CHINESE OPTICS LETTERS

Vol. 21, No. 9 | September 2023

Integrated diffractive optical neural network with space-time interleaving

Tingzhao Fu (符庭钊)†, Yuyao Huang (黄禹尧)†, Run Sun (孙 润), Honghao Huang (黄泓皓), Wencan Liu (刘文灿), Sigang Yang (杨四刚), and Hongwei Chen (陈宏伟)*

Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

*Corresponding author: chenhw@tsinghua.edu.cn

Received February 12, 2023 | Accepted June 2, 2023 | Posted Online August 22, 2023

Integrated diffractive optical neural networks (DONNs) have significant potential for complex machine learning tasks with high speed and ultralow energy consumption. However, the on-chip implementation of a high-performance optical neural network is limited by input dimensions. In contrast to existing photonic neural networks, a space-time interleaving technology based on arrayed waveguides is designed to realize an on-chip DONN with high-speed, high-dimensional, and all-optical input signal modulation. To demonstrate the performance of the on-chip DONN with high-speed space-time interleaving modulation, an on-chip DONN with a designed footprint of 0.0945 mm^2 is proposed to resolve the vowel recognition task, reaching a computation speed of about 1.4×10^{13} operations per second and yielding an accuracy of 98.3% in numerical calculation. In addition, the function of the specially designed arrayed waveguides for realizing parallel signal inputs using space-time conversion has been verified experimentally. This method can realize the on-chip DONN with higher input dimension and lower energy consumption.

Keywords: integrated diffractive optical neural networks; machine learning; arrayed waveguides.
DOI: 10.3788/COL202321.091301

1. Introduction

Artificial neural networks (ANNs) have received significant attention in many fields, including computer vision[1], natural language processing[2], decision-making[3,4], etc. Faced with complex tasks, the requirements of ANNs for computing power are more stringent, causing a heavy computation burden in existing electronic computing hardware[5–13] [e.g., the central processing unit (CPU), the graphical processing unit (GPU), the field-programmable gate array (FPGA), and the application-specific integrated circuit (ASIC)]. Therefore, an alternative novel computing architecture is imperative for performing heavy computation. Presently, optical neural networks (ONNs) have garnered tremendous interest[14–34] because of the advantages of their low power consumption, low latency, and ultrahigh bandwidth, which may solve the dilemma that the existing electronic computing architectures face.

The ONNs based on diffractive surfaces in free space can optically perform inference tasks[15,35,36], this kind of ONNs is not limited by high input dimensions. However, the ONNs based on spatial diffraction are often composed of discrete devices, such as 3D-printed metasurfaces, digital micromirror devices (DMDs), and spatial light modulators (SLMs), which are bulky and low in integration. Moreover, the unavoidable calibration process between the discrete devices may bring potential errors to the system. Consequently, these problems may limit the application scenarios of such kinds of ONNs to some extent. The ONNs based on the silicon-on-insulator (SOI) platform not only have the advantages of low power consumption, low latency, and ultrahigh computing bandwidth but also have the characteristics of small volume, light weight, good stability/portability, and unnecessary physical alignment process. There are several different implementations of the integrated ONNs, including a coherent approach based on Mach–Zehnder interferometer (MZI) mesh[14,20,21,27], wavelength-division multiplexing (WDM) processing with micro-ring resonators (MRRs), programmable routing enabled by a phase-change material (PCM)[16], and an on-chip diffractive approach based on sub-wavelength structures[17,18,29]. Among them, the ONNs designed based on MZIs, WDM-MRRs, and PCM have low integration, which makes achieving large-scale expansion difficult. The on-chip DONNs based on sub-wavelength structures can achieve high integration and large computational capacity. However, its massively parallel inputs are limited by the energy consumption and high speed. Therefore, it is significantly urgent to solve the limited input dimensions of the on-chip DONNs.

The calculation function of the DONNs is based on the mutual interference between the parallel input signals. Therefore, it must be ensured that the modulated signals enter the input section of the DONNs at the same time. Routinely, the parallel input signals can be realized by integrated phase shifters (i.e., the integrated heaters) on each waveguide[14,30,31,37]. However, it is inevitably necessary to continuously provide additional energy to maintain the normal operation of these shifters by this way. Thus, the on-chip power consumption would increase dramatically with the increase of the input dimensions. In this work, the parallel input multiple signals are realized by designing the true-delay lines of the arrayed waveguides. The interval length between the adjacent waveguides is determined by the time interval between the serial signals. The shorter the time interval, the smaller the length difference between two adjacent arrayed waveguides. Here, an on-chip DONN for the task of vowel recognition was theoretically verified, and the arrayed waveguides with different lengths were fabricated using electron beam lithography (EBL) technology based on an SOI platform. According to the modulation rate (10 Gbps), the interval difference between adjacent waveguides was set to 7.1429 mm corresponding to a 100 ps time delay. We fabricated the arrayed waveguides with a fixed length difference and experimentally demonstrated its performance. Based on the results of the fabricated arrayed waveguides, an on-chip DONN with two hidden layers (each with 70 neurons) is proposed, and the numerical calculation result of the blind test prediction of the vowel recognition dataset[38] is 98.3%. The aforementioned method for designing the on-chip DONN based on the standard complementary metal-oxide semiconductor (CMOS) process provides a solution for high dimensional inputs, low power consumption, and large capacity computing, which paves a way for promoting the applications of the on-chip DONN in various aspects.

2. Concept and Principle

2.1. Arrayed waveguides

The design of the arrayed waveguides is key to realizing the high dimensional parallel loading of signals using the space-time interleaving method. In Fig. 1, the length of ΔL is determined by the modulation rate of the input signal, which can be calculated by

$$\Delta L = \frac{c}{B_m \cdot n_g}, \tag{1}$$

where ΔL is the length interval between the adjacent arrayed waveguides, B_m is the modulation rate of the input signal, n_g is the group refractive index of the waveguide, and c is the speed at which light travels in the vacuum.

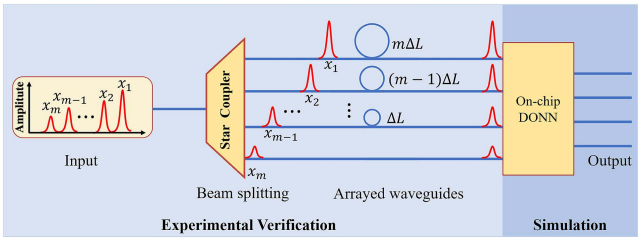

Fig. 1. Schematic diagram of converting a serial input signal into a parallel signal at a certain time using the space-time conversion method and feeding it into the on-chip DONN.

2.2. On-chip electromagnetic propagation model

In this work, an on-chip electromagnetic propagation model (OCEPM) is proposed by modifying the Huygens–Fresnel principle under restricted propagation conditions, which is shown as follows:

$$w^i_{n,m} = \frac{1}{j\lambda} \cdot \left(\frac{1+\cos\theta_{n,m}}{2r_{n,m}}\right) \cdot \exp\left(j\frac{2\pi r_{n,m} n_S}{\lambda}\right) \cdot A_p \exp(j\Delta\phi), \tag{2}$$

where i represents the ith layer of the network, n represents the nth neuron located at (x_n, y_n) of layer i, and m represents the mth neuron located at (x_m, y_m) of layer $i-1$. λ is the working wavelength, and $j = \sqrt{-1}$ is an imaginary unit. $\cos\theta_{n,m} = (x_n - x_m)/r_{n,m}$, $r_{n,m} = \sqrt{(x_n - x_m)^2 + (y_n - y_m)^2}$ is the distance between the mth neuron in layer $i-1$ and the nth neuron in layer i, and n_S is the effective refractive index (ERI) of the slab waveguide. A_p is a specific coefficient of the amplitude and $\Delta\phi$ is a fixed phase delay[18]. To verify the effectiveness of the OCEPM, the same input signal, with an amplitude distribution as shown in Fig. 2(a) and a phase distribution as shown in Fig. 2(b), is fed into the OCEPM and the 2.5D variational finite-difference time-domain (2.5D FDTD) solver for calculation. The calculation results are shown in Figs. 2(c) and 2(d). The calculated amplitude and phase distributions are consistent with each other.

3. Methods

3.1. Device fabrication

The arrayed waveguides were fabricated on an SOI (100 substrate) platform with a 220-nm-thick silicon (Si) top layer and a 3-μm-thick buried oxide layer. The fiber-grating coupler loss was optimized to 5 dB per input/output facet.

3.2. Optical measurements

A continuous-wave tunable semiconductor laser with a polarization controller was used to launch light onto the chip (32 mW). The output was monitored using two dual-channel optical power meters, and the minimum power detection limit was −75 dB. An external auxiliary circuit was provided by a direct current (DC) dual-tracking voltage-stabilizing source (DH1718E-5, 0–35 V).

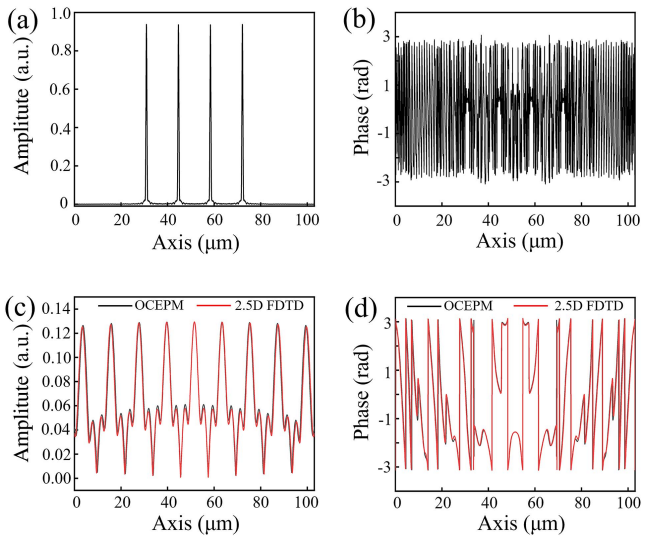

Fig. 2. (a) and (b) are the field intensity and phase distribution of the input signal, respectively. (c) and (d) are the field intensity and phase distribution of the input signal propagating 300 μm later in a slab waveguide (thickness is 220 nm, width is 105 μm) of the modified OCEPM (black line) and the 2.5D FDTD (red line), respectively.

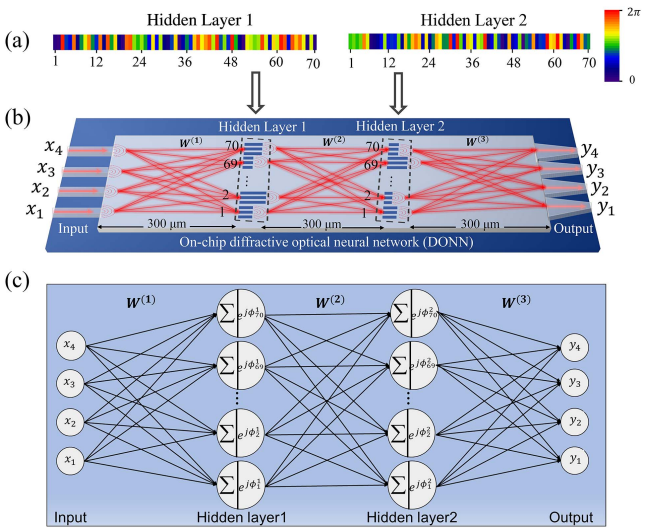

Fig. 3. (a) Weights of neurons on the two hidden layers of the on-chip DONN designed for the vowel recognition task. (b) The schematic of the on-chip DONN. Each diffractive unit on a given layer acts as a secondary source of a wave. Each diffractive unit is a slot group composed of three identical silicon slots filled with silicon dioxide and represents a single neuron in the DONN system. (c) Logic diagram of (b), which mathematically describes the physical calculation process of the on-chip DONN. $\mathbf{W}^{(k)}$ represents the kth diffraction matrix derived from the on-chip electromagnetic propagation model [Eq. (2)]. (x_1, x_2, x_3, x_4) is the input and (y_1, y_2, y_3, y_4) is the output.

3.3. Numerical simulations

The training process of the vowel recognition classification was conducted in Pytorch, which is a package for Python. The light diffraction connection in the process of forward and error backward propagation followed the modified Huygens–Fresnel principle. The data in the vowel recognition dataset were encoded onto the amplitude of light.

4. Numerical Calculation and Experiment

4.1. Structure design

Combining forward propagation, error backpropagation, and gradient descent algorithms, the structural parameters of an on-chip DONN can be obtained by pre-training through a computer in advance based on the OCEPM[18,39]. An on-chip DONN with two hidden layers is proposed, and the weight parameters on two hidden layers (HLs) are trained on the premise of fixing the super parameters of the DONN. The weight parameters on each HL obtained in the pre-training process, which are displayed in the form of pixels, as shown in Fig. 3(a), can be equivalently mapped onto the light in the form of the phase difference. The phase difference is realized by the optical path interval generated by the light passing through the slot group (composed of identical silicon slots filled with silicon dioxide) with different lengths. The length of the identical silicon slots in the slot group filled with silicon dioxide is calculated by

$$L_i = \frac{\Delta\varphi_i}{(n_E - n_S)\cdot k_0}, \tag{3}$$

where L_i is the length of the identical silicon slots filled with silicon dioxide in the ith group, n_E is the effective refractive index (ERI) of the slot group filled with silicon dioxide through which light passes, n_S is the ERI of the slab waveguide, $k_0 = 2\pi/\lambda$ is the wavenumber of light that travels in the vacuum, and $\Delta\varphi_i$ is the phase difference generated by the ith slot group filled with silicon dioxide[18,40]. Here, the phase difference $\Delta\varphi_i$ is pre-trained on the computer according to specific machine learning tasks.

In Fig. 3(b), the length of the HLs was 105 μm along the Y-axis. Each HL contained 70 neurons, and each value of the neurons is mapped by a slot group (consisting of three identical silicon slots filled with silicon dioxide). The center distance between the adjacent silicon slot filled with silicon dioxide (SSSD) is 500 nm, the period of the slot group is 1.5 μm, the width of the SSSD is 200 nm, and the thickness of the SSSD is 220 nm. The distance between two successive HLs was 300 μm along the X-axis. In addition, the input features are loaded onto the corresponding input single-mode waveguides and propagate directly into the slab waveguide, and then propagate 300 μm through the slab waveguide to reach the first HL. After light exits the last HL, it also propagates 300 μm until it reaches the output layer of the network, with four detector regions "y_i" $(i = 1, 2, 3, 4)$ arranged in the output section. Each output detector region is assigned a specific category. The width of each detector region was 8 μm, and the distance between the centers of the two neighboring detector regions was 8 μm. Therefore, for the designed on-chip DONN, the footprint of the

on-chip DONN can be calculated as $0.105 \times 3 \times 0.3 = 0.0945\,\mathrm{mm}^2$. The physical calculation process can be mathematically described in Fig. 3(c), and its specific formula expression is shown in Eq. (4), where "T" represents the matrix transpose. Our previous research[18,39] outlined the specific and detailed design method of the on-chip DONN structure.

$$\begin{bmatrix} y_4 \\ y_3 \\ y_2 \\ y_1 \end{bmatrix}^T = [\mathbf{W}^{(3)}]^T_{4\times 70} \begin{bmatrix} e^{j\phi^2_{70}} & 0 & 0 & \cdots & 0 \\ 0 & e^{j\phi^2_{69}} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & e^{j\phi^2_1} \end{bmatrix} [\mathbf{W}^{(2)}]^T_{70\times 70} \times \begin{bmatrix} e^{j\phi^1_{70}} & 0 & 0 & \cdots & 0 \\ 0 & e^{j\phi^1_{69}} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & e^{j\phi^1_1} \end{bmatrix} [\mathbf{W}^{(1)}]^T_{70\times 4} \begin{bmatrix} x_4 \\ x_3 \\ x_2 \\ x_1 \end{bmatrix}. \tag{4}$$

4.2. Numerical calculation

In this work, an on-chip DONN for the task of vowel recognition is proposed. Meanwhile, the vowel recognition data is obtained from the pronunciation of 10 types of letters by 90 different people, here, 4 types of phonemes are selected as the classification prediction categories, namely "hid", "hEd", "hYd", and "hOd". Therefore, there is a total of 360 sets of data, which is divided into a training set and a testing set by 1:1, that is, the training set and testing set have 180 sets of data, respectively. In addition, the features of the vowel recognition dataset are compressed into four features through a fully connected layer network. Then, these features are mapped onto the amplitude of light. Based on the OCEPM, the on-chip DONN with two HLs was optimized and used for classification on the vowel recognition dataset. Figure 4(a) shows the loss values for the training set and the accuracy values for the blind testing set during the learning procedure. The confusion matrix for the blind testing set in the numerical calculation is depicted in Fig. 4(b), and the prediction accuracy is 98.3%. The recognition results of the phonemes "hid", "hEd", "hYd", and "hOd" are shown in Figs. 4(c), 4(d), 4(e), and 4(f), respectively.

4.3. Experimental verification of space-time conversion performance

To verify the performance of the space-time conversion of the arrayed waveguides, four single-mode waveguides with fixed delay lines were fabricated using EBL technology based on an SOI platform, which is shown in Fig. 5. The incremental length

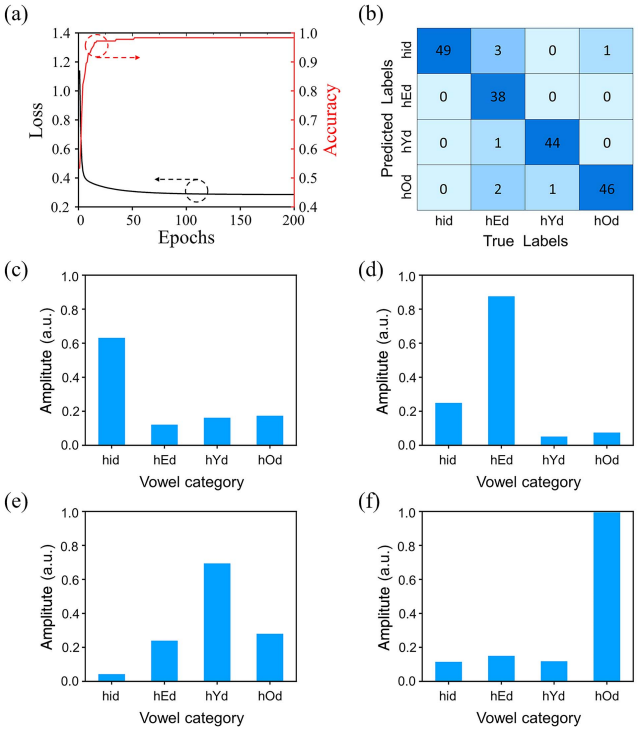

Fig. 4. (a) Loss curve on the training set (black line) and accuracy curve on the blind testing set (red line) for the optimized on-chip DONN during the learning procedure. (b) The confusion matrix by numerical calculation for the blind testing sets. (c)–(f) The display of the on-chip DONN classification results of the different types of vowel phonemes "hid", "hEd", "hYd", and "hOd", respectively.

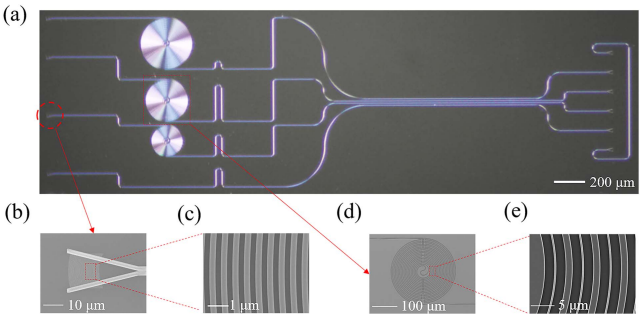

Fig. 5. (a) Microscopic view of the fabricated arrayed waveguides. (b) Vertically coupled grating. (c) Local close-up of the vertically coupled grating. (d) Ring delay line. (e) Local close-up of the ring delay line.

of the fixed delay line is calculated based on Eq. (1), which is 7.1429 mm, corresponding to a 100 ps time delay.

In this experiment, four chronological Gaussian pulses with different amplitudes generated by an arbitrary waveform generator (AWG) are loaded on the continuous-wave laser through the amplitude modulator (AM). The time slot of the input signal pulse is 100 ps, corresponding to the calculated time delay. Then, the input signals are coupled into the on-chip arrayed waveguides from the four vertically coupled gratings by a 1×4 fiber star-coupler outside the chip. Ultimately, the serial signals in

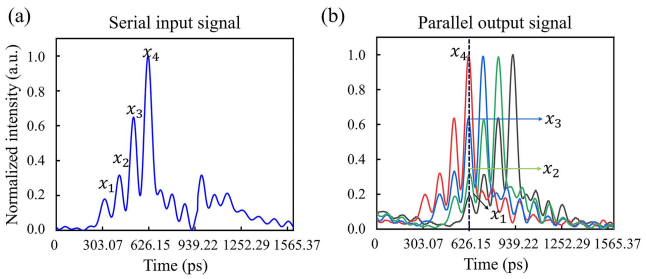

Fig. 6. (a) Serial input signals. (b) The output result of the serial input signals after passing through the arrayed waveguides.

chronological order appear simultaneously at a certain moment on the output interface through the fixed true-delay lines owned by the on-chip arrayed waveguides. Figure 6(a) is a serial input signal, and Fig. 6(b) shows the output result of the serial input signal appearing at 626.15 ps simultaneously after passing through the arrayed waveguides, which proves that the designed arrayed waveguides with true-delay lines can achieve better space-time conversion function. Meanwhile, in Fig. 5, before compensating for losses, the losses of the arrayed waveguides from bottom to top were 15.38, 17.93, 22.03, and 25.2 dB, respectively.

4.4. System experimental implementation

In this work, we designed a system experimental setup (Fig. 7) based on the experimental verification results of arrayed waveguides and theoretically implemented the numerical calculation of the on-chip DONN. When the serial input signals pass through the designed arrayed waveguides with a fixed true-delay line, they will appear simultaneously in the input interface of the on-chip DONN in parallel at a certain moment. Based on the input of the parallel signals at that time, the on-chip DONN will perform inference calculations according to the characteristics of the input signals and give calculation results at a specific time on the output interface of the on-chip DONN. The intensity of different light field distributions on the output interface of the on-chip DONN is coupled into the single-mode waveguide through an inverse taper and then collected by a high-speed photodetector from the vertical coupling grating. Finally, the eventual classification results are given in the form of light intensity. As for the research on on-chip DONNs, related theoretical simulations[17,18,33] and experimental verifications[29,39] have been conducted.

5. Discussion

5.1. Computation speed and energy efficiency

The proposed on-chip DONN architecture has the potential to process high-dimensional big data at high speeds and low power consumption. Once all the parameters have been trained and mapped onto the physical structures, the whole computing procedure is performed optically in a passive manner. Assuming that the loading signal modulation rate is f Gbps, the input signal is $M \times 1$ vector in parallel after the arrayed waveguides with a fixed delay on the DONN input interface. The on-chip DONN has N neurons at each HL, implementing m layers of the $N \times N$ matrix multiplication and operating at a D_{bw} GHz photodetection rate. The number of floating-point operations per second (FLOPS) to match the optical network is obtained using the following equation[14]:

$$R = 2m \times N^2 \times \min\left(\frac{f}{2M-1}, D_{bw}\right)\text{FLOPS}, \tag{5}$$

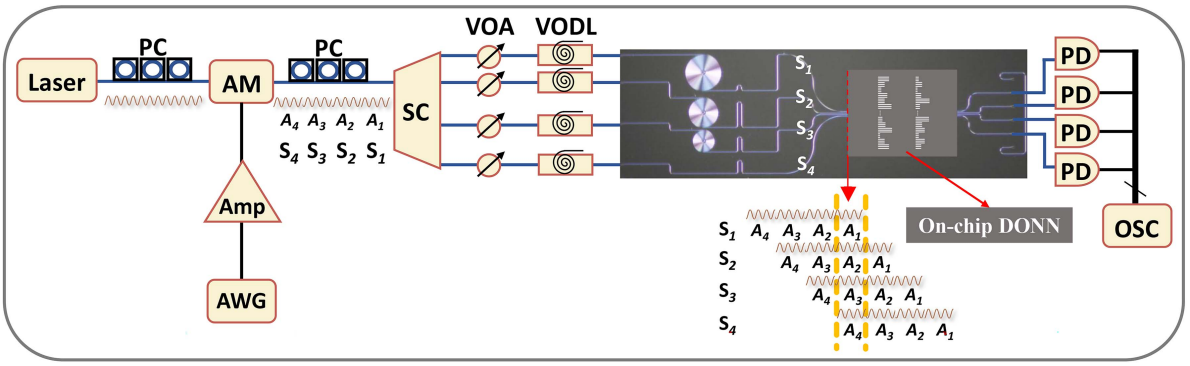

Fig. 7. (a) System experimental setup. Four chronological Gaussian pulses (S1, S2, S3, and S4) with different amplitudes generated by an arbitrary waveform generator (AWG) are loaded on the continuous-wave laser through the amplitude modulator (AM). The time slot of the input signal pulse is 100 ps corresponding to the calculated time delay. Then, the input signals with various amplitudes are coupled into the on-chip arrayed waveguides from the four vertically coupled gratings by a 1×4 fiber star-coupler (SC) outside the chip. Next, the serial signals in chronological order will exist in parallel at a certain moment on the output interface through the fixed true-delay lines owned by the on-chip arrayed waveguides. Ultimately, after the required parallel signals enter the on-chip DONN to complete the inference calculation, its output optical field will be coupled into the single-mode waveguides through the corresponding inverse tapers at the DONN output interface and received by the on/off-chip photodetector (PD). At this time, a classification task is completed. PC, polarization controller; Amp, electric amplifier; VOA, variable optical attenuator (which is to compensate the fabricating error of the on-chip delay lines); VODL, variable optical delay line; OSC, digital oscilloscope.

Table 1. Comparison between the Proposed on-Chip DONN and Other Integrated Works.

Study	Computing Unit	Signal Loading Method	Signal Loading Rate	On-Chip Power Consumption
Ref. [14]	Mach-Zehnder interferometer (MZI)	Multiple thermo-optic phase shifter (MTPS)	~ Kbps	1.92 W
Ref. [31]	MZI	MTPS	~ Kbps	3.92 W
Ref. [30]	MZI	MTPS	~ Kbps	7.7 mW
Ref. [37]	Multimode interferometer	MTPS	~ Kbps	2.04 W
Ref. [29]	Subwavelength unit (SWU)	DMD and Lens group	~ Kbps	Passive
Ref. [39]	SWU	MTPS	~ Kbps	120 mW
This work	SWU	Arrayed waveguides	~ 1.4 Gbps	Passive

where R is the number of operations per second, which is related to the modulation rate f, the scale of the arrayed waveguides M, the number of HLs m, the number of neurons on each HL N, and the detection frequency of the photodetectors D_{bw}. Therefore, in this work, the computation speed is approximately 1.4×10^{13} FLOPS calculated using Eq. (5), which is one order of magnitude higher than the performance of modern GPUs, which typically perform at 10^{12} FLOPS[21]. The on-chip part of the whole system, including the arrayed waveguides and the DONN, is fully passive during the system operation, and no additional energy supply is required except for the laser (32 mW). Therefore, excluding the energy consumed by peripheral electronic devices and drive circuits, the energy consumed by the chip in the optical computing process is about 2.286×10^{-17} J/FLOP.

5.2. Performance of proposed DONN framework

Recently, certain research on integrated ONNs has been conducted. Table 1 compares the proposed on-chip DONN with other integrated DONNs and ONNs. It is not difficult to find that the computing unit of on-chip DONNs is a sub-wavelength structure. Thus, its integration degree is higher than that of other on-chip ONNs. In addition, the signal loading method of the arrayed waveguides is all passive. Thus, the energy consumption only spends on the single modulator, and the energy consumption during the signal loading part would not increase with the increase of input dimensions. However, the other signal loading methods are very limited by energy consumption when dealing with high dimensional tasks because the energy consumption would increase sharply with the increase of the number of phase shifters used for signal loading.

6. Conclusion

A wholly passive optical on-chip DONN based on an SOI platform was proposed. The signal loading method of the proposed on-chip DONN is achieved using the arrayed waveguides, which can convert the serial input signals in chronological order into parallel signals at a certain time and feed them into the on-chip DONN. The advantage of this approach is that it overcomes the problem of the existing integrated ONNs, which are limited in the input of high-dimensional signals. In addition, the conversion of serial signals to parallel signals using space-time interleaving would not consume more energy because the whole conversion process is passive. The proposed method would make the on-chip DONNs more widely applicable and may, to a certain extent, promote the further development of silicon-based photon computing.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 62135009) and the Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (No. Z221100005322010).

†These authors contributed equally to this work.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**, 84 (2017).
2. T. Ananthanarayana, P. Srivastava, A. Chintha, A. Santha, B. Landy, J. Panaro, A. Webster, N. Kotecha, S. Sah, T. Sarchet, R. Ptucha, and I. Nwogu, "Deep learning methods for sign language translation," ACM Trans. Access. Comput. **14**, 1 (2021).
3. I. Kruglov, O. Mishulina, and M. Bakirov, "Quantile based decision making rule of the neural networks committee for ill-posed approximation problems," Neurocomputing **96**, 74 (2012).
4. G. Ozkan and M. Inal, "Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems, "Appl. Soft Comput. **24**, 232 (2014).
5. Y. Liu, K. Qian, K. Wang, and L. He, "Effective scaling of blockchain beyond consensus innovations and Moore's law: challenges and opportunities," IEEE Syst. J. **16**, 1424 (2022).
6. M. B. Taylor, "The evolution of bitcoin hardware," Computer **50**, 58 (2017).
7. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," Nature **529**, 484 (2016).

8. Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE J. Solid-State Circuits **52**, 127 (2016).
9. J. Misra and I. Saha, "Artificial neural networks in hardware: a survey of two decades of progress," Neurocomputing **74**, 239 (2010).
10. S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. Mckinstry, T. Melano, D. R. Barch, C. NolfoDi, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, "Convolutional networks for fast, energy-efficient neuromorphic computing," Proc. Natl. Acad. Sci. U.S.A. **113**, 11441 (2016).
11. C. S. Poon and K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities," Front Neurosci. **5**, 108 (2011).
12. A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," Nature **538**, 471 (2016).
13. A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (2016).
14. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441 (2017).
15. X. Lin, Y. Rivenson, N. T. Yardimici, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004 (2018).
16. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**, 208 (2019).
17. S. Zarei, M. R. Marzban, and A. Khavasi, "Integrated photonic neural network based on silicon metalines," Opt. Express **28**, 36668 (2020).
18. T. Z. Fu, Y. B. Zang, H. H. Huang, Z. M. Du, C. Y. Hu, M. G. Chen, S. G. Yang, and H. W. Chen, "On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model," Opt. Express **29**, 31924 (2021).
19. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52 (2021).
20. M. Y. S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. Deweese, "Design of optical neural networks with component imprecisions," Opt. Express **27**, 14009 (2019).
21. I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. H. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 7700412 (2020).
22. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, "All-optical nonlinear activation function for photonic neural networks [Invited]," Opt. Mater. Express **8**, 3851 (2018).
23. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**, 1132 (2019).
24. J. M. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**, 756 (2018).
25. E. Khoram, A. Chen, D. Liu, L. Ying, Q. Wang, M. Yuan, and Z. Yu, "Nanophotonic media for artificial neural inference," Photonics Res. **7**, 823 (2019).
26. D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).
27. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through *in situ* backpropagation and gradient measurement," Optica **5**, 864 (2018).
28. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Fourier-space diffractive deep neural network," Phys. Rev. Lett. **123**, 023901 (2019).
29. Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," Nat. Commun. **13**, 2131 (2022).
30. H. H. Zhu, J. Zou, H. Zhang, Y. Z. Shi, S. B. Luo, N. Wang, H. Cai, L. X. Wan, B. Wang, X. D. Jiang, J. Thompson, X. S. Luo, X. H. Zhou, L. M. Xiao, W. Huang, L. Patrick, M. Gu, L. C. Kwek, and A. Q. Liu, "Space-efficient optical computing with an integrated chip diffractive neural network," Nat. Commun. **13**, 1044 (2022).
31. H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. Yung, Y. Z. Shi, F. K. Muhammed, G. Q. Lo, X. S. Luo, B. Dong, D. C. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**, 457 (2021).
32. F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," Nature **606**, 501 (2022).
33. T. Yan, R. Yang, Z. Zheng, X. Lin, H. Xiong, and Q. Dai, "All-optical graph representation learning using integrated diffractive photonic computing units," Sci. Adv. **8**, eabn7630 (2022).
34. Z. Xu, X. Yuan, T. Zhou, and L. Fang, "A multichannel optical computing architecture for advanced machine vision," Light Sci. Appl. **11**, 255 (2022).
35. C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, and H. Chen, "Performing optical logic operations by a diffractive neural network," Light Sci. Appl. **9**, 59 (2020).
36. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367 (2021).
37. X. Zhao, H. Lv, C. Chen, S. Tang, X. Liu, and Q. Qi, "*On-chip Reconfigurable Optical Neural Networks*," PPR283160 (Research Square Preprints, 2021).
38. D. H. Deterding, "Speaker Normalisation for Automatic Speech Recognition," Ph.D. Thesis (University of Cambridge, 1990).
39. T. Fu, Y. Zang, Y. Huang, Z. Du, H. Huang, C. Hu, M. Chen, S. Yang, and H. Chen, "Photonic machine learning with on-chip diffractive optics," Nat. Commun. **14**, 70 (2023).
40. Z. Wang, T. Li, A. Soman, D. Mao, T. Kananen, and T. Gu, "On-chip wavefront shaping with dielectric metasurface," Nat. Commun. **10**, 3547 (2019).