the PCM, thus realizing the integration of storage and computing. The optical signals pass through MRR both before and after PCM, and the function of MRR is to realize the frequency division of optical signals. The purpose of optical signals passing through the PCM before MRR is to avoid the disturbance of different frequencies of optical signals during calculating at PCM. The aim of optical signal passing through the PCM after the MRR is to reduce the scattering distortion of the calculated optical signals. Both the settings above are to improve the accuracy of optical calculation.

By implementing some layers of traditional neural networks, HONNs can perform more intensive matrix multiplication operations in traditional neural networks at the speed of light, which share some of heavy computational load to a certain degree. However, photoelectric conversion introduces additional system delays and optical power losses, which weakening the high rate computational advantages of optical neural networks. Therefore, it is a transitional implementation method.

3. ALL-OPTICAL NEURAL NETWORK (AONN)

AONN is a kind of ONNs that does not apply photoelectric conversion devices to transmit data in the process of computing. All the computing is realized through optical devices. According to the main use of optical devices, AONN can be divided into three types of neural networks, which are nanophotonic neural network, passive diffraction neural network and optical scattering neural network.

In 2017, Shen et al. [12] proposed a new architecture for an AONN using a cascaded array of 56 MZIs in a silicon photonic integrated circuit to form a programmable nanophotonics processor (PNP) for speech recognition. The architecture consists of an input layer, hidden layers, and an output layer. The hidden layers include an Optical Interference Unit (OIU) and an Optical Nonlinear Unit (ONU), which act as a matrix multiplier and a nonlinear activation function respectively. The MZI consists of two 3 dB directional couplers in front and back connecting the upper and lower silicon waveguide branches. The inner phase shifter controls the output spectral ratio by changing the refractive index of the waveguide, while the outer phase shifter controls the differential output and phase delay. ONU can be implemented by optical devices with nonlinear characteristics such as saturated absorbers.

In 2018, Hughes et al. [13] [14] proposed a method of efficient and local training of ONN. Adjoint Variable Method (AVM) is used to obtain the parameters of an optical path in backward propagation, which is similar to the way traditional neural networks computing gradients. The training of the ONN operates in the optical circuit with a tunable beam separator. And the parameters are adjusted by changing the settings of the optical phase shifter. The laser beams encoding the information to be processed are launched into the optical circuit and transmitted by the optical waveguide through the beam splitter. After each iteration is completed, a loss value is calculated based on the root-mean-square error function. A new optical signal will be generated based on the obtained accompanying variables in feed-forward propagation, the error function value, and its differentiation. Then the new optical signal will be sent back through this optical network in the opposite direction. By measuring the light intensity around each beam separator, it can detect how the performance of the neural network changes with each beam shifter setting in parallel. The phase shifter settings can be changed based on the parameters of the propagation process, and the process can be repeated until the neural network outputs the desired results.

In 2018, Lin et al. [15] proposed an all-optical diffraction deep learning framework called Diffractive Deep Neural Network (D^2NN), and the network structure is shown in Figure 3.a. The D^2NN consists of multiple layers of diffractive surfaces. By cooperating with these diffraction surfaces, it is possible to achieve computational functions in the form of photons. The training of D^2NN is completed on an electronic computer. After the training is finished, 3D models of each diffraction layer of the network are generated by Poisson Surface Reconstruction using the parameters of D^2NN. Then the models are printed by a 3D printer. The obtained neural network diffraction layers are shown in Figure 3.b and 3.c. Each printed layer can be viewed as a projective or reflective layer. And the points on a layer represent neurons capable of reflecting and transmitting light waves. These points are connected to subsequent layers by optical diffraction, which enabling a forward propagation process that can perform a variety of complex function operations at the speed of light.

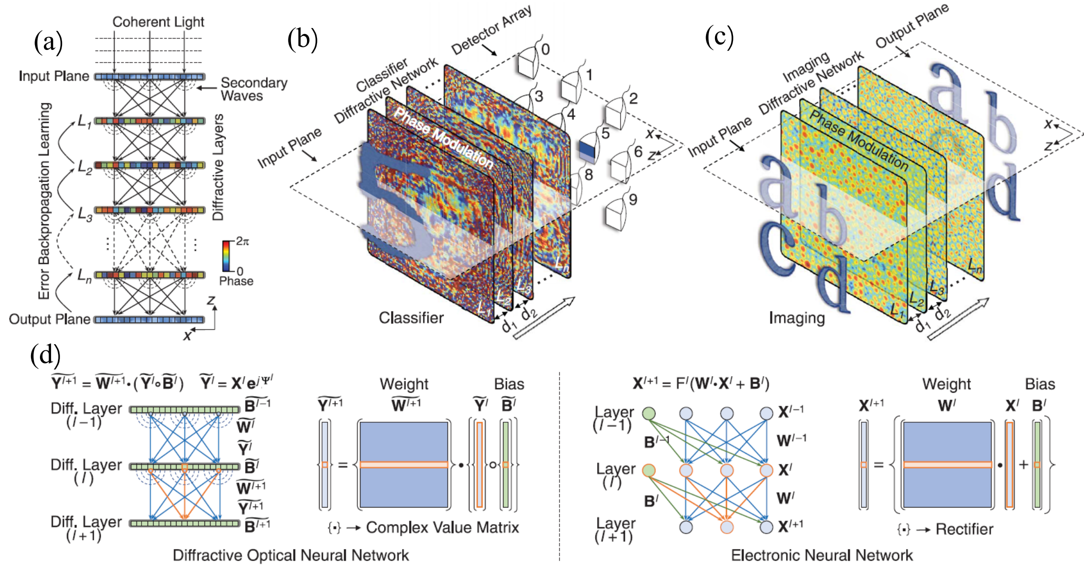

Figure 3 The Structure of D²NN [13]. (a) The overall structure of D²NN. (b) The usage of D²NN for image classifier. (c) The usage of D²NN for imaging. (d) The difference between diffractive ONN and electronic NN

The inputs of the neurons in D²NN are complex values. The weighting parameters are determined by diffraction in free space and related to the physical distance between layers. And multiplicative biases are determined by the transmission or reflection coefficient of each neuron, as shown in Figure 3.d. Each neuron in D²NN participates in modulating the amplitude and phase of input optical waves, and each neuron will output a secondary wave. The output of each neuron in the previous layer provides unique interconnectivity to the network through wave propagation and coherent or partially coherent interference coupling. Although the secondary wave generated by the neurons diffracts to all angles, the intensity of the wave attenuates with propagation distance. Each neuron in the next layer can only receive waves within a finite radius. And light leaking outside the finite radius will cause optical power loss in the entire network. The interconnectivity of D²NN is related to the network layer spacing, the incident beam light intensity, the detection SNR, the coherent length, and the radius of the light source [13].

The authors of this paper trained the classifier and imager on the MNIST dataset and ImageNet dataset respectively. Testing of designed D²NN containing 5 diffraction layers resulted in 91.75% classification accuracy on the classifier while adding two more layers to the 5-layer network resulted in 93.39% classification accuracy. As for the imager, the D²NN of the trained 5-layer diffraction layer could map the signal at the output plane to the same position as the input plane, which is completely different from the effect of signal diffraction in free space.

AONNs need to solve three problems of intensive matrix arithmetic, implementation of nonlinear activation functions, and model trainability simultaneously. D²NN can indeed efficiently compute the most common matrix arithmetic in neural networks, but since no nonlinear activation function is introduced in D²NN, it is not truly an implementation of neural networks. Trainable optical neural networks load practical application effects directly into the optimization process of the structure, which makes it easier and more reliable to optimize the structure. And due to the ability to introduce nonlinear terms directly, the nonlinear activation functions in the neural network can be effectively implemented. Currently, AONNs still have a critical challenge, which is the inability to directly implement convolutional computation using all-optical way. The output of optical convolution is still a tiled two-dimensional image instead of stacked eigenvalues at high latitudes. which makes the advantage of all-optical neural networks over traditional neural networks not outstanding, and it is still a great challenge to achieve efficient convolutional computation at the all-optical level and fully recover convolutional neural networks.

4. APPLICATIONS AND CHALLENGES

ONN shows great potentials in analog computing and information processing due to high speed, high parallelism, low energy consumption, and low latency of optical transmission. Many studies mention different structures of neural networks, such as convolution neural networks [16], recurrent neural networks [17] [18], and spiking neural networks [19] [20] [21]. But the fact is that large-scale application of ONNs currently still has many challenges, which requires more subsequent researches to overcome one by one.

4.1 Intelligent marine in-situ imaging

Conventional underwater in-situ imaging techniques require various types of cameras or detection devices to obtain image data from the surrounding underwater environment. But these devices do not work underwater for long periods due to power consumption and other factors. Additionally, the redundancy of the directly acquired images is high. The transmission of all the image data will increase the transmission burden and cost, so some preprocessing of the acquired image data at the sensing end is needed. In 2020, Mennel et al. [22] proposed a method of etching neural networks in an image sensor and produced a 3×3 intelligent image sensor. While collecting light signals, the acquired image can be used for edge extraction and other operations at the sensor automatically, which greatly improves the frame rate of image acquisition and decreases the redundancy of the image data during the transmission. The sensor can get power by photoelectric conversion so that it works without external power, but the sensor still requires an external high-voltage circuit for setup, which is limited in practical applications. However, this study has great inspiration for the application of ONNs in marine in-situ imaging. ONNs can process optical signals directly without analog-digit conversion, so there is no need of any external circuit for setup. By taking advantage of optical transmission, ONNs act as both a sensor and a processor. So imaging devices integrated ONNs can operate underwater in a low power consumption status for longer periods, which is truly an implementation of marine in-situ imaging.

4.2 Intelligent receiver of underwater optical communication

The absorption and scattering of the optical signal by water as well as the turbulence and bubbles in the water will seriously affect the transmission rate of underwater optical communication. And in most cases, it is necessary to adopt signal processing methods at the receiver to minimize the influence of water on the optical signal. But it will also increase the power consumption of the receiver and the system delay. In 2020, Yu et al. [23] proposed a binary coherent optical receiver to achieve the recovery of modulated signals from the transmitter. Signals after QPSK modulation pass through the input layer into the binary AONN. Then the output optical signals of binary AONN pass through the photodiode and analog-to-digital converter into the electrical neural network. The electrical neural network is used to recovery the modulated signals. AONN in the receiver is equivalent to the optical domain filter. Its function is to recovery distorted optical signals during transmission. It can reduce the burden of signal processing and the power consumption of the system. Additionally, since the system uses a 1-bit precision analog-to-digital converter for complex modulated signals recovery, the cost of the optical receiver can be greatly reduced. Hence, ONNs will have great power in underwater optical communication to overcome complicated situations in the water.

4.3 Brand new neural network computing accelerator

The need for computing of deep learning is endless, but due to the energy consumption and physical limits of the electrical components, they are gradually unable to meet the demands of deep learning tasks. Therefore, there is an urgent need to develop new neural network accelerators. Optical neural networks can greatly accelerate the training speed of models and significantly shorten the training time and decrease energy consumption by taking advantage of the superior characteristics of optical transmission. Additionally, unlike traditional neural network accelerators, ONNs do not rely on external memory to store parameters, which can lower the cost. The cost of an electric neural network accelerator is mainly due to a large number of memory and computing units. And the best electric neural network accelerator can cost thousands of dollars, such as NVIDIA GeForce RTX 3090. ONN is still in need of further development, but desirable characteristics of optical transmission make it more suitable for high-speed parallel computing. ONN will be more competitive in the future than GPUs etc. in terms of parallel computing and cost.

4.4 Challenges

The implementation of ONN generally contains three major components: matrix multiplication, nonlinear activation function, and backpropagation training. There are many previous studies on the three topics, but there are still some challenges that need to be solved.

(1) The optical implementation of nonlinear activation function

Artificial neural networks consist of artificial neurons, and the cores of the mathematical model of the neuron are matrix multiplication and nonlinear activation function. Matrix multiplication can be implemented by a variety of passive lossless optical devices, such as MZI. Most studies of nonlinear activation function of artificial neurons use materials such as saturated absorbers to fit the characteristics of nonlinear function by changing the material properties of light. But the use of special materials will also cause a partial loss of optical power, which affects the accuracy of subsequent calculations. However, there is few study focusing on the nonlinear phenomena that occur in the transmission process of the optical signal itself. Therefore, finding a way to fit the nonlinear activation function from nonlinear characteristics of optical signals is needed.

(2) The optical implementation of loss function

The performance of artificial neural networks depends on the evaluation of the loss function. Due to different tasks oriented, there are many kinds of loss functions. Some common loss functions are absolute error function, root mean square error function, cross-entropy error function, etc. How to use optical methods to implement some common loss functions is also one of the problems that need to be solved.

(3) Accuracy improvement of optical computing

The strong fitting problem capability of artificial neural networks benefits from high computational accuracy of dielectric devices. However, optical computing belongs to analog computing so that there will be power loss and scattering in the propagation process of optical signals, which leads to the limited accuracy of the current computation. Improving the accuracy of optical computing is conducive to improving the fitting and generalization capabilities of ONNs to some degree. However, the improvement of accuracy will also bring the problem of increased complexity of the calculation, so accuracy and system latency need to have a good balance.

5. CONCLUSION

ONN is a product of the intersection of optoelectronics technology and artificial intelligence, which has been studied for 35 years. ONNs can break through the bottlenecks of electrical device neural networks with the combination of the high-speed, high-bandwidth, low-latency, and low-power characteristics of optical transmission and the strong problem-fitting capabilities of artificial intelligence. In this paper, we summarize and analyze previous researches on optical neural networks according to different physical implementations. And we conclude that most studies apply the characteristics of special materials to realize the dense matrix multiplication and nonlinear activation function of ONN. Less research focuses on the nonlinear characteristics inherent in the optical signal transmission to realize important components of traditional neural networks. ONN shows great potentials in analog computing and information processing, such as marine in-situ imaging and optical receiver of underwater optical communication. And ONN is possible to be a new generation of neural network accelerator. But the fact is that large-scale application of ONNs currently still has many challenges, which requires more subsequent researches to overcome one by one. The large-scale application of ONNs requires more studies in optical implementation of nonlinear activation function and loss function and accuracy improvement optical computing.

6. ACKNOWLEDGMENT

This work was supported by National Key Research and Development Program of China (2016YFC1401202, 2017YFC0306601, 2017YFC0306100).

REFERENCES

[1] Zaharia, M., Chowdhury, M., Das, T., "Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing," Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation, 15-28 (2012).

[2] Reck, M., Zeilinger, A., Bernstein, H. J., "Experimental realization of any discrete unitary operator," Physical Review Letters, 73(1), 58-61 (1994).

[3] Clements, W. R., Humphreys, P. C., Metcalf B. J., "Optimal design for universal multiport interferometers," Optica, 3(12), 1460-1465 (2016).
[4] Farhat, N. H., Psaltis, D., "Optical implementation of the Hopfield model," Applied Optics, 24(10), 1469-1475 (1985).
[5] Wagner, K., Psaltis, D., "Multilayer optical learning networks," Applied Optics, 26(23), 5061-5076 (1987).
[6] Owechko, Y., "Optoelectronic resonator neural networks," Applied Optics, 26(23), 5104-5111 (1987).
[7] Anderson, D., "Material Demands for Optical Neural Networks," MRS Bulletin, 13(8), 30-35 (1988).
[8] Chen, H., Yu, Z., Zhang, T., "Advances and challenges of optical neural networks," Chinese Journal of Laser, 47(05), 80-91 (2020).
[9] Chen, H. G., Jayasuriya, S., Yang, J., "ASP Vision: Optically Computing the First Layer of Convolutional Neural Networks Using Angle Sensitive Pixels," Proce. CVPR, 903-912 (2016).
[10] Chang, J., Sitzmann, V., Dun, X, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Science Report, 8(1), 12324 (2018),.
[11] Miscuglioa, M., Sorgerb, V. J., "Photonic tensor cores for machine learning," Applied Physics Review, 7, 031404 (2020).
[12] Shen, Y., Harris, N., Skirlo, S., "Deep learning with coherent nanophotonic circuits," Nature Photonics, 11, 441–446 (2017).
[13] Hughes, T. W., Minkov, M., Shi, Y., "Training of photonic neural networks through in situ backpropagation and gradient measurement," Optica, 5, 864-871 (2018).
[14] Hughes, T. W., Minkov, M., Williamson, I. A., "Adjoint method and inverse design for nonlinear nanophotonic devices," ACS Photonics, 5(12), 4781–4787 (2018).
[15] Lin, X, Rivenson, Y, Yardimci, N. T., "All-Optical Machine Learning Using Diffractive Deep Neural Networks," Science, 361(6406), 1004-1008 (2018).
[16] Bagherian, H., Skirlo, S., Shen, Y. C., "On-chip optical convolutional neural networks," 16 Aug 2018, https://arxiv.org/abs/1808.03303 (24 Octobor 2020).
[17] Feldmann, K., Vandoorne, J., Dambre, D., "Parallel Reservoir Computing Using Optical Amplifiers," IEEE Transactions on Neural Networks, 22(9), 1469-1481 (2011).
[18] Paquot, Y., Duport, F., Smerieri, A., "Optoelectronic Reservoir Computing," Science Report, 2, 287 (2012).
[19] Maass, W., "Networks of spiking neurons: The third generation of neural network models," Neural Networks, 10(9): 1659–1671 (1997).
[20] Shastri, B., Nahmias, M., Tait, A., "Spike processing with a graphene excitable laser," Science Report, 6, 19126 (2016).
[21] Feldmann, J., Youngblood, N., Wright, C. D., "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature, 569(7755), 208–214 (2019).
[22] Mennel, L., Symonowicz, J., Wachter, S., "Ultrafast machine vision with 2D material neural network image sensors," Nature, 579, 62–66 (2020).
[23] Yu, Z., Zhao, X., Yang, S., "Binarized Coherent Optical Receiver Based on Opto-Electronic Neural Network," IEEE Journal of Selected Topics in Quantum Electronics, 26(1), 1-9 (2020).