applied sciences

MDPI

Review

Key Technologies of Photonic Artificial Intelligence Chip Structure and Algorithm

Li Pei [1,*], Zeya Xi [1], Bing Bai [1,2], Jianshuai Wang [1], Xiaoyan Zuo [1], Tigang Ning [1], Jingjing Zheng [1] and Jing Li [1]

[1] Key Laboratory of All-Optical Networks and Modern Communication Networks of Ministry of Education, Institute of Lightwave Technology, Beijing Jiaotong University, Beijing 100044, China; 19120149@bjtu.edu.cn (Z.X.); baibing@photoncounts.com (B.B.); wangjsh@bjtu.edu.cn (J.W.); 19111051@bjtu.edu.cn (X.Z.); tgning@bjtu.edu.cn (T.N.); jjzheng@bjtu.edu.cn (J.Z.); lijing@bjtu.edu.cn (J.L.)

[2] Photoncounts (Beijing) Technology Co. Ltd., Beijing 100081, China

* Correspondence: lipei@bjtu.edu.cn; Tel.: +86-010-51683625

Abstract: Artificial intelligence chips (AICs) are the intersection of integrated circuits and artificial intelligence (AI), involving structure design, algorithm analysis, chip fabrication and application scenarios. Due to their excellent ability in data processing, AICs show a long-term industrial prospect in big data services, cloud centers, etc. However, with the conceivable exhaustion of Moore's Law, the size of traditional electronic AICs (EAICs) is gradually approaching the limit, and an architectural update is highly required. Photonic artificial intelligence chips (PAIC) utilize light beam propagation in the silicon waveguide, contributing to a high parallelism configuration, fast calculation speed and low latency. Due to light manipulation, PAICs perform well in anti-electromagnetic interference and energy conservation. This invited paper summarized the recent research on PAICs. The characteristics of different hardware structures are discussed. The current widely used training algorithm is given and the Photonic Design Automatic (PDA) simulation platform is introduced. In addition, the authors' related work on PAICs is presented and we believe that PAICs may play a critical role in the deployment of data processing technology.

Keywords: photonic artificial intelligence chip; neural network; training algorithm

Citation: Pei, L.; Xi, Z.; Bai, B.; Wang, J.; Zuo, X.; Ning, T.; Zheng, J.; Li, J. Key Technologies of Photonic Artificial Intelligence Chip Structure and Algorithm. *Appl. Sci.* **2021**, *11*, 5719. https://doi.org/10.3390/app11125719

Academic Editors: Yufei Ma and Yong Zhao

Received: 26 May 2021
Accepted: 15 June 2021
Published: 20 June 2021

1. Introduction

The capacity of computing systems is in an arms race with the massively growing amount of data. The AIC is considered to be an effective way to embrace the data explosion era and provides excellent ability in data computing. However, the traditional EAIC is based on electronic computing, which is gradually entering the bottleneck period with the upcoming limit of Moore's law [1,2]. Integration size and large power consumption bring big challenges in calculation ability and data processing capacity for EAIC. Unlike electrical interconnects in EAICs, the PAIC is constructed by optical waveguide interconnects, forming optical neural networks (ONN) or optical elemental devices to realize optical linear or nonlinear computing. It offers great potential for orders of magnitude improvement in energy conservation and computing capacity due to its natural parallel processing, being less susceptible to interference, free superposition, etc. [3]. In addition, with the growth of nanophotonics technology, the PAIC has been applied in various practical scenarios, assisted by the prospect of an algorithm and material [4–6].

PAICs include the hardware's structural design and the matched optimization algorithm. The hardware's structure refers to accomplishing the computing by changing the phase and intensity of the light signal through optical devices. The matching algorithm refers to data training and module control. In PAIC design, the compatibility of the structure and algorithm is also significant. Before fabrication, in order to obtain a workable PAIC, a system-level simulation platform is necessary to simulate the whole performance.

The hardware structure of PAIC is also called ONN. The ONN in PAIC is constructed with optical waveguides or devices, such as Mach–Zehnder interferometers (MZI), semiconductor optical amplifiers (SOAs), micro-ring resonators (MRRs) or some 3D waveguides. Based on the hardware ONN structure or specialized devices, PAIC can easily implement linear or nonlinear operation [7]. Assisted by the weight training algorithm, PAIC highly benefits big data services and high-capacity communication networking.

Specifically, the main advantages of PAICs are as follows: (1) PAICs utilizing optical signal propagation show a high clock frequency and, according to the waveguide interconnects, the data can be labeled from one dimension to multi-dimension [3]. These contribute to fast data processing and low latency [8]. (2) The data propagation and calculation take place simultaneously. Thus, it reduces the data's repeated handling and eliminates the "memory wall" issue. (3) The flexible multiple waveguide channel design represents a truly parallel implementation of PAICs, expanding the data capacity and also accelerating the processing speed. (4) The calculation in PAICs is achieved by a passive optical waveguide. There is no need for an extra energy supply and PAICs are cost-effective in energy consumption [6].

In this invited paper, firstly, three linear ONNs are listed and we made a comparison concerning their advantages and disadvantages. Next, two nonlinear operation hardware structures are presented. Then, the current widely used training algorithm (gradient descent algorithm) is introduced. After that, the simulation platform of the PAIC system is described. Finally, the research work of our group in the related field is presented. An ONN structure is designed and an algorithm is optimized to form PAICs. It can complete artificial intelligence computing, such as image recognition and classification. The test results show that it runs well.

2. Hardware Structures for Linear Operation

ONN is formed by passive optical waveguide interconnects and has natural advantages for the linear operation. Due to the different computing tasks, the ONN is mainly classified into three types: feedforward neural network (FNN), convolutional neural network (CNN), and recurrent neural network (RNN). The networks can be implemented by optical waveguides. Among them, the response function of a single MZI can be expressed as a Jones matrix [9] and thus, the MZI array is an excellent candidate for linear matrix multiplication [10–12]. In the case of SOA, modulation and amplification are used to control the intensity or phase of the input signal and complete the matrix multiplication [13,14]. MRR can also be operated according to different wavelengths. In MRR, optical signals of different wavelengths are modulated, and the calculation is completed [15–17]. Furthermore, complex 3D-routed waveguides are created by two photon polymerization [18,19], which efficiently connects many IO channels. When the channels are added by a combiner, the 3D waveguide works as a summator. Above all, these waveguides are deployed for ONN, intending to realize linear operation. For the three types of ONN, the scheme is shown in Table 1.

Table 1. The scheme of main neural network.

Scheme / Network	MZI Array	SOA	MRR	3D Waveguide
PFNN	√	√	√	√
PCNN	√		√	
PRNN	√	√	√	

2.1. Photonic Feedforward Neural Network (PFNN)

The feedforward neural network (FNN) is the simplest one-way neural network, including an input layer, hidden layer and output layer. The signal transmits unidirectionally from the input layer to the output layer. The essence of realizing the matrix multiplication

function is to complete the multiplication and accumulation (MAC). PFNN can be realized by the MZI array, SOA, MRR and 3D waveguide.

A. MZI array.

In the case of PFNN, the MZI is cascaded without reflection. The light stays in forward propagation. In PAICs, linear computing is implemented in the MZI array as shown in Figure 1 [20]. The input optical signal with multiple channels propagates in the paralleled MZI waveguide. The phase changes when passing through the MZI array. In this way, the output matrix can be obtained by completing the linear operation. In 2017, Shen proposed and demonstrated the first photonic interference computing unit chip based on the MZI array [21]. The whole network uses an array of 56 MZIs and 213 phase shifters to complete the matrix operation through abundant phase changes.

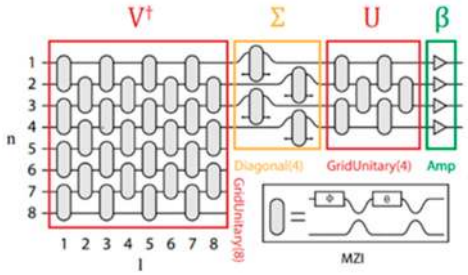


Figure 1. MZI array [20]. A universal 8 × 4 optical linear multiplier with two unitary multipliers marked red. It consists of MZIs in a grid-like layout and a diagonal layer marked yellow.

B. Semiconductor optical amplifier (SOA).

In the basic idea of SOA, the MAC operation is realized by a single amplifier. Once the amplifier is cascaded, it enables a large-scale MAC operation. Generally, weighting is expressed by the attenuation or gain. At the output, the light of the different channels with multiple weighting is summed through the wavelength division multiplexer. Thus, SOA can complete the N × M matrix multiplication. The FNN based on SOA is illustrated in Figure 2. By adjusting the drive current of the SOA, the gain coefficient changes. It can be considered that the variation of the gain coefficient corresponds to the change in the transmission matrix coefficient. Then, through the arrayed waveguide grating (AWG) or multimode interference optical coupler (MMI), the weighted results of multiple SOAs are summed. In this way, the weighted summation of each row in the matrix is obtained. This realizes the matrix's multiplication operation [22].

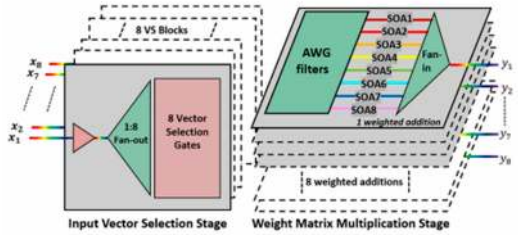


Figure 2. SOA wavelength division multiplexer [22]. The chip integrates 8 weighted addition circuits, which are used for 8 WDM input vectors and provide 8 WDM outputs. SOA is used to assist the weight sum of each layer. Highlight the block in use in gray.

C. the micro-ring resonator (MRR).

The implementation scheme based on MRR is similar to that of SOA. MRR works at different wavelength when the MRR's length or the radius are adjusted. The MRR is featured in wavelength selection. The light after different MRRs forms different channels with forward propagation. Then, through the arrayed waveguide grating (AWG) or multimode interference optical coupler (MMI), the light is summated with multiple channels.

A matrix operation is completed. The transmittance of MRR can be adjusted by thermal adjustment [23,24], ESC [25], phase adjustable materials, etc.

The PFNN based on MRR is illustrated in Figure 3 [26]. The input signals are ported into different MRRs, and then, each channel carries a signal with a specific wavelength. After nonlinear modulation, the output matrix is obtained. In this structure, photonic neurons' output signals are fixed to certain wavelengths. With the WDM signal porting in the network, each connection between a pair of neurons is independently configured by an MRR weight, and each channel has a signal monitor. After that, the operation is completed.

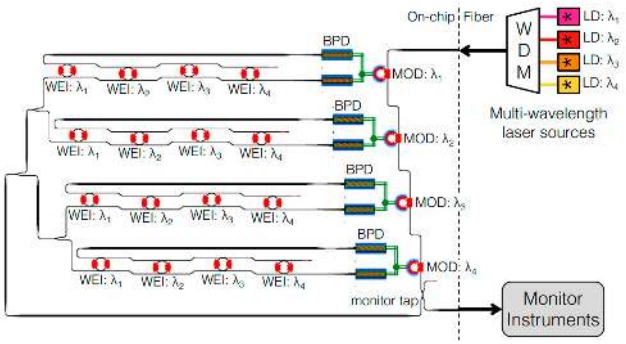


Figure 3. MRR wavelength division multiplexer [26]. MRR is mainly responsible for the weight configuration of neural network, and the red part indicates that the weight can be changed through the external environment. The whole network is integrated except for bit pumped lasers.

D. 3D waveguide.

In addition, Yu et al. [27] and Moughames et al. [28] construct a feedforward neural network through a 3D waveguide written directly by a laser, as shown in Figure 4. At present, there are only two layers of network, so the signal is a one-way transmission. The multiple IO channels are finally combined into one output port. Three-dimensional waveguide achieves the goal of dimension expansion, but the signal is still a one-way transmission. The input signal is summed by the N × 1 beam combiner. This completes the interconnection between different layers and the waveguides works as a summator.

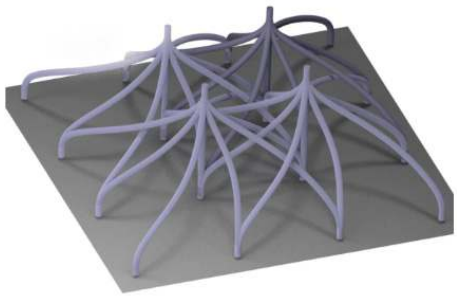


Figure 4. 3D waveguide feedforward neural network [28]. A small network hosting simple coupler. Chirality of the connections avoids the intersection of individual waveguides between the input and output ports.

2.2. Photonic Convolutional Neural Network (PCNN)

The main task of the convolution neural network is to complete convolution operations. The schemes in convolutional neural networks are MZI array [29] and MRR array [30–33].

A. MZI array.

The basic principle is similar to the corresponding photonic feedforward neural network. The essence of convolution is matrix operation. The MZI array is used to implement matrix multiplication operations. The signal after matrix decomposition is input into the MZI array by segments. The front and back operations are implemented by

using cascaded MZI. This completes a convolution operation. In 2018, Bagherian utilized chips to extend the original simple fully connected neural network to high-dimensional image recognition. It uses time division multiplexing to complete convolution computing. The main steps are as follows: matrix multiplication based on the MZI array, image convolution by time division multiplexing, and construction of the convolutional neural network layers (five convolutional layers and one fully connected layer). This realizes the recognition of colored numbers 0–9 [29].

B. MRR

In the CNN [34], MRR completes matrix operation by the weight of the micro-ring. The multiplexed wavelengths enter the MRR array. The amplitude of each wavelength is multiplied by its corresponding micro-ring weight, and then, output. The multiplication is realized by adjusting the resonant peaks of the micro-rings to their respective laser wavelengths [31]. Figure 5 shows the convolution operation with input feature mapping. This greatly facilitates the calculation.

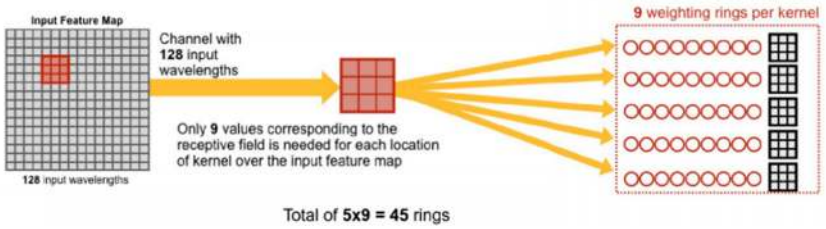


Figure 5. MRR structure is for an input feature map of size 16 × 16. After filtering the map with input feature mapping, the convolution is completed [33].

2.3. Photonic Recurrent Neural Network (PRNN)

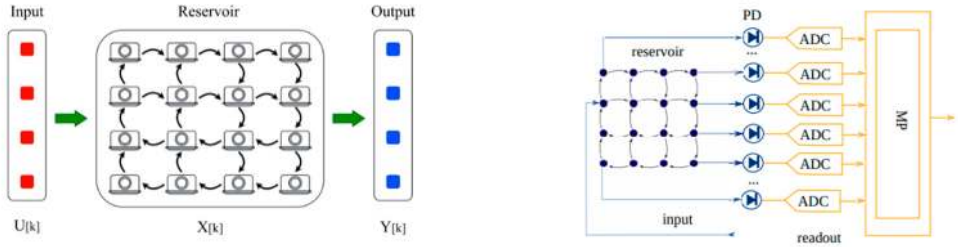
The recurrent neural network is also known as the reservoir calculation. The information transmits forward and backward to form a loop structure. The reservoir is mainly composed of an input layer, middle layer and output layer. Only the output weights are trained in the RNN. For photonic storage pool network, there are two structures: one is the parallel scheme, and the other is the serial scheme, as shown in Figure 6 Different from the previous two neural networks, the reservoir network is mainly used for dimension transformation of data.

A. Parallel scheme

The nodes in the parallel structure's reservoir can be implemented by SOA [35], silicon-based micro-ring resonators [36], silicon-based waveguide delay lines [37], etc. Optical reservoirs based on SOA and MRR, respectively, utilize the gain saturation of SOA and the nonlinear effects of MRR (free carrier dispersion and thermo-optical effects). No matter what kind of device, the simplest way to realize interconnection is to return the output signal to the input node. In this way, feedback can be achieved within the network. When there are multiple inputs, parallel operation is realized by inputting different signals.

B. Serial scheme

The nodes of the serial structure's reservoir operation can be implemented using modulators, SOA, etc. In recent years, the delay of serial loop RNN based on MRR or MMI has made great progress. At the same time, this type of photonic RNN also tries to use multi-stage or more complex time division multiplexing to further accelerate computing speed.



(**a**) (**b**)

Figure 6. Optical reservoir plan (**a**) MRR structure [36]. The data matrix is input into the reservoir network composed of MRR, and after network circulation, the data matrix is output. (**b**) Silicon-based delay line [37]. The data are input into a network composed of delay lines, which make the front and back input signals interact with each other. PD is used to detect and convert optical signal into electrical signal, which is processed by a micro-processor.

2.4. Comparison of three Linear ONNs

The above is the analysis of the structural characteristics of the three neural networks.

From the analysis, it can be seen that PFNN is the simplest of the three networks. It is the easiest structure to implement. However, because of its simple structure, the computation of the matrix is limited.

The PCNN can complete convolution operation on account of its complex hidden layers, which is the core of convolution. However, it also causes the complexity of the structure. Furthermore, because the hidden layers participate in the operation, the operation time will be longer than PFNN.

PFNN and PCNN mostly use MZI and MRR to realize matrix operation. For an MZI, its system is relatively simple and has stronger versatility. However, due to excessive loss during cascading, it is not suitable for large-scale integration. For MRR, its size is relatively small and large-scale integration is easier. However, it is more sensitive to temperature. There are great challenges in achieving precise control.

Different from PFNN and PCNN, PRNN has the structural characteristics of a recurrent network. Its application is to enrich or compress data dimensions. It can also be used to deal with tasks related to time series. Its structure uses a reservoir. For the reservoir, the internal principle utilizes random projection to transform the dimensionality of the data. Therefore, there is no need for complicated control, and it has strong fault tolerance for integrated technology. However, due to this special principle, it is difficult to apply to most occasions.

3. Hardware Structures for Nonlinear Operation

The linear operation of a neural network can solve relatively simple problems. The nonlinear activation function is the root of the artificial neural network's powerful expression ability. This affects the speed of network convergence and the accuracy of recognition. As shown in Figure 7, the position of nonlinearity in the system is behind the linear neural network.

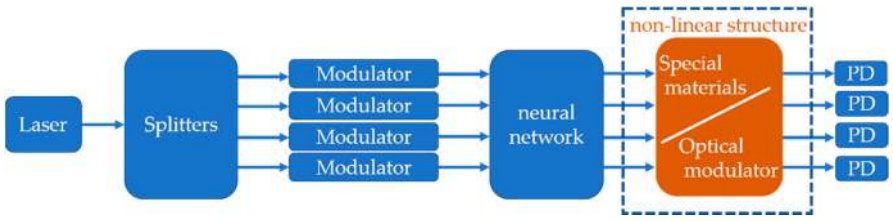


Figure 7. Ideal PAIC system implementation diagram. The dotted line is nonlinear structure.

There are a lack of practical photonic devices that express nonlinear functions. Currently, photon calculation is an optoelectronic hybrid computing architecture. The nonlinear

calculation part is all performed in the electrical domain. In this way, multiple photoelectric and electro-optical conversions are involved in the network. This not only limits the speed of the photonic neural network, but also brings additional energy consumption. Ideally, some optical nonlinear devices can realize the nonlinear calculation, and they have the features of low threshold, reconfigurability, easy integration, and fast response. The current ideas are divided into two types: special materials (saturated absorbers and two-dimensional graphene materials) and the combined structure of the optical modulator (MZI, MRR and SOA).

3.1. Nonlinear Operation Based on Special Material

The main special materials used to realize nonlinear operation are a saturated absorber and two-dimensional graphene materials. Nonlinearity can be realized by placing the special material in the dotted line of Figure 7.

The saturated absorber mainly uses its transmission characteristics. The absorption coefficient has a reverse relationship with the intensity of the incident light. The transmission coefficient increases with the growth of the optical power. Thus, the Relu function can be realized [21]. Two-dimensional graphene materials work analogous to the saturated absorber. However, compared with the saturated absorber, graphene has the advantages of low threshold and that it is easy to be excited in nonlinear effects.

3.2. Nonlinear Operation Based on Optical Modulator

Another way to realize nonlinear structure is to use the optical modulator. The optical modulators that can realize nonlinearity are placed in the dotted line of Figure 7. The basic principle is that the light signal of the weighted sum is converted into a voltage by the photodetector, and then, applied to the optical modulator. This affects the transmission spectrum. The realization of the optical nonlinear function is by changing the transmittance of the optical signal through the modulator. The structures of optical modulators can be implemented by MZI modulators [38], electro-absorption modulators [39], and micro-ring resonator modulators [40], as shown in Figure 8. Different nonlinear functions can be realized by changing the bias voltage of the modulator.

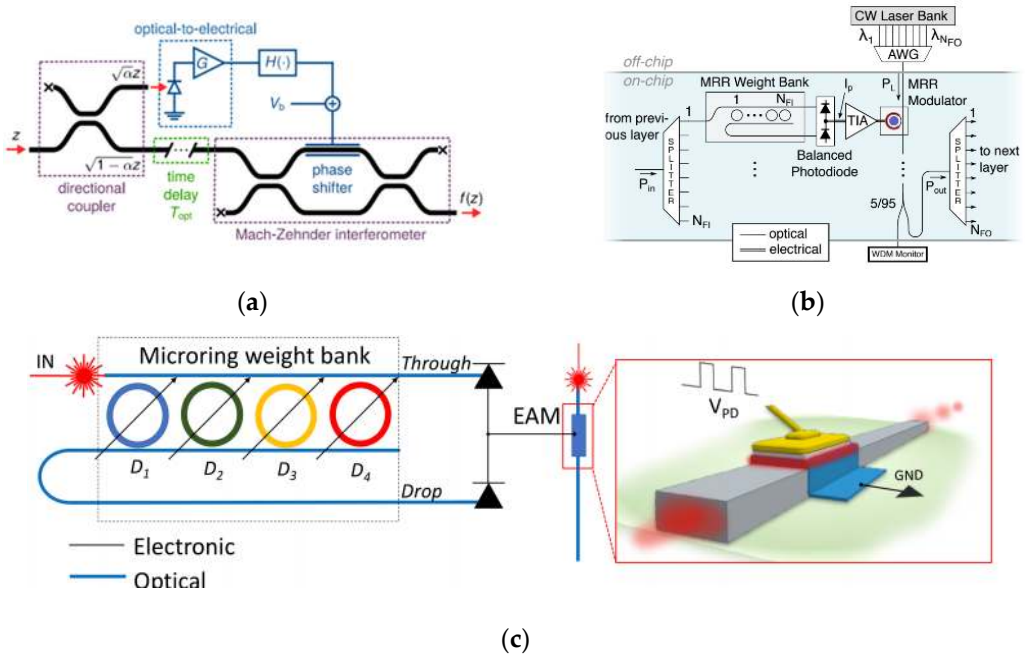


Figure 8. Optical modulator implementation diagram. (**a**) MZI modulator [38], (**b**) electro-absorption modulator [39], (**c**) micro-ring cavity modulator [40].

In 2019, Alexandris et al. realized the sigmoid nonlinear function based on the serial structure of MZI composed of two SOAs and a single SOA [41]. The schematic diagram is shown in Figure 9. This structure is mainly based on the cross-phase modulation effect and cross-gain modulation effect of SOA. When the input pulse width is very small, within the integration window, the arrangement of the input light pulse signal and the number of pulses will affect the pulse width of the output signal. SOA has a high nonlinear coefficient, and at the same time, has a gain effect on optical signals.

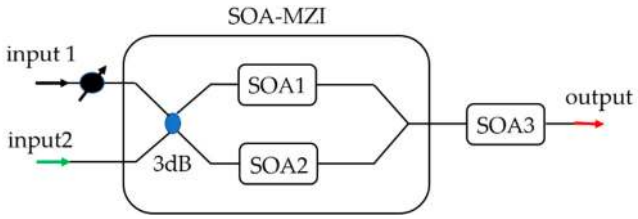


Figure 9. SOA-based implementation of the nonlinear function. SOA is placed in the two arms of MZI, which mainly performs phase modulation. The input signal and the reference signal go through the 3dB coupler, then are input into the two SOA for phase modulation. Cross-modulation is completed in SOA3.

3.3. Comparison of Two Nonlinear Types

From the above analysis, we can see that the two types of nonlinear structures have their own advantages and disadvantages.

The nonlinear structure based on the special materials is easier to be integrated into the chip. However, it is difficult to obtain, and it also has higher requirements for the integrated environment.

The nonlinear structure based on the optical modulator is easier to be adjusted and obtain nonlinearity. However, the structure is larger and it is harder to integrate.

4. Training Algorithm

Currently, in terms of algorithm training, the simulation model of the photonic computing network is trained on the electronic computer. Then, the trained model parameters are loaded onto the photonic chip. However, even if it is trained in the electrical domain, its effect is still restricted by two aspects: the accuracy of the simulation model's description and the computing speed. The training algorithm includes forward propagation [42], finite difference calculation gradient (MIT), in situ back propagation [43] and gradient measurement (Stanford). These attempts have essentially been completed at the level of computer simulation, and have not been used for training on actual physical chips.

The training algorithm problem of the photonic network is a restrictive factor for expanding the application of the PAIC. Photons cannot be stored like electrons. We cannot directly record the state of photons. Therefore, backward propagation algorithms that are widely used in electrical neural network training are difficult to directly transplant to photonic artificial intelligence network training. In order to solve this problem, Hughes et al. proposed an on-chip training algorithm in 2018 [43]. By recording the light field distribution and the phase distribution of the phase shifter, we can obtain the gradient value that decreases toward the convergence direction. Then, we calculate the phase configuration of the chip phase shifter in the next iteration to gradually converge to a better result. Hughes et al. trained a specific two-optical interference unit (OIU) neural network on the chip through simulation. It implements exclusive OR logic to verify the effectiveness of the algorithm.

Zhang et al. proposed an effective training algorithm based on neuro-evolution strategy in 2019 [44]. It uses genetic algorithms (GA) and particle swarm optimization (PSO) to train the hyperparameters in ONNs and optimize the connection weights. The trained ONNs are used to complete the classification task for performance evaluation. The calculation results show that its accuracy and stability are sufficient to compete with

traditional learning algorithms. The system also uses the photonic artificial intelligence network to realize the classification of the modulation format of the communication signal. In the future, this algorithm can be further expanded and transplanted to larger-scale PAICs. It can gradually obtain the best configuration of the chip through on-chip training to complete specific functions. Figure 10 shows the results of autonomous learning using a gradient descent algorithm. It can be seen that with iterative learning, signal recovery is improving [42].

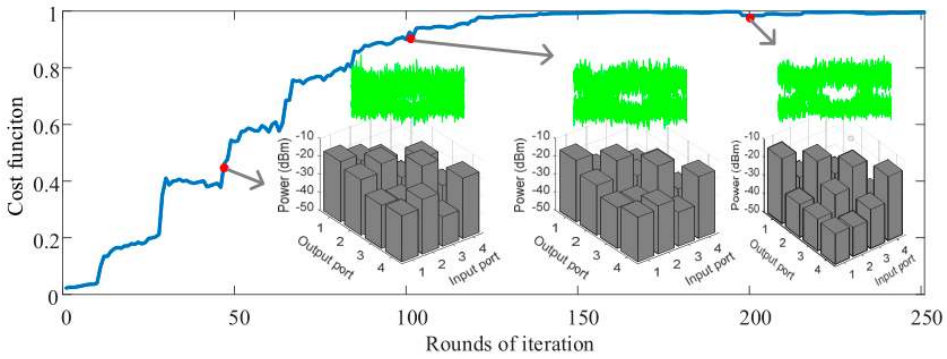


Figure 10. Results of algorithm training [42]. Three different training times are selected with 50 times,100 times and 200 times.

5. Software Simulation Platform

Efficient algorithms and a variety of network models are important cornerstones to support the continued development of photonic chips. At present, photonic computing networks are mainly simulated and trained through electronic computers. Then, the trained model parameters are loaded onto the photonic chip. By far, there are two main software simulation platforms in use: one is IPKISS [45], and the other is INTERCONNECT. In addition to the above two commonly used simulation tools, the Institute of Microelectronics of the Chinese Academy of Sciences has designed a system-level simulation and verification tool, named PDA.

PDA designers use Python to package various models of optical devices. Users can modify the parameters and interface of the devices, and connect the device by a simple function statement. It can realize the extremely complex network structure at the link level, or the simulation task of the framework. After the completion of the system, each optical path can be monitored and the simulation diagram in the time domain can be output. This greatly facilitates the observation and analysis of the experimental results. Because of the flexibility of Python, PDA can also interact with MATLAB to complete the simulation according to the operators' different needs. PDA also has the advantage that it can be used in conjunction with the open-source layout tool Klayout to realize the complete chip design process of the layout driver and schematic driver. This provides great convenience for the overall development and design of the chip. It is worth mentioning that it is easy to operate.

6. Our Work in Lab

Although the PAIC has great potential, there is still no mature system technology for it. In order to seek a new breakthrough, we have performed a considerable amount of research work on the system-level function of PAIC.

First, we have designed and optimized the aspect of hardware structure. It is composed of MZI and a reservoir to complete the calculation and transform the dimension of data. This is the important part for the preprocessing of input data. According to the design structure, an integrated optical operation module is built, as shown in Figure 11a.

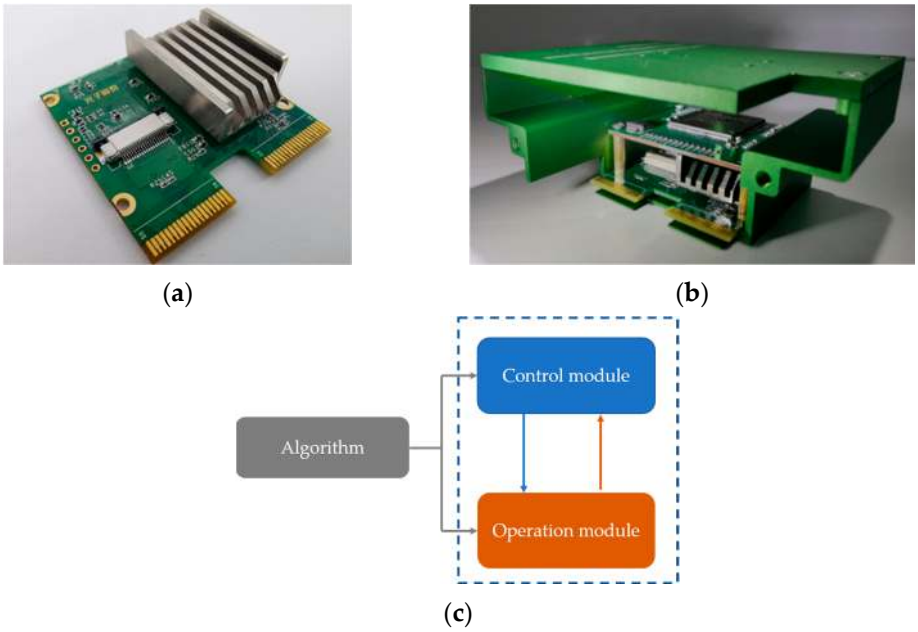


Figure 11. (**a**) The optical operation module. (**b**) The integrated module. It is a two-layer module with a separate control and calculation. The top is the control module, and the bottom is the optical operation module. (**c**) The work principle.

Second, we have written a training algorithm to realize the cooperation of software and hardware. In the aspect of the algorithm, a gradient descent algorithm is designed to train the output data. This training algorithm functions on both the control module and the operation module, which is shown in Figure 11b. The work's principle is shown in Figure 11c.

Lastly, based on the comprehensive analysis of devices, circuits and algorithms, a PAIC is formed, as shown in Figure 12a. It is mainly composed of a control module and optical operation module, and the data preprocessing of the reservoir network is included.

Figure 12. (**a**) Self-developed PAIC. (**b**) Application of this system to complete the data monitoring of the construction site.

The self-developed PAIC is applied to intelligent computing. The results show that it can complete image segmentation, image recognition and other functions effectively. The processed image is shown in Figure 12b. It mainly monitors whether workers are wearing safety helmets. In a specific example test, the marked 6057 image data are divided into five groups and input into the algorithm model. The average test accuracy is 89.7%, and the energy efficiency ratio of algorithm deployment is 1.23 Tops/W.

7. Conclusions

In conclusion, PAICs have great advantages due to their power consumption and small size. They have aroused considerable attention from scholars. This paper mainly

summarizes the structural design, algorithm matching of PAICs and the software platforms that can be used in large-scale simulation. Additionally, we put forward the related work of our laboratory, an integrated information processing system (Self-developed PAIC), which provides one useful solution for the in-depth research of PAICs. They contain much significance and unlimited possibilities. We believe that with the help of PAICs, the optical computer is in sight.

Author Contributions: Conceptualization, L.P. and B.B.; data curation, J.W. and X.Z.; writing—original draft preparation, Z.X.; writing—review and editing, L.P., T.N., J.Z. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities, grant number 2019JBM343.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jizhi, X.; Xinyan, Z.; Jianwei, L. Application of Artificial Intelligence in the Field of Power Systems. In Proceedings of the The 4th International Conference on electrical engineering, mechanical engineering and automation, February 2019, Xian, China, 29–31 May 2019.
2. Sui, X.; Wu, Q.; Liu, J.; Chen, Q.; Gu, G. A Review of Optical Neural Networks. *IEEE Access* **2020**, *8*, 70773–70783. [CrossRef]
3. Shen, Y.; Meng, X.; Cheng, Q.; Rumley, S.; Abrams, N.; Gazman, A.; Manzhosov, E.; Glick, M.S.; Bergman, K. Silicon Photonics for Extreme Scale Systems. *J. Light. Technol.* **2019**, *37*, 245–259. [CrossRef]
4. Yao, K.; Unni, R.; Zheng, Y. Intelligent nanophotonics: Merging photonics and artificial intelligence at the nanoscale. *Nanophotonics* **2019**, *8*, 339–366. [CrossRef]
5. Piccinotti, D.; Macdonald, K.F.; A Gregory, S.; Youngs, I.; Zheludev, I.N. Artificial intelligence for photonics and photonic materials. *Rep. Prog. Phys.* **2021**, *84*, 012401. [CrossRef] [PubMed]
6. Shastri, B.J.; Tait, A.N.; de Lima, T.F.; Pernice, W.H.P.; Bhaskaran, H.; Wright, C.D.; Prucnal, P.R. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **2021**, *15*, 102–114. [CrossRef]
7. Mosca, E.P.; Griffin, R.D.; Pursel, F.P.; Lee, J.N. Acoustooptical matrix-vector product processor: Implementation issues. *Appl. Optics* **1989**, *21*, 3843–3851. [CrossRef] [PubMed]
8. Mirza, A.; Avari, S.M.; Taheri, E.; Pasricha, S.; Nikdast, M. Opportunities for Cross-Layer Design in High-Performance Computing Systems with Integrated Silicon Photonic Networks. In Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020, Grenoble, France, 9–13 March 2020.
9. Cartwright, S. New optical matrix-vector multiplier. *Appl. Opt.* **1984**, *23*, 1683–1684. [CrossRef]
10. Tait, A.N.; De Lima, T.F.; Zhou, E.; Wu, A.X.; Nahmias, M.A.; Shastri, B.J.; Prucnal, P.R. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **2017**, *7*, 1–10. [CrossRef]
11. Reck, M.; Zeilinger, A.; Bernstein, H.J.; Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **1994**, *73*, 58–61. [CrossRef]
12. Pai, S.; Bartlett, B.; Solgaard, O.; Miller, D.A.B. Matrix Optimization on Universal Unitary Photonic Devices. *Phys. Rev. Appl.* **2019**, *11*, 064044. [CrossRef]
13. Liboiron-Ladouceur, O.; Bergman, K.; Boroditsky, M.; Brodsky, M. Polarization-Dependent Gain in SOA-Based Optical Multistage Interconnection Networks. *J. Light. Technol.* **2006**, *24*, 3959–3967. [CrossRef]
14. Stubkjaer, K. Semiconductor optical amplifier-based all-optical gates for high-speed optical processing. *IEEE J. Sel. Top. Quantum Electron.* **2000**, *6*, 1428–1435. [CrossRef]

15. Tan, X.; Yang, M.; Zhang, L.; Jiang, Y.; Yang, J. A Generic Optical Router Design for Photonic Network-on-Chips. *J. Lightwave Technol.* **2012**, *30*, 368–376. [CrossRef]
16. Zhu, A.; Chen, D.; Xu, C.; Cong, H.; Zhi, L. Research of MRR fault detection in photonic network-on-chip. *J. Electron. Meas. Instrum.* **2017**, *31*, 1200–1205.
17. ZhiHua, Y.; Qi, Z.; Xin, J.; Juan, Z.; Hadi, B.; Selviah, D.R. Microring resonator-based optical router for photonic networks-on-chip. *Quantum Electron.* **2016**, *46*, 655–660. [CrossRef]
18. Deubel, M.; Von Freymann, G.; Wegener, M.; Pereira, S.; Busch, K.; Soukoulis, C.M. Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. *Nat. Mater.* **2004**, *3*, 444–447. [CrossRef]
19. Yang, L.; Münchinger, A.; Kadic, M.; Hahn, V.; Mayer, F.; Blasco, E.; Barner-Kowollik, C.; Wegener, M. On the Schwarzschild Effect in 3D Two-Photon Laser Lithography. *Adv. Opt. Mater.* **2019**, *7*, 1901040. [CrossRef]
20. Fang, M.Y.-S.; Manipatruni, S.; Wierzynski, C.; Khosrowshahi, A.; Deweese, M.R. Design of optical neural networks with component imprecisions. *Opt. Express* **2019**, *27*, 14009–14029. [CrossRef]
21. Shen, Y.; Harris, N.C.; Skirlo, S.; Prabhu, M.; Baehr-Jones, T.; Hochberg, M.; Sun, X.; Zhao, S.; LaRochelle, H.; Englund, D.; et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **2017**, *11*, 441–446. [CrossRef]
22. Shi, B.; Calabretta, N.; Stabile, R. Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect. *IEEE J. Sel. Top. Quantum Electron.* **2019**, *26*, 1–11. [CrossRef]
23. Tait, A.N.; De Lima, T.F.; Nahmias, M.A.; Shastri, B.J.; Prucnal, P.R. Multi-channel control for microring weight banks. *Opt. Express* **2016**, *24*, 8895–8906. [CrossRef]
24. Ma, P.Y.; Tait, A.N.; De Lima, T.F.; Huang, C.; Shastri, B.; Prucnal, P.R. Photonic independent component analysis using an on-chip microring weight bank. *Opt. Express* **2020**, *28*, 1827–1844. [CrossRef]
25. Tait, A.N.; Nam, S.; Mirin, R.P.; Shastri, B.; Prucnal, P.R.; De Lima, T.F.; Shainline, J.M.; Buckley, S.M.; McCaughan, A.N.; Nahmias, M.A.; et al. Neuromorphic Silicon Photonics on Foundry and Cryogenic Platforms. In Proceedings of the 2019 IEEE Photonics Society Summer Topical Meeting Series (SUM), Beijing, China, 8–10 July 2019; pp. 1–2.
26. Tait, A.N.; De Lima, T.F.; Nahmias, M.A.; Miller, H.B.; Peng, H.-T.; Shastri, B.; Prucnal, P.R. Silicon Photonic Modulator Neuron. *Phys. Rev. Appl.* **2019**, *11*, 064043. [CrossRef]
27. Yu, H.; Zhang, Q.; Gu, M. Three-dimensional direct laser writing of biomimetic neuron structures. *Opt. Express* **2018**, *26*, 32111–32117. [CrossRef]
28. Moughames, J.; Porte, X.; Thiel, M.; Ulliac, G.; Larger, L.; Jacquot, M.; Kadic, M.; Brunner, D. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks. *Optica* **2020**, *7*, 640. [CrossRef]
29. Cohen, E.; Malka, D.; Shemer, A.; Shahmoon, A.; Zalevsky, Z.; London, M. Neural networks within multi-core optic fibers. *Sci. Rep.* **2016**, *6*, 29080. [CrossRef] [PubMed]
30. Bagherian, H.; Skirlo, S.; Shen, Y.; Meng, H.; Soljacic, M. On-Chip Optical Convolutional Neural Networks. *arXiv* **2018**, arXiv:1808.03303.
31. Xu, S.; Wang, J.; Zou, W. High-energy-efficiency integrated photonic convolutional neural networks. *arXiv* **2019**, arXiv:1910.12635.
32. Bangari, V.; Marquez, B.A.; Miller, H.; Tait, A.N.; Nahmias, M.A.; De Lima, T.F.; Peng, H.-T.; Prucnal, P.R.; Shastri, B.J. Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs). *IEEE J. Sel. Top. Quantum Electron.* **2019**, *26*, 1–13. [CrossRef]
33. Mehrabian, A.; Al-Kabani, Y.; Sorger, V.J.; El-Ghazawi, T. PCNNA: A Photonic Convolutional Neural Network Accelerator. In Proceedings of the 2018 31st IEEE International System-on-Chip Conference (SOCC), Arlington, VA, USA, 4–7 September 2018.
34. Liu, W.; Liu, W.; Ye, Y.; Lou, Q.; Xie, Y.; Jiang, L. HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers. In Proceedings of the 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, 25–29 March 2019; pp. 1483–1488.
35. Vandoorne, K.; Dambre, J.; Verstraeten, D.; Schrauwen, B.; Bienstman, P. Parallel Reservoir Computing Using Optical Amplifiers. *IEEE Trans. Neural Netw.* **2011**, *22*, 1469–1481. [CrossRef]
36. Coarer, F.D.-L.; Sciamanna, M.; Katumba, A.; Freiberger, M.; Dambre, J.; Bienstman, P.; Rontani, D. All-Optical Reservoir Computing on a Photonic Chip Using Silicon-Based Ring Resonators. *IEEE J. Sel. Top. Quantum Electron.* **2018**, *24*, 1–8. [CrossRef]
37. Freiberger, M.; Katumba, A.; Bienstman, P.; Dambre, J. Training Passive Photonic Reservoirs with Integrated Optical Readout. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 1943–1953. [CrossRef] [PubMed]
38. Williamson, I.A.D.; Hughes, T.W.; Minkov, M.; Bartlett, B.; Pai, S.; Fan, S. Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks. *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 1–12. [CrossRef]
39. De Lima, T.F.; Tait, A.N.; Saeidi, H.; Nahmias, M.A.; Peng, H.-T.; Abbaslou, S.; Shastri, B.J.; Prucnal, P.R. Noise Analysis of Photonic Modulator Neurons. *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 1–9. [CrossRef]
40. Amin, R.; George, J.K.; Sun, S.; De Lima, T.F.; Tait, A.N.; Khurgin, J.B.; Miscuglio, M.; Shastri, B.; Prucnal, P.R.; El-Ghazawi, T.; et al. ITO-based electro-absorption modulator for photonic neural activation function. *APL Mater.* **2019**, *7*, 081112. [CrossRef]
41. Shi, B.; Calabretta, N.; Stabile, R. SOA-Based Photonic Integrated Deep Neural Networks for Image Classification. In Proceedings of the 2019 Conference on Lasers and Electro-Optics (CLEO), Munich, Germany, 23–27 June 2019; OSA Technical Digest (Optical Society of America, 2019). pp. 1–2. [CrossRef]
42. Zhou, H.; Zhao, Y.; Wang, X.; Gao, D.; Dong, J.; Zhang, X. Self-Configuring and Reconfigurable Silicon Photonic Signal Processor. *ACS Photonics* **2020**, *7*, 792–799. [CrossRef]

43. Hughes, T.W.; Minkov, M.; Williamson, I.A.D.; Shi, Y.; Fan, S. Training of Photonic Neural Networks through In Situ Backpropagation. In Proceedings of the 2019 Conference on Lasers and Electro-Optics (CLEO), Munich, Germany, 5–10 May 2019; OSA Technical Digest (Optical Society of America, 2019). pp. 1–2.
44. Zhang, T.; Wang, J.; Dan, Y.; Lanqiu, Y.; Dai, J.; Han, X.; Sun, X.; Xu, K. Efficient training and design of photonic neural network through neuroevolution. *Opt. Express* **2019**, *27*, 37150–37163. [CrossRef] [PubMed]
45. Bogaerts, W.; Dumon, P.; Lambert, E.; Fiers, M.; Pathak, S.; Ribeiro, A. IPKISS: A parametric design and simulation framework for silicon photonics. In Proceedings of the the 9th International Conference on Group IV Photonics (GFP), San Diego, CA, USA, 29–31 August 2012; pp. 30–32.