Physics – Uspekhi **66** (12) 1211 – 1223 (2023)

REVIEWS OF TOPICAL PROBLEMS

PACS numbers: 07.05.Mh, 42.79.Hp

Photonics approaches to the implementation of neuromorphic computing

A I Musorin, A S Shorokhov, A A Chezhegov, T G Baluyan,
K R Safronov, A V Chetvertukhin, A A Grunin, A A Fedyanin

DOI: https://doi.org/10.3367/UFNe.2023.07.039505

Contents

Abstract. Physical limitations on the operation speed of electronic devices has motivated the search for alternative ways to process information. The past few years have seen the development of neuromorphic photonics — a branch of photonics where the physics of optical and optoelectronic devices is combined with mathematical algorithms of artificial neural networks. Such a symbiosis allows certain classes of computation problems, including some involving artificial intelligence, to be solved with greater speed and higher energy efficiency than can be reached with electronic devices based on the von Neumann architecture. We review optical analog computing, photonic neural networks, and methods of matrix multiplication by optical means, and discuss the advantages and disadvantages of existing approaches.

Keywords: neuromorphic photonics, artificial intelligence, machine learning, reservoir computing, matrix–vector multiplication, photonic computing, neural networks, optical coprocessor, photonic tensor computing, optical Fourier transform, integrated photonics, Mach–Zehnder interferometer, ring resonators, waveguides

1. Introduction

The 21st century is inextricably related to the information technology (IT) industry. The explosive increase in the volume of information after the advent of social networks, the trend towards cloud-based data storage, the development of Internet resources and the entertainment sector, and increased security of banking and financial transactions stimulated the development of new methods and algorithms for processing and transmitting data, including with the use of light [1]. The rapid expansion of IT in all areas of activity requires a considerable amount of computing resources and power, which in turn gives rise to urgent tasks to accelerate processors, develop new computer architectures, reduce energy consumption, and miniaturize systems [2–5]. A number of these problems can be solved by using photonics [6], which is attracting the attention of scientists due to the high frequency of electromagnetic waves, a wide bandwidth, and the possibility of parallelization. In addition, progress in industrial manufacturing technologies for microprocessors and optoelectronic components has led to the appearance of private companies ready to design integrated photonic chips. Such devices are in demand because of the formation of the optical computing market due to the increasing number of customers interested in upgrading their fiber optic communication lines and data centers. An increase in the number of consumers leads to economic growth in this area, and mass production reduces the production costs for photonic elements. Mathematical operations can then be performed faster and implemented more easily and cheaply, not electronically in digital form, but using analogue optical signal processing. This stimulates the development of neuromorphic photonics [7].

Photonics methods can be successfully used to complement electronic processors. A light signal, especially in free space, has physical properties that are valuable for computation. With the help of optical systems, using lenses, it is possible to carry out the Fourier transform, use the phenomenon of interference to add complex quantities, use the phenomenon of diffraction to transform the original signal, perform a nonlinear quadratic transformation during detection, and add the intensities [8–10]. In addition, when working with wide inhomogeneous parallel beams whose cross section represents a multidimensional matrix, it is possible to operate on the entire matrix at a speed independent of its size. These properties of an optical signal carrying information allow

A I Musorin, A S Shorokhov, A A Chezhegov, T G Baluyan, K R Safronov, A V Chetvertukhin, A A Grunin, A A Fedyanin [(a)]
Lomonosov Moscow State University, Faculty of Physics, Leninskie gory 1, str. 2, 119991 Moscow, Russian Federation
E-mail: [(a)] fedyanin@nanolab.phys.msu.ru

Received 8 November 2022, revised 4 July 2023
Uspekhi Fizicheskikh Nauk **193** (12) 1284 – 1297 (2023)
Translated by S Alekseev

speeding up, first and foremost, the operation of vector–matrix and matrix–matrix multiplications, which underlie the operation and training of multilayer neural networks.

Bottlenecks in using an optical coprocessor are data input and output operations [11]. Digital-to-analogue and analogue-to-digital converters (DACs and ADCs), together with wiring, are often the limiting factors in terms of both speed and power consumption. The advantage of optical methods is the possibility of parallel processing of information and working with high-resolution images.

The purpose of this review is to describe the methods developed for photonic computation and matrix multiplication by optical means, and to consider modern approaches to the implementation of photonic neural networks.

2. Optical computation on an integrated photonics platform

2.1 Vector–matrix operations and their implementation by integrated photonics methods

Integrated photonics is one of the most promising platforms for implementing optical computations on an industrial scale. The key factor is that such devices can be built based on the hardware component of the existing microelectronics industry using well-established semiconductor technology methods. According to Yole Group research [12, 13], the integrated silicon photonics market will increase to nearly $4 billion by 2025, while hardware accelerators and photonic interposers (silicon chips whose main role is to electrically connect the tracks between the memory and the processor) can occupy a significant part of it thanks to major players such as Nvidia. At the moment, there are a number of companies producing custom silicon-based integrated photonic chips on a silicon nitride insulator, but the key difficulties still lie with lasers on a chip (combining silicon with III–V semiconductors) and combining photonics and CMOS (complementary metal–oxide–semiconductor) electronics in a single device. Many manufacturers have already achieved significant progress, as can be seen from recent announcements, e.g., by Global Foundries [14]. Among the companies specializing in solutions in the field of hardware accelerators based on integrated photonics, we note American Light-Matter, Lightelligence, and Fathom Computing, and the British Salience labs. The most common operation to be accelerated is vector–matrix multiplication, but there are also examples of combinatorial problems being solved [15]. Photonic hardware accelerators based on a platform of integrated photonics can be built directly into network solutions, for example, in large data centers, where communication between computing clusters is realized via optical interconnects. This provides an additional advantage compared with cloud computing tasks and tasks to ensure security and prevent cyber attacks [16].

2.1.1 Various architectures of photonic hardware accelerators and their features. Most often, three main architectures are distinguished for implementing vector–matrix calculations using integrated photonics: (a) a photonic network based on Mach–Zehnder interferometers (MZIs), similar photonic matrices being used in implementations of optical integrated quantum computers; (b) a photonic network based on microring resonators; (c) a crossbar array photonic network involving optical memristors, the topology of a fully connected coupling matrix based on a matrix commutator. The matrix is called fully connected because any input port can be connected to any output port, as in memristive electrical circuits. The first architecture is characterized by its versatility and low sensitivity to manufacturing errors, but it requires a large number of elements and occupies a large area on the chip, with a negative effect as regards scalability. The second architecture, based on microring resonators, can be implemented in a much more compact design; it is initially optimized for parallel operation with wavelength division multiplexing (WDM), but is very sensitive to deviations in manufacturing and requires special tuning (passive or active). The crossbar array architecture also has the advantages of working with WDM, the possibility of being implemented in a compact form factor, and not suffering from variations in manufacturing parameters, but is limited by the efficiency of cross-connects and splitters, which incur high optical loss and reduce the ability to scale the photonic circuit for real-world applications. We discuss each of these architectures in more detail.

2.1.2 Photonic matrix based on Mach–Zehnder interferometers. The first and currently most common architecture is the MZI photonic matrix, which allows realizing any predefined matrix (for example, the weight matrix for a fully connected neural network or the kernel matrix for convolution neural networks). It is known that, using the singular value decomposition, any matrix can be represented as a product of two unitary matrices and one diagonal matrix,

$$M = UZV^T,$$

where M is the original matrix, U and V are unitary matrices, and Z is a diagonal matrix. A unitary matrix can in turn be represented as a product of rotation matrices (Fig. 1), implemented using a grid of integrated MZIs [17]. This approach is also widely used in optical quantum computing problems (for example, quantum normalization) both on a chip and in three-dimensional execution on an optical table.

Each 2×2 interferometer can be described by a rotation matrix

$$T = \mathrm{i}\exp\left(\frac{\mathrm{i}\theta}{2}\right)\begin{pmatrix} \exp(\mathrm{i}\varphi)\sin\theta/2 & \cos\theta/2 \\ \exp(\mathrm{i}\varphi)\cos\theta/2 & -\sin\theta/2 \end{pmatrix}.$$

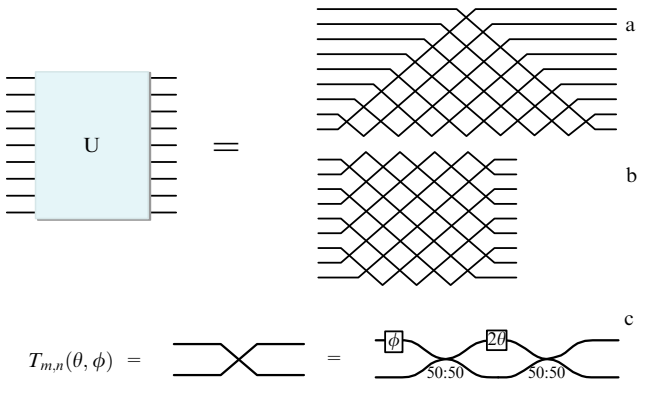

Figure 1. Representation of a unitary matrix in the form of an MZI grid. Each site is a 2×2 interferometer whose properties are determined by the external and internal phase in one of the arms (φ and θ) [17].

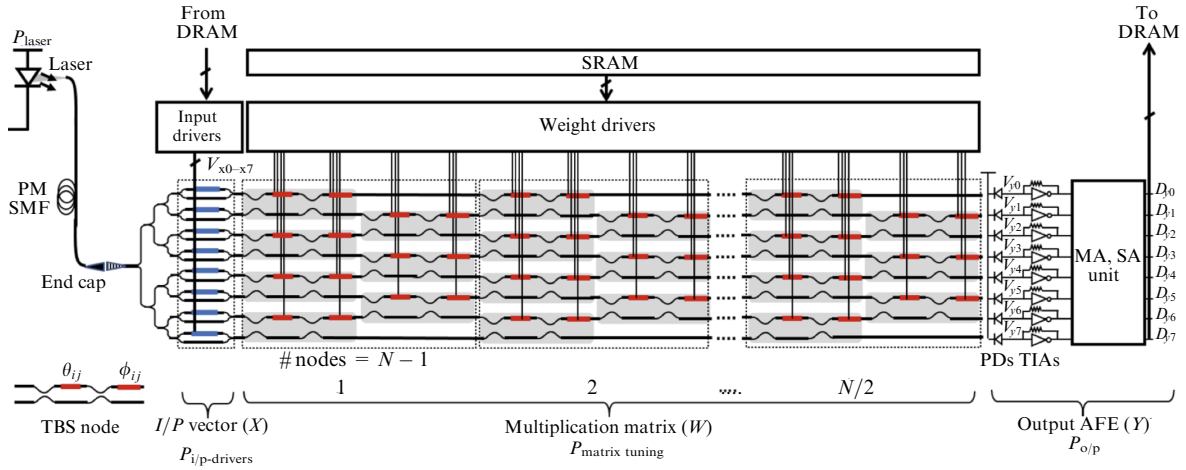

Figure 2. Schematic representation of the architecture of a photonic accelerator based on an MZI. Network of interferometers implements one unitary matrix. A single MZI unit is highlighted in gray, and modulators that set φ and θ are shown in red. Amplitude modulators that form the input vector are shown in blue [18].

Following the procedure described in [17], it is possible to build a photonic network of similar interferometers corresponding to any unitary $N \times N$ matrix [18]. In total, this requires $N(N-1)/2$ MZI elements (Fig. 2). To represent any matrix in accordance with its singular value decomposition, two such grids are required together with an intermediate layer of amplitude modulators that implement the diagonal matrix. An additional layer of phase modulators is needed to correct the first unitary matrix. This must be done in the case of a physical implementation of the entire matrix M, but one MZI network is often used in practice for sequential realization of the U and V matrices (in which case phase correction is not necessary).

In addition to the photonic MZI matrix itself, a device is also required to generate the input vector encoded in the amplitude of the light wave in each grid channel. For this, an external laser source of continuous generation is divided into the required number of inputs, in each of which an amplitude modulator is installed, controlled by an external electrical driver that uses a DAC to fix a sequence of data taken from memory. The phases of the photonic MZI matrix are also fixed by similar drivers in accordance with a similar procedure. The light passing through the photonic matrix is detected by photodiodes, which then transmit the signal through transimpedance amplifiers to the ADC. Converters return it to the digital representation and store the data in external memory.

The main limitation on the operating frequency of the device is the maximum frequency of the DAC–ADC drivers. The modulators themselves can operate at frequencies of tens, and, in the future, even hundreds of GHz. Therefore, fast and energy-efficient low-resolution DAC–ADCs are often used. Significant gains in speed and reductions in energy consumption can be achieved only for problems where the same weight matrix can be multiply reused (as in the case of convolutional neural network problems, where the convolution kernel is fixed) and a parallel computation at several wavelengths is possible. Due to manufacturing errors, together with noise and deviations in the device operation, the currently existing solutions work with low-bit numbers (up to 8 bits), but this is sufficient for a wide range of problems.

The first architecture implemented in hardware accelerators for deep machine learning tasks was proposed by the group of Marin Soljačić in [19]. As a model, a fully connected neural network was considered that classifies input data (a light signal encoding sound sequences corresponding to different letters). The network consisted of two layers corresponding to the multiplication of the input vector by weight matrices, with an intermediate nonlinear activation layer implemented separately using a standard microprocessor. The vector–matrix product was done using a photonic matrix, while one MZI network was used twice to sequentially define two unitary matrices (OIU1 and OIU2, Fig. 3). The network was pretrained using the standard backpropagation method on a computer, after which, using the found weights, the phases φ and θ were calculated for each MZI forming the photonic matrix (following the method described in [17]). The accuracy of solving the classification problem in such a system was about 77%, compared to 91% obtained using a standard central processor. Low efficiency may be associated with a number of factors, among which it is worth noting thermal induction in MZI elements from neighboring cells, the final accuracy of fixing the phases in different arms of the interferometers, detector noise during reading, and deviations of the geometric parameters of the structure from the model ones during manufacturing. Errors can be corrected in an already fabricated structure [20], which can potentially improve the accuracy of the classification problem in such devices.

Mach–Zehnder interferometers, LightMatter project. One example of the commercial implementation of this architecture was presented by the American company LightMatter, which was founded by former graduates of Soljačić's group. To date, the company has registered more than 30 patents related to photonic hardware accelerators based on integrated optoelectronic devices. The current product is the Envise processor [21], presented on the company's official website, with specifications and bench test results (Fig. 4). The main part of the device is made of two MZI-based photonic matrices, manufactured using silicon technology of the 90-nm standard. The light source is represented by an external laser; radiation is transmitted to the chip via an array of optical fibers. The optical core is combined with an electronic logical CMOS chip manufactured using 12-nm technology. This chip is based on the RISC-V architecture

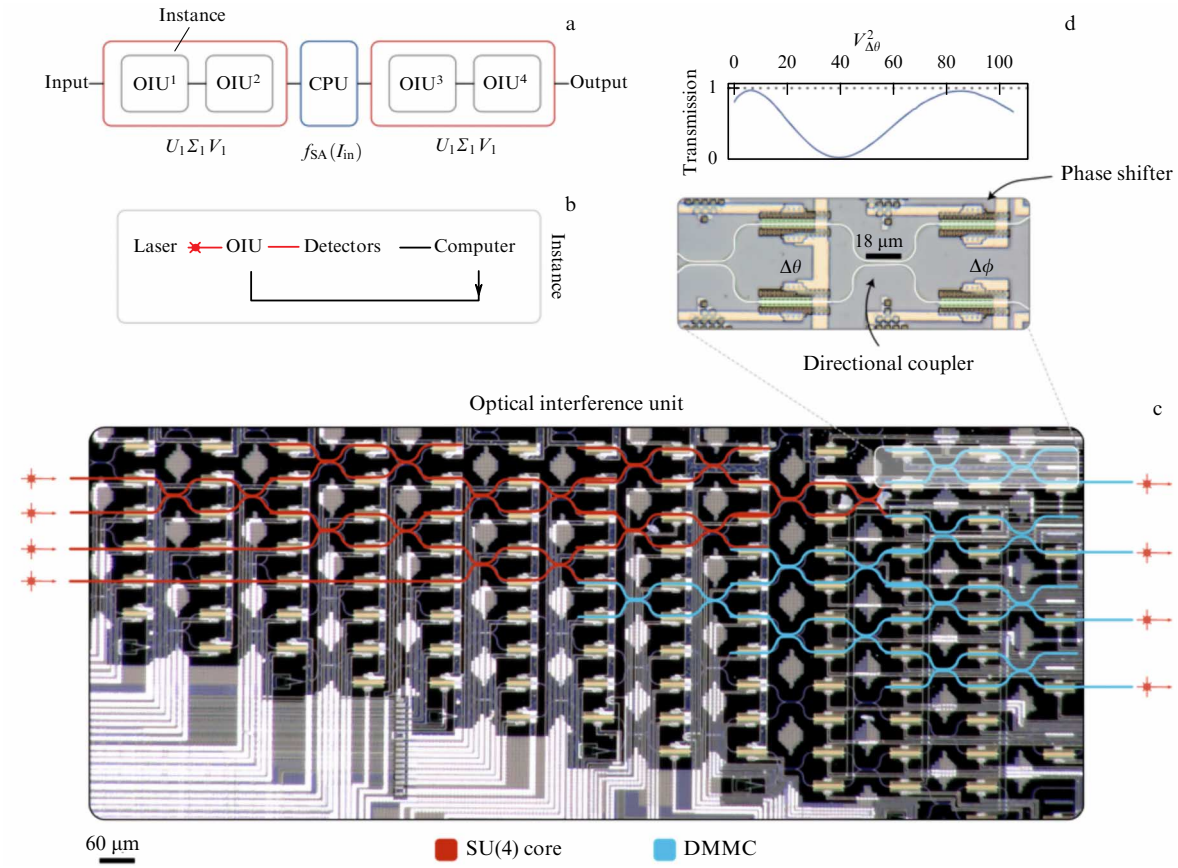

Figure 3. Example of the use of MZI architecture to implement a fully connected neural network for classification problems [19].

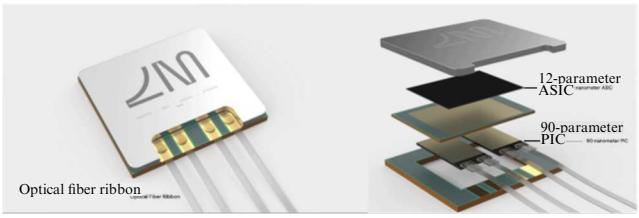

Figure 4. Optoelectronic hardware accelerator ENVISE from LightMatter [21].

and contains up to 256 computing cores. The chip also houses SRAM memory with a capacity of up to 500 MB. Data can be presented in various formats; it is possible to work with 8-bit and 16-bit representations. In addition, options are available for integration with the most popular software platforms (PyTorch and TensorFlow). Such a hardware accelerator can be used for server solutions, and the key operations are vector–matrix and tensor calculations. High operating speed and low power consumption are achieved due to a number of factors, among which we note WDM (up to eight parallel channels) and NOEMS (nano-opto-electro-mechanical system) micromechanical integrated modulators [22–24], which significantly reduce the power consumption compared to thermo-optical modulators, as well as modulators based on carrier injection, while providing a high switching frequency of up to 1 MHz. In addition, this modulator allows significantly increasing the MZI density on the chip due to its compact size (about 25 μm). We also note that, to ensure a high frequency of operations (sampling), high-speed DAC–ADC modules with medium resolution (8- or 16-bit data representation) are used. This can be justified in most machine learning and artificial intelligence applications, where a coarser representation of the data does not affect the final accuracy of the problem solution [25, 26].

Mach–Zehnder interferometers, Lightelligence project. A similar hardware solution is being developed by Lightelligence [27], which is also affiliated with Soljačić's group. As with Envise, an MZI array is used to implement a 64×64 photonic matrix. The transition between the digital electronic and analogue photonic domains is implemented using a microelectronic CMOS chip. The technology for combining chips is so-called flipchip bonding, which allows placing all the necessary elements on a single platform. The current product is the PACE (Photonic Arithmetic Computing Engine) hardware accelerator [28], operating at a system clock frequency of up to 1 GHz (Fig. 5). As with Envise, the main operation is vector–matrix multiplication, but, among the key tasks, the company highlights Max-Cut, Min-Cut, and the Ising problem, in which the authors promise acceleration up to three orders of magnitude compared to standard GPU-based solutions. A possible strategy proposed by Lightelligence to scale the technology is not the use of standard MZIs but of multimode interferometers, including those designed via inverse optimization of their shape [29].

2.1.3 Integrated optics: microresonators. In contrast to optical computing circuits with logic elements based on MZIs, the same circuits made of resonators allow not only controlling the signal with the power and/or phase of the original (or

Figure 5. PACE optoelectronic hardware accelerator from Lightelligence [28].

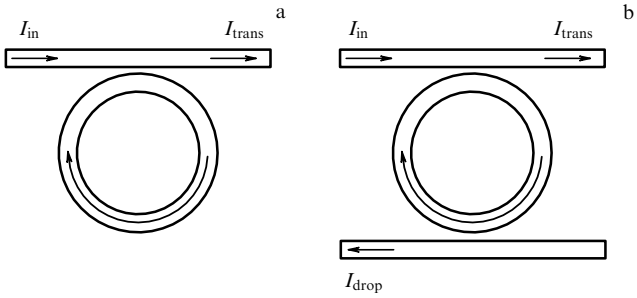

Figure 6. Two basic configurations of microring resonators: (a) simple connected microring, (b) doubly connected microring [30].

control) signal but also ensuring control and selectivity due to wavelength, in one form or another. Such schemes are often more compact than interferometric ones, because they allow one waveguide to be used for different channels encoded by the radiation wavelength.

Figure 6 illustrates typical layouts for incorporating a ring resonator into an optical circuit [30]. Direct waveguides are usually referred to as buses; the number of buses determines the mode in which the device can operate. One bus turns the microresonator into a frequency filter that does not transmit a signal in the vicinity of wavelengths corresponding to the resonator eigenvalues. The second bus, in turn, allows either the signal to be output from the resonator and transmitted further or an additional signal to be introduced into the ring, which affects the phase and power values at the output, thus making a kind of optical transistor.

For such circuits to operate efficiently on a large scale, the resonators must have high quality factors (about 10^7–10^8) and the buses must ensure low radiation losses during propagation. The quality factor of a resonator, as well as the throughput of the buses, depends on the geometric parameters (height, width, and radius) and the materials used. The choice of materials is wide, from silicon and its compounds to polymeric substances. The efficiency of signal introduction and output from the bus to the resonator and from the resonator to the bus is determined by the coupling constant, which can be controlled by changing the distance between the ring and the waveguide. All the parameters described above also depend on the wavelength of the radiation transmitting the signal. Thus, when designing resonator circuits, the parameter optimization problem must be solved and the signal radiation wavelength must be chosen.

2.1.4 Photonic crossbar matrix based on phase-change materials for mode switches. The third microarchitecture option is a crossbar array-type photonic matrix. This version of a photonic accelerator was first presented in [31] by Harish Baskaran's group. The key element of the structure is a nonvolatile photonic memory based on a phase-change material (PCM) [32]. Using such memory, the elements of the photonic matrix (transmittance and reflection coefficients) can be fixed, which allows significantly reducing energy consumption in the case of its repeated use without the need for updating, for example, in convolution problems with a known kernel (in convolutional neural networks). The photonic processor structure is shown schematically in Fig. 7. A similar architecture is used by the British company Salience labs, which was founded by graduates of the above-mentioned scientific group. A similar crossbar architecture was also mentioned in a patent [22].

In that study, the radiation source was chosen as an optical comb realized on an integrated chip based on silicon nitride under microring resonator pumping. Also, several lasers at different wavelengths could be used, with their radiation multiplexed before introducing it into an optical chip. Each channel is independently modulated in amplitude (in accordance with the values transmitted from external memory through the DAC), after which radiation at several wavelengths is introduced into each input waveguide of the photonic matrix. The input vector, encoded using the amplitude of the light wave in each horizontally oriented channel of the photonic matrix (X_1–X_4 in Fig. 7), is then divided equally between the vertical channels. For this, the division ratio of the DC dividers increases from the left edge to the right, so as to ensure the necessary fractions. It is worth noting that DC dividers are the most problematic element of the photonic matrix, because, first, they have significant dispersion and divide radiation at different wavelengths unevenly, and second, they are highly sensitive to the spread of geometric parameters during manufacturing. The second circumstance must be taken into account for the correct operation of the accelerator; in particular, a special normalization of the photonic matrix elements of the already created device must be used. New efficient dividers developed using machine learning and genetic optimization approaches could also be helpful in solving such problems.

After the input vectors are divided, they are weighed using PCM units. In different phase states, these materials have different absorptions: in amorphous states, it is typically lower than in crystalline ones. In addition, due to optical rearrangement, i.e., local heating during the propagation of nanosecond pulses through the unit, stable intermediate states with partial crystallization or amorphization can be achieved, which allows controlling the transmission in each waveguide with a certain accuracy through fixed levels (in the study under discussion, with discretization up to 5 bits). Thus, each element of the input vector experiences a certain attenuation in accordance with the weights of the assumed filter matrix. After that, radiation from different horizontal channels is mixed in vertical waveguides and enters the detector, having previously undergone a demultiplexing procedure during parallel operation at several wavelengths. The signal from the detectors after the transimpedance amplifier enters an ADC, and the result is then transferred to the external memory of the device. As with previous architectures, the ADC–DAC accuracy and speed are the performance limiting factor in photonic accelerators. Accord-

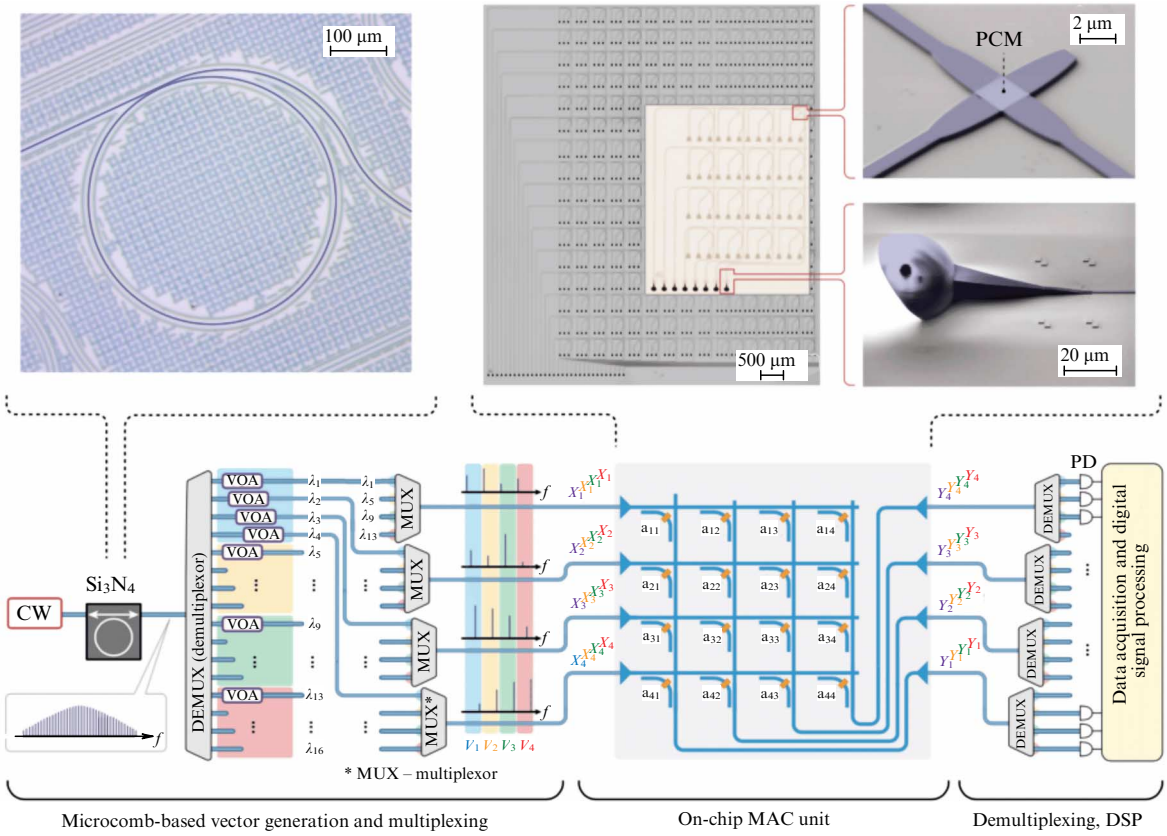

Figure 7. Schematic representation of a photonic processor based on a crossbar array. Orange areas a_{11}–a_{44} correspond to optical memory elements based on a PCM material (the actual SEM image is shown in the top-right inset), encoding the transmission in each individual channel and the corresponding weights of the kernel matrix used in the convolution. With WDM, parallel processing of several input vectors at once is possible (V_1–V_4, shown in different colors). For this, a microcomb based on a silicon nitride ring microresonator is used as the radiation source (implemented on a separate chip) [31].

ing to [31], such a device is capable of performing up to 10^{15} MAC operations per second with an energy consumption of about 20 fJ per operation (taking only optical losses into account). However, the problems listed above and additional external wiring must also be taken into account, which can significantly reduce the expected performance and increase the energy consumption of the entire system.

An extension of the presented architecture can be exemplified by an approach involving not only different wavelengths in WDM multiplexing but also different mode states in integrated optical waveguides. This idea was proposed in [33]. The key difference from the preceding case is the use of a nanostructured metasurface of a phase-change material instead of a continuous layer on the waveguide (Fig. 8), which allows not only using an additional degree of freedom but also achieving a larger number of stable levels when tuning (with a 6-bit discretization rather than a 5-bit one in the preceding study). The size of individual metasurface elements is selected so as to match the effective refractive indices for the TE_0 and TE_1 modes of a multimode silicon nitride waveguide in the crystalline state of the material. When the phase state of the material changes, the matching worsens, resulting in the preservation of the mode composition in the passing light wave. Thus, by changing the state of the metasurface through intermediate levels, it is possible to vary the radiation mode contrast. This is used to specify the photonic matrix elements instead of the usual transmission, as was implemented in the preceding study.

Combining such metasurfaces into a crossbar array allows considering it as a photonic processor for vector–matrix calculations (Fig. 9). The WDM approach is also used here, and the circuit itself works on similar principles. According to [31], such a device is capable of performing up to 164 teraoperations per second, but it is worth recalling that all the difficulties outlined above for a photonic crossbar array are relevant in this case as well. In addition, the metasurface is also sensitive to the radiation wavelength, which affects the options for scaling the system with respect to spectrum (the number of WDM channels that can actually be used in parallel without degrading the accuracy of the calculations).

2.2 Photonic Ising machine on an integrated platform

Examples of photonic accelerators for high-performance computations are not limited to MAC operations. Another relevant application could be solving combinatorial problems, such as determining the maximum cut of a graph (MAX-CUT). In particular, in [34], an integrated photonic accelerator for the Ising problem was demonstrated based on the MZI grid architecture discussed above. Particular interest in this problem is motivated by the possibility of reducing many popular combinatorial problems to the Ising problem [35]. The operating principle of a photon accelerator for the Ising problem is schematically shown in Fig. 10. At each algorithmic step, the spin state vector encoded using amplitude modulators in each waveguide channel is received

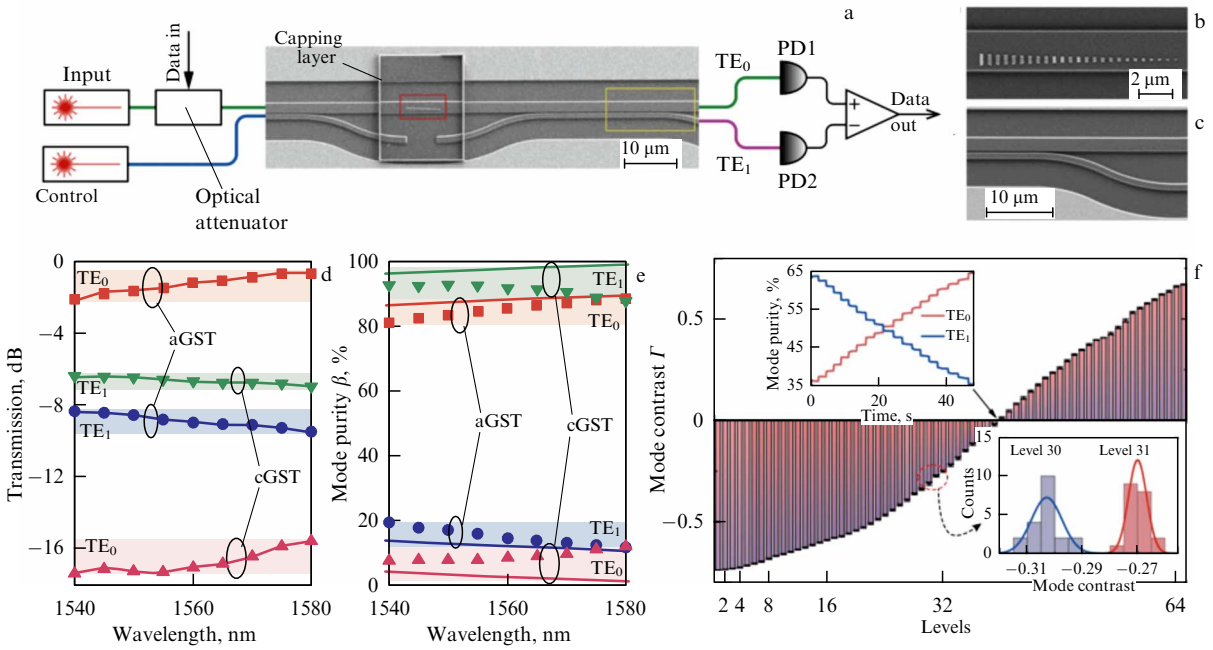

Figure 8. Integrated metasurface based on a PCM material for on-chip mode conversion. In the completely crystalline state of the PCM, transmitted radiation is effectively converted from the TE_0 to the TE_1 mode; in the completely amorphous state, almost no changes occur. Inset at the top right shows a SEM image of a metasurface on a multimode silicon nitride waveguide [33].

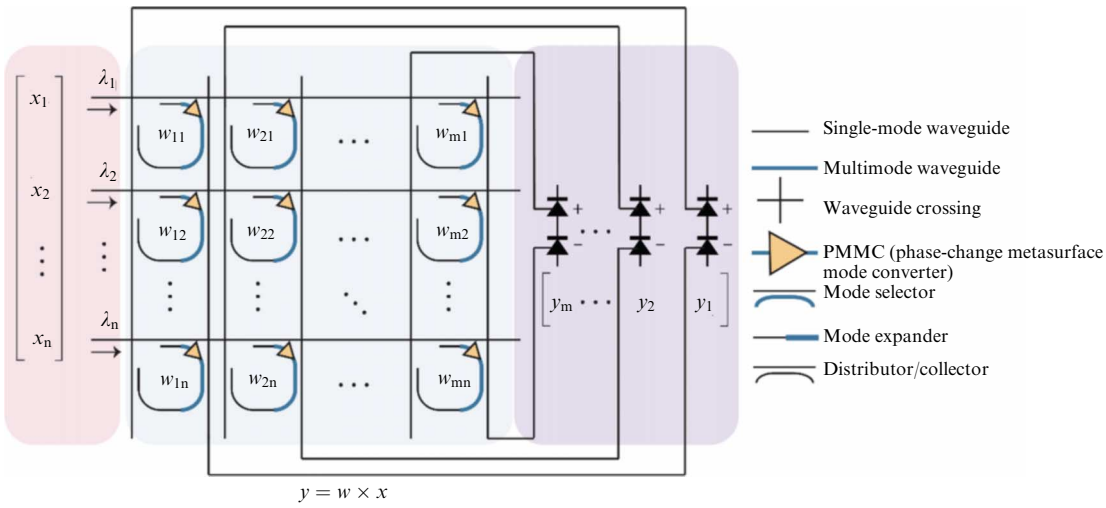

Figure 9. Crossbar array-type photonic matrix using integrated PCM metasurfaces (marked as PMMC). In contrast to a previous study, balanced photodetectors must be used here, which slightly complicates the design [33].

at the input of a photonic accelerator. The radiation corresponding to a given vector passes through a photonic matrix consisting of MZIs implementing the C_{ij} matrix related to the original spin coupling matrix K_{ij}. At the output of the matrix, the result is contaminated with Gaussian noise with a standard deviation φ, which can be implemented in both optical and electronic forms. After this, using a threshold function, a new state vector is formed, which again arrives at the input of the photonic matrix. After many iterations, regardless of the initial vector, the result converges to the Gibbs distribution for a particular spin coupling matrix. According to [34], when operating at a frequency of the order of 1 GHz, the energy per operation scales as $9/N$ pJ, where N is the number of spins in the problem, whereas a standard GPU shows a value of the order of 2.2 pJ per operation. However, it is worth noting that, as in the case of a photonic matrix for matrix multiplication, the limit value of N is bounded by optical losses, which increase sharply with increasing matrix size. Currently, there are references to an experimental implementation of 64 input channels, but further scaling may be difficult in the framework of the presented architecture and existing photonic chip elements.

3. Optical computations in free space

3.1 Diffraction neural networks

The propagation of electromagnetic radiation in space is described by the wave equation [36]. During propagation, radiation undergoes spatial changes: diffraction occurs. If the

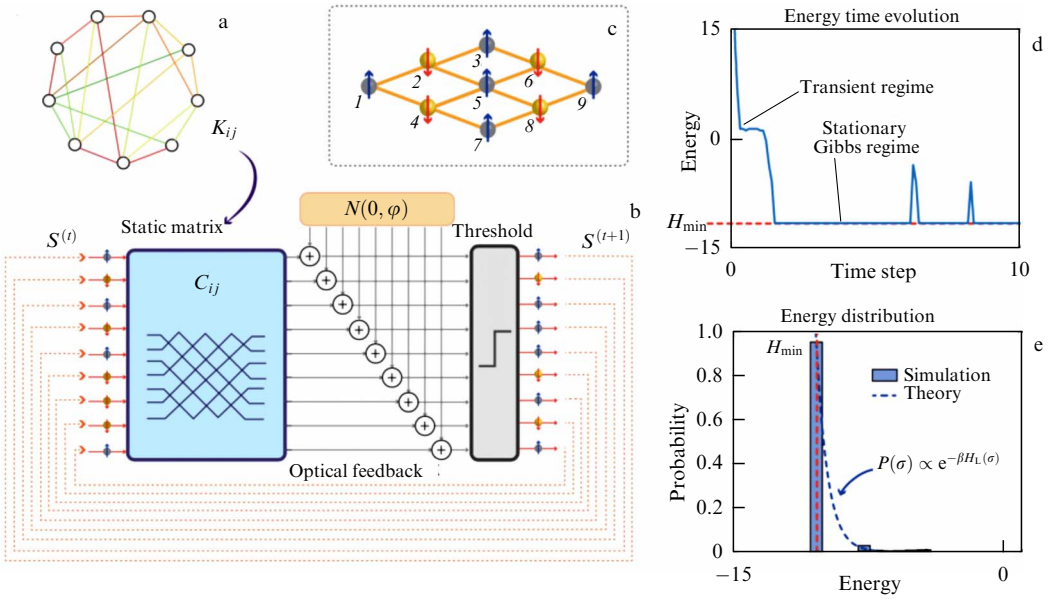

Figure 10. Photonic accelerator for the Ising problem [34].

electromagnetic field distribution $U(x, y)$ in a plane P is known, then the field distribution $U'(x', y')$ in a plane P' located at a distance z from P can be found in accordance with the Huygens–Fresnel principle as [37]

$$U'(x', y') = \frac{z}{\mathrm{i}\lambda} \iint_P U(x, y) \frac{\exp(\mathrm{i}kr)}{r^2} \, \mathrm{d}x \, \mathrm{d}y \,, \qquad (1)$$

where $r = [z^2 + (x - x')^2 + (y - y')^2]^{1/2}$. Formula (1) can be rewritten as

$$U'(x', y') = \iint_P U(x, y) \, f(x - x', y - y') \, \mathrm{d}x \, \mathrm{d}y \,,$$

where $f(x - x', y - y') = z \exp(\mathrm{i}kr)/(\mathrm{i}\lambda r^2)$. This expression is equivalent to the formula for the convolution of two functions. Thus, the propagation of the field distribution $U(x, y)$ in free space over a distance z is equivalent to applying the convolution operation with a fixed kernel.

This unique property of electromagnetic radiation allowed the development of an all-optical neural network, the diffraction neural network (DNN) [38]. Its basic idea is shown in Fig. 11. The network consists of an amplitude mask at the input (input level in Fig. 11a) and several phase (amplitude–phase) masks. The coherent radiation illuminates the amplitude mask, which defines the input field distribution. In the study under discussion, handwritten numbers from the MNIST (Modified National Institute of Standards and Technology) dataset were used as images [39]. Furthermore, according to the Huygens–Fresnel principle, each point of the amplitude mask is a point-like source of secondary waves. Their interference forms the field distribution on the first phase mask located at a certain distance. Phase masks consisted of individual pixels that introduced some phase delay at each point in space. Each pixel of the phase mask in turn acts as a point-like source of secondary waves, and this continues throughout the entire DNN. Phase

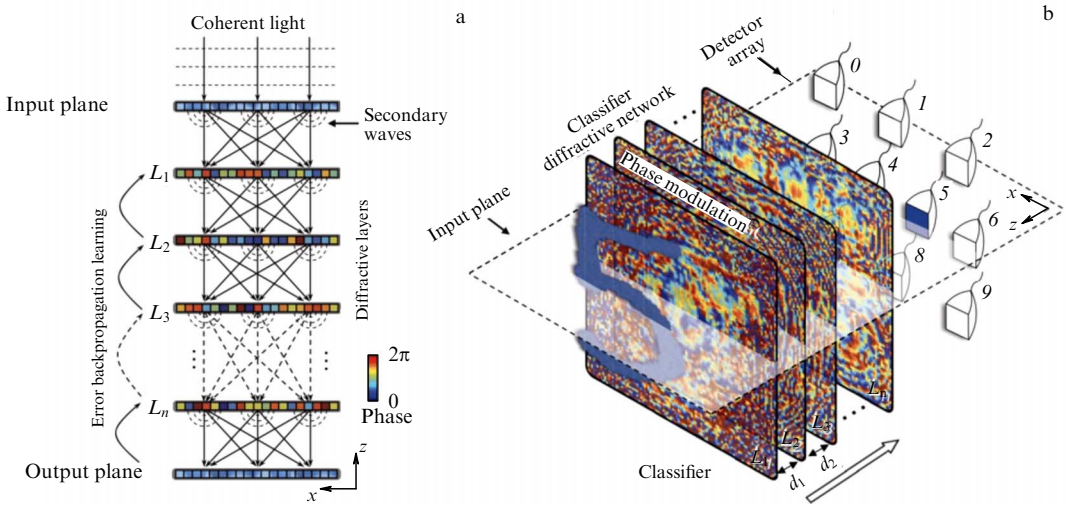

Figure 11. (a) Schematic representation of the DNN operation principle. (b) DNN operation diagram for the problem of classifying digits from the MNIST dataset [38].

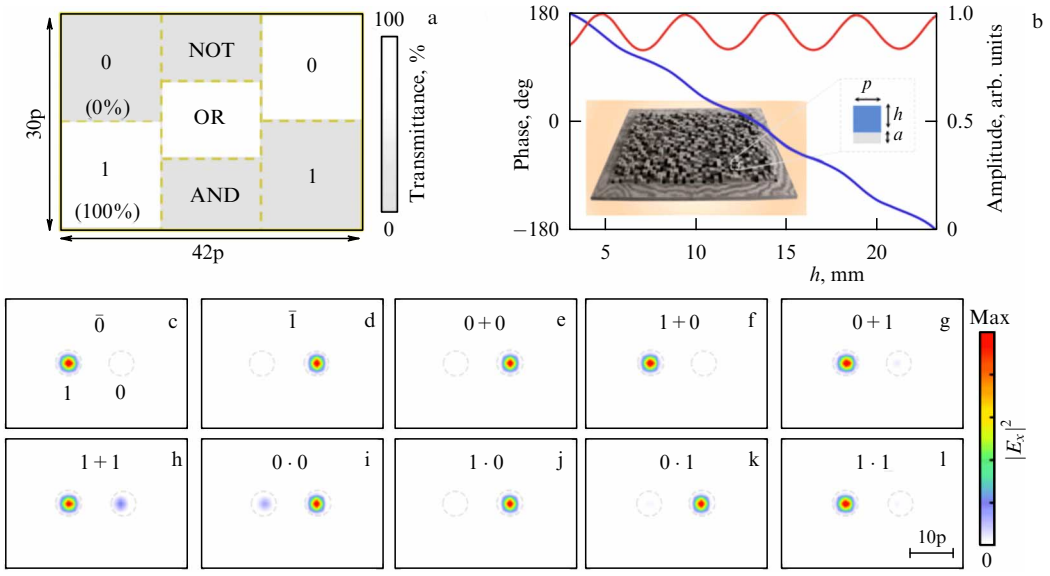

Figure 12. (a) Amplitude mask at the DNN input for encoding a logic operation. (b) Dependence of phase and amplitude of transmission of one metasurface pixel on pixel height. (c–l) Examples of field distribution at the DNN output when performing various logic operations [40].

masks allow the interference condition for secondary waves to be controlled, which ultimately allows obtaining the desired field distribution at the DNN output. In particular, the authors solved the problem of classifying handwritten digits from the MNIST dataset (Fig. 11b). For this, 10 detectors were placed at the output of the DNN, corresponding to numbers from 0 to 9. The number of the detector at which the maximum intensity was observed was used as a prediction made by the DNN.

To understand which phase masks need to be used to successfully solve the classification problem, the authors implemented and trained a DNN on a computer. They then experimentally implemented a DNN made of five phase masks for electromagnetic radiation with a frequency of 0.4 THz. Each phase mask consisted of 200 × 200 neurons, each 400 μm in size, and the distance between the masks was 3 cm. The masks were fabricated with a 3D printer, and the required phase delay was introduced by the thickness of the material. As a result, the experimentally implemented DNN shows an accuracy of 88% with the calculated accuracy given by 91.7%.

A DNN differs from standard deep neural networks by being a physical and completely optical network. In addition, it has some special features in its architecture. First, the input signals to neurons are complex-valued, determined by wave interference and the complex transmittance/reflectance of all masks. Second, the individual function of a neuron is to phase- and amplitude-modulate its input signal to generate a secondary wave, in contrast to other nonlinear functions of a neuron (a sigmoid, ReLU, or the hyperbolic tangent) used in modern deep neural networks. Third, the output signal of each neuron is delivered to the next-layer neurons (phase mask) via wave propagation and coherent (or partially coherent) interference, providing a unique form of interconnection within the network. For example, it is well known that, in modern convolutional neural networks, the receptive field is controlled by the dimension of the convolution kernel and the number of successive convolution operations. In a DNN, the perceptual field depends on the axial distance between different phase masks and the spatial and temporal coherence properties of the source of illumination. The secondary wave of each neuron is theoretically scattered at all angles, in principle affecting all neurons of the next layer. However, for a given distance between successive DNN layers, the intensity of the wave emitted by a neuron attenuates to below the detection noise level after a certain distance, which effectively defines the DNN perception field and can be physically tuned by changing it between network layers, by the intensity of the input optical radiation, or by the coherence length and diameter of the light source.

In addition to classification problems, DNNs successfully solve optical computing problems, such as implementing a controlled optical gate [40]. For this, the authors used a DNN consisting of two amplitude–phase masks. At the DNN input, the amplitude mask that specifies the necessary logic operation and operands was placed (Fig. 12a). For the demonstration, the NOT, OR, and AND operations were selected and the electromagnetic radiation frequency was fixed at 17 GHz. The role of amplitude-phase masks was played by metasurfaces — structured surfaces with the property that their parameters of reflection are determined by the collective effects of a specially created structure. The transmission of each pixel was determined by the thickness of the material (Fig. 12b). Furthermore, depending on the distribution of the field passing through the input amplitude mask, the radiation was focused in one of two areas of the output screen corresponding to the logical 0 and 1. Thus, the implementation of logic operations was essentially reduced to a binary classification of the input field distribution. The experimentally implemented DNN successfully coped with calculating the result of all logic operations, and the contrast between signals from the areas corresponding to the correct and incorrect result of the operation did not fall below 9.6 dB. The idea proposed by the authors can be extended to all basic binary logic operations, and the wavelength can be shifted to the visible or infrared ranges. The advantage of this approach is that a multifunctional logic gate is implemented not as a set of individual gates but as a single gate that can easily be controlled by the input field distribution.

DNNs are capable of processing an optical signal with a wide spectrum, i.e., in parallel for many wavelengths [41]. The authors showed that varying the loss function of a neural network allows achieving the desired functionality, such as spectrum filtering or (de)multiplexing. As an example, they took a pulse with an initial frequency of 0.25 THz and a final frequency of 1 THz, passing through a three-layer DNN. They showed that a narrow-band filter could be implemented with a fixed central frequency and the quality factor determined by the DNN loss function. To achieve this goal, the DNN was trained to focus radiation with a given frequency into an output aperture 2 mm in size. This idea can be extended to multiple frequencies that can be focused either into a single output aperture (multibandwidth filter) or into different output apertures depending on the wavelength (analogous to a spectral demultiplexor). In addition, changing the distance between the DNN layers allows shifting the central filtering frequency of the already manufactured DNN layers. This implementation of a DNN may be in demand for problems of optical information processing.

A comprehensive analysis of the DNN performance in image classification problems is presented in [42]. First, the authors show the effect of several parameters on the final accuracy of the DNN. In particular, they demonstrate that the loss function used in the DNN training process has a decisive effect on the final classification accuracy. For example, replacing the mean-square error function with cross entropy allows increasing the accuracy of a five-layer DNN from 91% to 97% in MNIST data, and from 81% to 89% in Fashion MNIST data [43]. However, such an increase in accuracy is accompanied by a drop in the output radiation intensity from the DNN, because, when using cross entropy, the network learns to transform the incoming image such that the detector corresponding to the relevant class, e.g., the number 1, receives more radiation than other detectors, but no conditions are imposed on radiation that does not reach the detectors. The authors also examined the dependence of DNN accuracy on the separation between network layers and found that, when it decreases from 40 to 4 wavelengths, the accuracy in the MNIST dataset dropped by 3%. Second, the authors show that a DNN can be connected to a standard 'digital' neural network such that the classification accuracy of the MNIST and Fashion MNIST datasets increases to 99% and 90%, which is comparable to the accuracy of modern fully digital neural networks. It is worth noting that the energy consumption of modern digital convolutional neural networks is of the order of $10^{-3}-10^{-4}$ J/image (for the ResNet network [44, 45]), while a hybrid DNN consumes about 10^{-9} J/image.

The dependence of the DNN operation quality on the number of layers was studied theoretically in [46]. The complexity of the DNN and its ability to produce arbitrary transformations of incident radiation were studied there. The matrix defining the optical transformation was shown to have the rank given by the product of the number of pixels at the input times the number of pixels at the output of the DNN. A DNN made of a finite number of layers corresponds to a lower-rank matrix in general. As the number of layers increases, the rank of the DNN matrix increases linearly with the number of layers until it reaches the maximum permissible value. Thus, implementing a DNN with the greatest learning ability, i.e., with the greatest complexity, requires increasing both the number of layers and the number of pixels (neurons) at the DNN input and output.

To implement a DNN in the visible range, metasurfaces were proposed in [47]; they are dielectric structures consisting of elements of a subwavelength size. A metasurface consisting of parallelepipeds made of TiO_2 was considered in [47]. Because the parallelepiped is asymmetric with respect to the orthogonal axes running along its sides, the polarization response is anisotropic. In addition, changing the parallelepiped dimensions allows changing the amplitude and phase of the radiation that is scattered on it. Due to these two factors, a DNN can be implemented that is capable of performing various tasks depending on the polarization of incident radiation; in other words, it is possible to implement a DNN with polarization-driven multiplexing of radiation. In particular, the authors demonstrated a DNN capable of simultaneously classifying images from MNIST and Fashion MNIST data. For this, images of each data set enter the DNN input with their own polarization, and then, depending on the image class, the radiation is focused onto an area corresponding to a given class. For example, when the number 1 is at the input, the radiation is focused in the upper-left corner of the camera, and when the number is 3, in the lower-right corner. Depending on the data set and the class, radiation should be focused onto different places in the matrix. Another innovation in [47] is the integration of a DNN with a detection system. For this, the metasurface was manufactured directly above the camera surface, at a distance of 100 μm, which facilitates scaling up the production of such DNNs.

3.2 Optical Fourier transform

It is known [37] that, in the paraxial approximation, a thin lens introduces a phase delay given by $t = \exp\left[-\mathrm{i}(k/2f)\times(x^2+y^2)\right]$, where k is the wave vector of radiation, f is the focus of the lens, and x and y are coordinates relative to the axis of the lens. If a certain field distribution $U(x,y)$ is incident on the lens, then the field distribution $U(x,y) * t$ is observed after the lens, and the field distribution in the focal plane of the lens can be found using formula (1):

$$U'(x',y') = \frac{z}{\mathrm{i}\lambda}\int\int_P U(x,y)\exp\left[-\mathrm{i}\,\frac{k(x^2+y^2)}{2f}\right] \times \frac{\exp(\mathrm{i}kr)}{r^2}\,\mathrm{d}x\,\mathrm{d}y\,.$$

In the paraxial approximation, the condition $z \gg x, y$ is satisfied, and r can be expanded in a Taylor series,

$$r = \sqrt{z^2+(x-x')^2+(y-y')^2} \approx z\left(1+\frac{(x-x')^2}{2z^2}+\frac{(y-y')^2}{2z^2}\right).$$

Then, the field distribution in the focal plane of the lens ($z = f$) can be found as

$$\begin{aligned} U'(x',y') &= \frac{\exp(\mathrm{i}kf)}{\mathrm{i}\lambda f}\int\int_P U(x,y)\exp\left[-\frac{\mathrm{i}k(x^2+y^2)}{2f}\right] \\ &\quad\times\exp\left[\frac{\mathrm{i}k\left((x-x')^2+(y-y')^2\right)}{2f}\right]\mathrm{d}x\,\mathrm{d}y \\ &= \frac{\exp(\mathrm{i}kf)\exp\left[\mathrm{i}k(x'^2+y'^2)/(2f)\right]}{\mathrm{i}\lambda f} \\ &\quad\times\int\int_P U(x,y)\exp\left[\frac{\mathrm{i}k(xx'+yy')}{f}\right]\mathrm{d}x\,\mathrm{d}y\,. \end{aligned}$$

Thus, $U'(x', y')$ is proportional to the Fourier transform of the field distribution incident on the lens.

Using lenses, a Fourier diffraction neural network (FDNN) can be constructed [48]. The operating principle of such a network is as follows. The input field distribution is incident on the lens that forms the Fourier image of this field distribution in its focal plane. Next, an optical element that modulates the radiation — a spatial light modulator or an array of digital micromirror devices — is placed there. Such modulation is equivalent to multiplying the Fourier transform of the input signal by some complex matrix. The modulating element is at the focus of a second lens, which produces the inverse Fourier transform of the modulated radiation. As is known, the Fourier transform of the product of two quantities (in this case, of the input radiation and the modulating element matrix) corresponds to the convolution operation. This makes the operating principle of FDNNs similar to that of convolutional neural networks. An array of micromirrors with a resolution of 1920×1080 pixels, updated at a frequency of 20 kHz, was used as a modulating element in [48]. To achieve a high FDNN operation speed, the input was not a single image from the data set but 16 images combined into a 4×4 matrix, with each image having a size of 208×208 pixels, making the final composite image 832×832 pixels in size. Thanks to this combination, all 16 images can undergo convolution simultaneously. The operation speed is limited only by the pixel update rate of the micromirror array (20 kHz) and the signal reading frequency of the camera (1 kHz). The authors demonstrated the functionality of the FDNN concept using the example of a single-layer network for which 16 convolution kernels were trained; for the FDNN, therefore, 16 configurations of the micromirror array were specified and were applied sequentially. The FDNN was followed by a single-layer fully connected digital neural network with a nonlinear activation function. After training, the FDNN demonstrated a classification accuracy of 98% for MNIST and 63% for CIFAR-10 [49]. The authors showed that their FDNN model is capable of calculating the convolution operation of large matrices 10 times faster than modern graphics accelerators. Computations can be done even faster using advanced developments in micromirror arrays.

A continuation of this idea is the implementation of convolution operations by optical methods and of all other operations of a neural network by digital methods [50]. The authors propose using an array of microlenses with an individual amplitude–phase mask placed at the focus of each microlens. This approach allows simultaneously calculating the convolution for several kernels. As an example, the authors tried to replace the first convolutional layer of the well-known convolutional neural network AlexNet [51] with an optically implemented layer. The remaining layers of the neural network were implemented on a computer. It was found that, when replacing the digital convolutional layer with an optical one, the classification accuracy for Kaggle's Cats and Dogs dataset [52] decreases from 96% to 87%. This is most probably caused by the lack of a nonlinear activation function for the optical convolution layer and the small size of the training data set. With the more famous MNIST dataset, the accuracy of a neural network with an optical convolutional layer differs from the accuracy of a fully digital neural network by less than 0.5%. The main motivation for replacing the digital convolutional layer with an optical one is to speed up the neural network. The time T_{latency} of performing the convolution operation with an optical circuit is a sum, $T_{\text{latency}} = T_{\text{source}} + T_{\text{prop}} + T_{\text{detect}} + T_{\text{data}}$, where T_{source} is the time required to generate the input image, T_{prop} is the time it takes the light to pass through the optical circuit, T_{detect} is the time required to detect a signal, and T_{data} is the time required to transmit the detected signal to the software that implements the rest of the neural network. The time T_{source} is determined by the speed of the image generation system and was 1 ms (a 1-kHz update rate) in the study under discussion. The time T_{prop} is a few picoseconds and can be ignored. The time T_{detect} is determined by the speed of the detecting element and is about 1 ms for a CCD camera. The time T_{data} is determined by the time of signal transmission from the detecting element to the computer. For a USB 3.0 connection with a bandwidth of 2500 Mbps^{-1} and a 100 KB image, T_{data} is 0.32 ms. Thus, the total time to calculate the result of the convolution operation is 2.32 ms. It is important to note that this time is not sensitive to changes in the image resolution. As a result, a graph (Fig. 13) can be obtained where the operation speed of convolutional layers implemented entirely digitally is compared with that using optical tools. The result produced by a single convolution layer implemented using an optical circuit is calculated faster than with a graphics accelerator when the image is larger than 500×500 pixels in size.

The authors also evaluated the power consumption of an optically implemented convolution layer and a convolution layer placed on a graphics accelerator. In the optical implementation, power consumption is independent of the size of the convolution kernel, while the power consumption of the convolution layer on a graphics accelerator increases in proportion to the number of pixels. We can conclude that the optical implementation of even one convolution layer is called for only when working with high-resolution images and when using high-dimensional convolution kernels.

A major obstruction to using DNNs is the lack of a simple way to implement an optical nonlinear activation function. In [53], a thin plate of photorefractive material SBN:60 installed at the end of a multilayer FDNN was proposed for this purpose (Fig. 14). The main motivation for the use of the SBN:60 material is its large nonlinear response. The refractive index of that material depends on the intensity of the incident

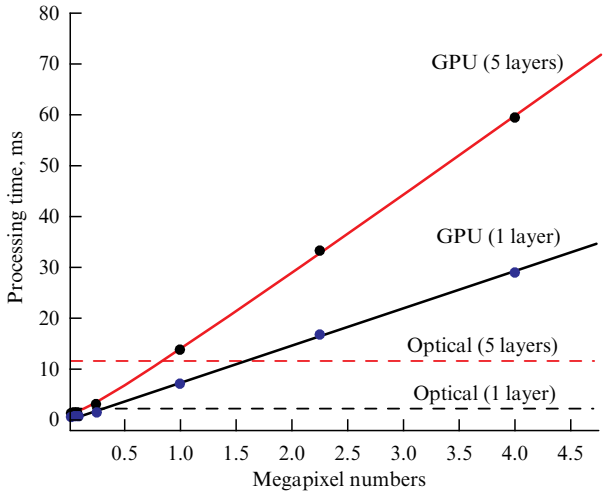

Figure 13. Dependence of the operation speed of convolution layers placed on a graphics accelerator (GPU) and implemented using an optical circuit (Optical) [50].

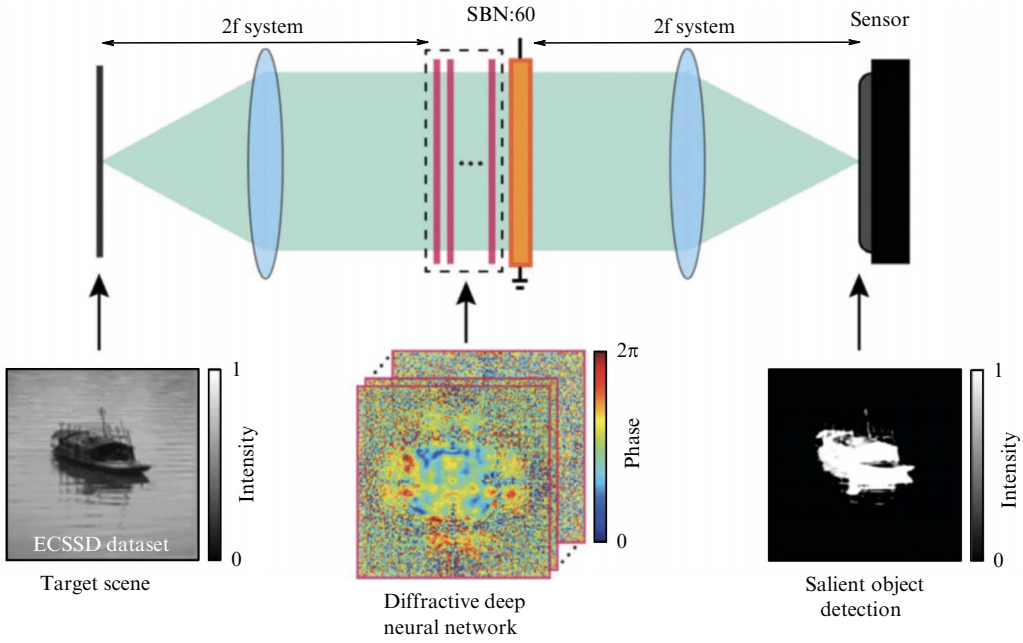

Figure 14. Scheme of an FDNN with a plate of photorefractive material SBN:60 [53].

radiation as

$$n = \kappa E_{\text{app}} \frac{\langle I \rangle}{1 + \langle I \rangle} ,$$

where n is the change in the refractive index; κ is a constant determined by the refractive index of the material, the electro-optical coefficient, and the intensity $\langle I_0 \rangle$ of uniform illumination of the material; E_{app} is the applied static electric field; and $\langle I \rangle$ is the intensity perturbation relative to uniform illumination $\langle I_0 \rangle$. It was shown that changing n allows changing the phase of transmitted radiation from 0 to π for a plate 1 mm in thickness with a static voltage of 972 V applied to it and an incident radiation intensity of 0.1 mW mm^{-2}. The introduction of a nonlinear layer allowed the authors to implement an FDNN capable of solving problems of segmentation and detection of objects in images (Fig. 14). Such an FDNN can operate in real time and detect objects directly on video.

4. Conclusions

We have examined advanced approaches to implementing analogue photonic computing and photonic intelligent systems.

The first group of approaches is based on the use of elements of integrated photonics to implement vector–matrix operations, which allows introducing photonic neural networks into traditional computing systems as a coprocessor to speed up calculations and increase energy efficiency. This approach has already been put into practical use and has a significant energy efficiency advantage, up to 20 fJ per operation, which is two orders of magnitude higher than the current energy efficiency of standard central processors. We have also discussed the challenges associated with this approach: its efficiency, the maximum frequency of optoelectronic conversions, and the bit depth of numbers used in computation.

The second group of approaches is based on optical calculations in free space, which allows performing mathematical operations in one computation cycle with the entire data array by using the physical properties of optical radiation. These approaches have not yet found their practical application in computing, partly due to the complexity of their integration with current computing algorithms, but they certainly deserve attention and have significant potential for working with large data sets.

The advantages of photonic computers over electronic ones include a two-orders-of-magnitude gain in energy efficiency due to the possibility of parallel operation at several wavelengths at once, using physically the same array, with the same number of weight elements in the photonic and electronic crossbar arrays, increased throughput thanks to signal modulation at a much higher frequency, and reduced requirements for heat removal from the photonic chip. The advantage of electronic components over optical ones is their compactness, i.e., the potential to place more elements and compensate for the difference in performance. However, the problem of resistance drift remains open, which reduces the accuracy of hardware accelerator computations.

Optical computing for certain tasks opens up the prospect of not only increasing energy efficiency but also accelerating operation by the use of low-bit high-speed DAC–ADCs at low power consumption and reducing the number of electro-optical conversions when working with optical convolutional neural networks or with signals that are initially represented in the optical domain.

This research was carried out in the framework of the scientific program of the National Center for Physics and Mathematics (project National Center for Studying Supercomputer Architectures) and with the support of the Intellect Nonprofit Foundation for the Development of Science and Education.

References

1. Xu R et al. *Opt. Laser Technol.* **136** 106787 (2021)
2. Marković D et al. *Nat. Rev. Phys.* **2** 499 (2020)

3. Xu M et al. *Adv. Funct. Mater.* **30** 2003419 (2020)
4. Zhang J et al. *Adv. Intell. Syst.* **2** 1900136 (2020)
5. Sunny F P et al. *ACM J. Emerg. Technol. Comput. Syst.* **17** (4) 61 (2021)
6. Yao K, Unni R, Zheng Y *Nanophotonics* **8** 339 (2019)
7. Ferreira de Lima Th et al. *Nanophotonics* **6** 577 (2017)
8. Goodman J W, Dias A R, Woody L M *Opt. Lett.* **2** 1 (1978)
9. Zuo Y et al. *Optica* **6** 1132 (2019)
10. Liu J et al. *PhotoniX* **2** 5 (2021)
11. Shastri B J et al. *Nat. Photon.* **15** 102 (2021)
12. Silicon Photonics. From Technologies to Markets. Market and Technology Report 2021. Yole Group, https://s3.i-micronews.com/uploads/2021/05/YINTR21175-Silicon-Photonics-2021-Sample.pdf
13. Silicon Photonics 2022. Market and Technology Trends. Yole Group, https://www.yolegroup.com/product/report/silicon-photonics-2022/
14. GlobalFoundries Announces Next Generation in Silicon Photonics Solutions and Collaborates with Industry Leaders to Advance a New Era of More in the Data Center. March 7, 2022. GlobalFoundries Press Releases, https://gf.com/gf-press-release/globalfoundries-announces-next-generation-silicon-photonics-solutions-and/
15. Huang C et al. *Adv. Phys. X* **7** 1981155 (2022)
16. Kirtas M et al. ''Early detection of DDoS attacks using photonic neural networks'', in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop* (*IVMSP*) (Piscataway, NJ: IEEE, 2022) https://doi.org/10.1109/IVMSP54334.2022.9816178
17. Clements W R et al. *Optica* **3** 1460 (2016)
18. Al-Qadasi M A et al. *APL Photon.* **7** 020902 (2022)
19. Shen Y et al. *Nat. Photon.* **11** 441 (2017)
20. Bandyopadhyay S, Hamerly R, Englund D *Optica* **8** 1247 (2021)
21. Envise. Lightmatter, https://lightmatter.co/products/envise/
22. ''High-efficiency multi-slot waveguide nano-opto-electromechanical phase modulator'', Grant US-11281068-B2, https://app.dimensions.ai/details/patent/US-10884313-B2
23. Feng Y et al. *Opt. Express* **28** 38206 (2020)
24. Baghdadi R et al. *Opt. Express* **29** 19113 (2021)
25. Jacob B et al., in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, June 18–22, 2018* (Piscataway, NJ: IEEE, 2018) p. 2704
26. Machupalli R, Hossain M, Mandal M *Microprocess. Microsyst.* **89** 104441 (2022)
27. Lightelligence, https://www.lightelligence.ai
28. Lightelligence. PACE: Photonic Arithmetic Computing Engine, https://www.lightelligence.ai/index.php/product/index/2.html
29. Qu Y et al. *Sci. Bull.* **65** 1177 (2020)
30. Chao C-Y, Fung W, Guo L J *IEEE J. Sel. Top. Quantum Electron.* **12** 134 (2006)
31. Feldmann J et al. *Nature* **589** 52 (2021)
32. Burr G W et al. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **6** 146 (2016)
33. Wu C et al. *Nat. Commun.* **12** 96 (2021)
34. Prabhu M et al. *Optica* **7** 551 (2020)
35. Lucas A *Front. Phys.* **2** 5 (2014) https://doi.org/10.3389/fphy.2014.00005
36. Akhmanov S A, Nikitin S Yu *Fizicheskaya Optika* (Physical Optics) 2nd ed. (Moscow: Izd. Mosk. Univ., 2004)
37. Goodman J W *Introduction to Fourier Optics* (Englewood, CO: Roberts and Co., 2005)
38. Lin X et al. *Science* **361** 1004 (2018)
39. Yann LeCun's Home Page, http://yann.lecun.com/exdb/mnist/
40. Qian C et al. *Light Sci. Appl.* **9** 59 (2020)
41. Luo Y et al. *Light Sci. Appl.* **8** 112 (2019)
42. Mengu D et al. *IEEE J. Sel. Top. Quantum Electron.* **26** 3700114 (2020) https://doi.org/10.1109/JSTQE.2019.2921376
43. Papers with Code. Fashion-MNIST, https://paperswithcode.com/dataset/fashion-mnist
44. Papers with Code. Residual Network, https://paperswithcode.com/method/resnet
45. He K et al. ''Deep residual learning for image recognition'', in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, 27–30 June 2016* (Piscataway, NJ: IEEE, 2016) p. 770, https://doi.org/10.1109/CVPR.2016.90
46. Kulce O et al. *Light Sci. Appl.* **10** 25 (2021)
47. Luo X et al. *Light Sci. Appl.* **11** 158 (2022)
48. Miscuglio M et al. *Optica* **7** 1812 (2020)
49. The CIFAR-10 dataset, https://www.cs.toronto.edu/~kriz/cifar.html
50. Colburn S et al. *Appl. Opt.* **58** 3179 (2019)
51. Krizhevsky A, Sutskever I, Hinton G E, in *Advances in Neural Information Processing Systems* Vol. 25 (Eds F Pereira et al.) (Red Hook, NY: Curran Associates Inc., 2012) p. 1097
52. Kaggle. Datasets. Cats-vs-Dogs: image dataset for binary classification, https://www.kaggle.com/datasets/shaunthesheep/microsoft-catsvsdogs-dataset
53. Yan T et al. *Phys. Rev. Lett.* **123** 023901 (2019)