First experimental demonstration of highly scalable and reconfigurable optical convolution computing based on wavelength routing

Jialin Cheng*[1], Chong Li[2], Jun Dai[1], Chu Yayan[2], Xinxiang Niu[2], Xiaowen Dong[2], Jian-Jun He[1]

[1]State Key Laboratory of Extreme Photonics and Instrumentation, Centre for Integrated Optoelectronics, College of Optical Science and Engineering, Zhejiang University, Hangzhou 310027, China; [2]Huawei Technologies Co., Ltd., Bantian, Longgang, Shenzhen, Guangdong 518000, China; *12230069@zju.edu.cn, phone 18772272503

ABSTRACT

A highly scalable and reconfigurable optical convolution paradigm based on wavelength routing is proposed, which leverages the unique sliding property of an arrayed waveguide grating router (AWGR) to execute the sliding window operation of convolution in the wavelength-space domains. By directly loading two input vectors onto two modulator arrays, the convolution result is instantaneously generated at a photodetector array at the speed of light propagation. This enables the entire convolution computation to be executed within one clock cycle, eliminating the necessity for preprocessing or decomposition into elementary MAC operations. The proposed optical convolution unit (OCU) has striking advantages of high scalability, high speed, and processing simplicity compared to those based on optical matrix-vector multipliers (MVM). A proof-of-concept experiment employing standalone optical components is devised to validate optical convolution computing principles with one-bit accuracy. The classification of ten handwritten digit classes sourced from the MNIST database is experimentally demonstrated, achieving a precision of 4-bit. New algorithms for data splitting and reorganization were concurrently introduced to facilitate the convolution calculation of two-dimensional image data. Notably, through Field-Programmable Gate Array (FPGA) across varying data transmission speeds of 1 MHz, 5 MHz, and 10 MHz, inference accuracy rates of 97.32%, 96.25%, and 94.50% were respectively achieved, demonstrating the robustness and versatility of the proposed paradigm.

Keywords: optical convolution computing, arrayed waveguide grating router, optical convolution unit, MNIST database

1. INTRODUCTION

With artificial intelligence (AI) technologies based deep learning[1] advance rapidly, the use of deep neural networks is expanding across numerous fields. However, this growth escalates swiftly the computational demand for training deep neural networks. As traditional electronic computing encounters limitations due to Moore's Law, the search for computing architectures that are faster and more energy-efficient, tailored for deep neural networks, becomes increasingly paramount. Optical methods show great potential for the next wave of neural network accelerators due to their advantages of ultra-wide bandwidth, low power consumption, and inherent parallelism, making them compelling candidates for accelerating deep learning hardware. Conventional experimental configurations for handling input dimensions of N require N^2 optical elements, including Mach-Zehnder interferometers (MZI)[2] or micro-ring resonators (MRR)[3], to execute multiply-accumulate (MAC) operations. These methodologies struggle to achieve both high efficiency and scalability to support large networks sizes[4].

To address the above issues, we propose a highly scalable and reconfigurable optical convolution paradigm based on wavelength routing, leveraging the unique sliding property of an arrayed waveguide grating router (AWGR)[5] to execute the sliding window operation of convolution in the wavelength-space domains. The AWGR is a N×N device that can be used for wavelength-routing and distributed optical switching in wavelength division multiplexing (WDM) systems[6, 7]. By directly loading two input vectors onto two modulator arrays, the convolution result is instantaneously generated at the output photodetector array at the speed of light propagation . This enables the entire convolution computation to be executed within one clock cycle, eliminating the necessity for preprocessing or decomposition into elementary MAC operations. The proposed optical convolution unit (OCU)[8] has striking advantages of high scalability, high speed, and processing simplicity compared to those based on optical matrix-vector multipliers (MVM)[2, 3]. A proof-of-concept experiment employing standalone optical components was devised to validate the principles of optical convolution computing based AWGR with one-bit accuracy. The classification of ten handwritten digit classes sourced from the MNIST database[9] is experimentally demonstrated, achieving a precision of 4-bit. New algorithms for data splitting and reorganization were concurrently introduced to facilitate the convolution calculation of two-dimensional image data. Notably, through Field-Programmable Gate Array (FPGA) across varying data transmission speeds of 1 MHz, 5 MHz, and 10 MHz, inference accuracy rates of 97.32%, 96.25%, and 94.50% were respectively achieved, demonstrating the robustness and versatility of the proposed paradigm.

2. PRINCIPLE

Figure 1 illustrates the basic structure of the N×K AWGR , which includes N input waveguides, an input star coupler (the first free propagation region, FPR1), a grating composed of an array of waveguides with equal length differences, an output star coupler (FPR2), and K output waveguides. When an optical signal is transmitted from an input waveguide to the input FPR1, the light diverges due to diffraction and is then coupled to the arrayed waveguides. After propagating though the arrayed waveguide grating, the light from each arrayed waveguide is diffracted in the output FPR2 and then focused onto a specific position on the imaging surface according to its wavelength, due to the interference effect. Consequently, the light of different wavelengths is coupled into different output waveguides. As depicted in Figure 1, M represents the number of wavelength, N represents the number of inputs ports of the AWGR and K is number of the output ports of the output ports of the AWGR. For executing full convolution computation, these numbers satisfy the relationship K = M + N – 1. A distinctive feature of the AWGR is its “slide property”, wherein the transition of the multi-wavelength input signals from one input port to an adjacent input port results in an equivalent shift in the corresponding output ports by the same number of channels. This property closely mirrors the concept of sliding window operation in vector-vector convolution computations.

Building upon this pivotal sliding property of the AWGR for wavelength routing, we propose a novel Optical Convolution Unit (OCU) for efficiently executing the convolution computing in the optical domain. As shown in **Figure 2**, within the OCU framework, vector ***A*** is encoded into intensity signals of different wavelengths via a directly modulated or externally modulated light source array denoted as ***A***. Simultaneously, vector ***B*** is loaded onto another modulator array denoted as ***B***. All the signals of different wavelengths of vector A are combined via a multiplexer and then sent to each modulator (designated as Bj) of the modulator array B through a power splitter. Each modulator multiplies the vector element Bj to its input signals of different wavelengths carrying vector ***A***, and then send them to the corresponding input port of the AWGR. As illustrated by the dashed frames in Figure 1, the AWGR demultiplexes the signals of different wavelengths from an input port to different output ports with the above-mentioned space invariant sliding property.

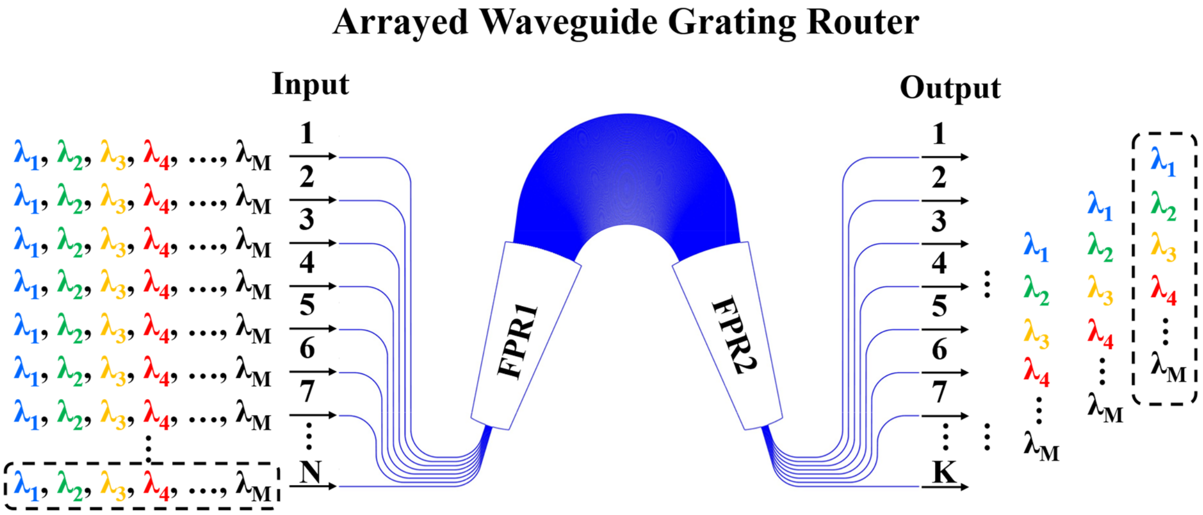

Arrayed Waveguide Grating Router
Input
Output
$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_M$
1
2
3
4
5
6
7
N
FPR1
FPR2
K
λ_1
λ_2
λ_3
λ_4
λ_M

Figure 1. Schematic diagram of a N×K AWGR wavelength router.

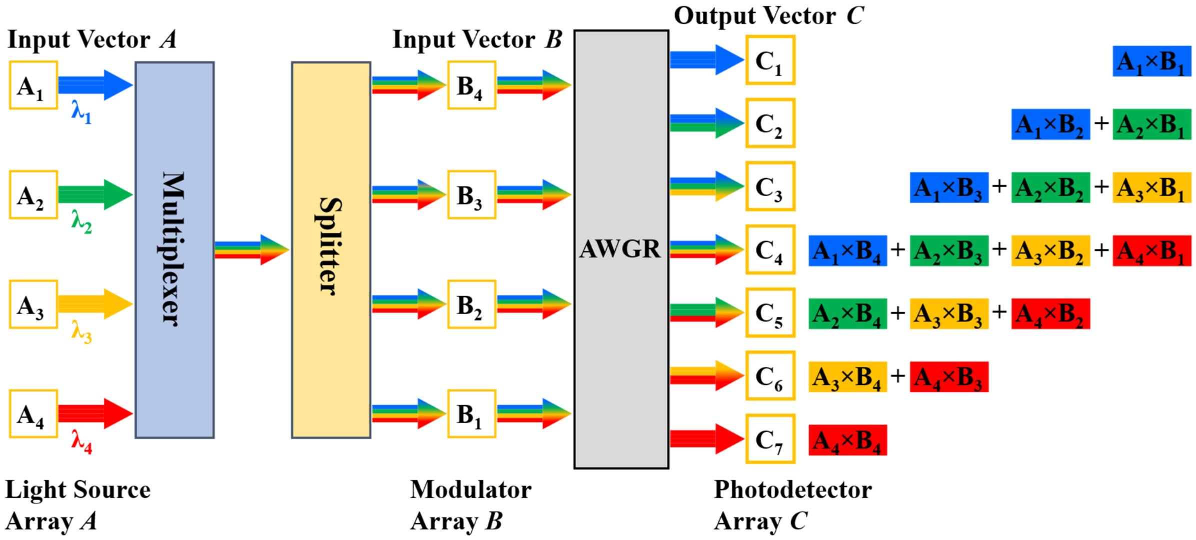


Figure 2. Optical Convolution Unit (OCU)

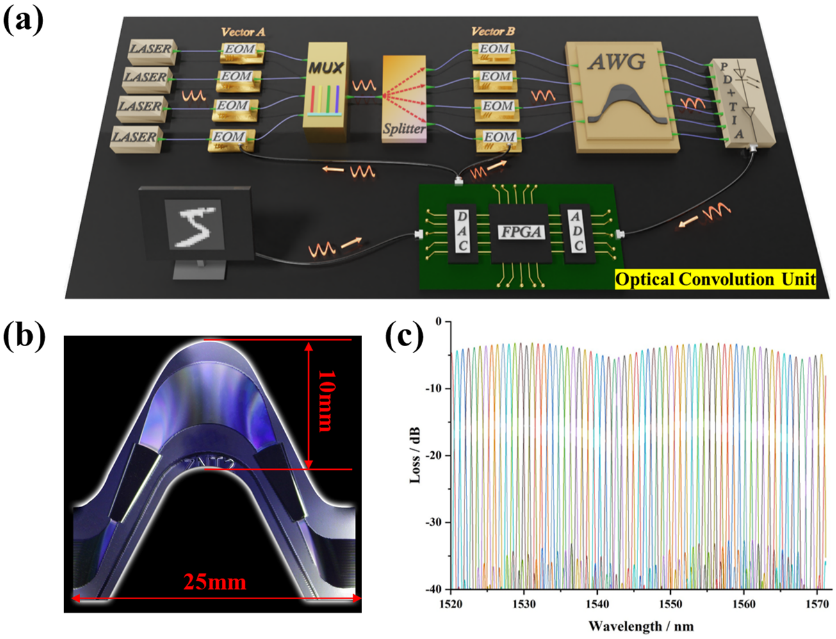


Figure 3. (a) OCU experimental setup. (b) Microscopic image of the fabricated (b) Microscopic image of the fabricated 32×32 AWGR with a footprint of 25×10 mm². (c) Measured transmission spectra from one input port to 32 output ports.

3. EXPERIMENT RESULT

3.1 Experimental setup

Figure 3(a) provides a detailed experimental demo setup of the proposed OCU paradigm, which comprises a laser array, EOM (Electro-Optic Modulator) array A, MUX (Wavelength-Division Multiplexer), splitter, EOM array B, AWGR, a photodetector (PD) array integrated with a transimpedance amplifier (TIA) circuit, and an FPGA driver board equipped with digital-to-analog converter (DAC) and analog-to-digital converter (ADC). The AWGR used in the experiment is fabricated on the silica-on-silicon platform. It features a configuration with 32 input and 32 output ports, enabling it to perform vector convolution calculations with a maximum length of 16 for both vectors. Operating at a central wavelength of 1550 nm, the 32×32 AWGR characterizes a channel spacing of 100 GHz (0.8 nm). The AWGR module used in the current experiment includes a built-in heater to ensure the wavelength stability of its channels, which can potentially be eliminated by using an athermal design to reduce the power consumption. A photograph of the AWGR chip is shown in **Figure 3(b)**. It has a footprint of 25×10 mm^2 with curved profile. **Figure 3(c)** shows the measured transmission spectra of the AWGR for input port #17. The 32×32 AWGR is designed with cyclic characteristics, with a free spectral range (FSR) equal to 3200 GHz (25.6 nm). The insertion loss for the central channel is about 3.5 dB, with a channel non-uniformity of 2.5 dB. Notably, the observed crosstalk is below -31 dB.

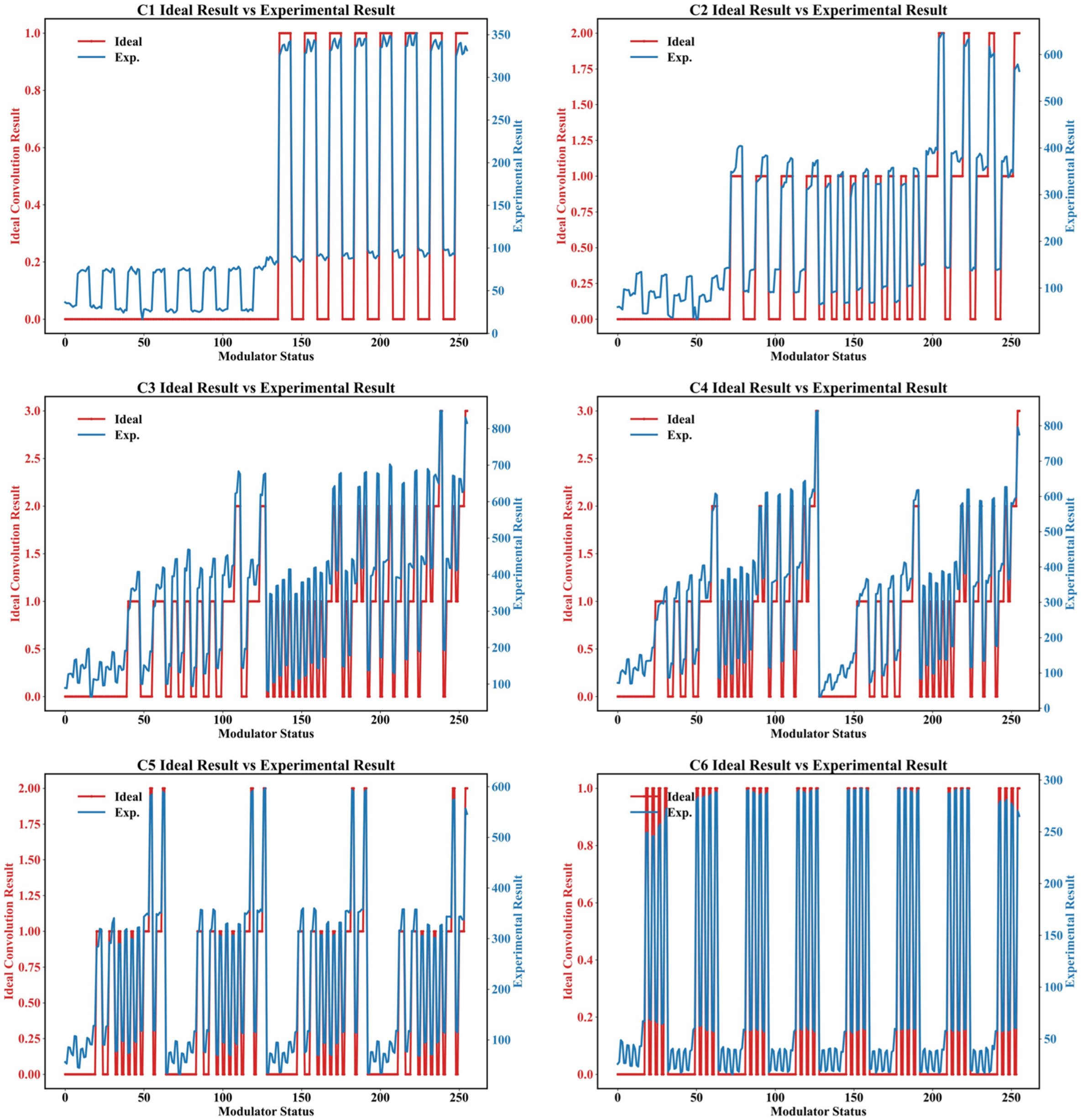


Figure 4. AWGR C$_1$ to C$_6$ output channel one-bit optical convolution results

3.2 One-bit optical convolution

Initially, a 4×4 vector-vector optical convolution computing with one-bit accuracy was conducted, which necessitated 8 modulators and thus 256 modulation states in the system. However, due to a defect in one modulator, the actual setup resulted in a 4×3 convolution instead. Consequently, the AWGR was configured with 6 output channels (C$_1$ to C$_6$). Channels C$_3$ and C$_4$, which overlapped with 3 wavelength signals, exhibited higher noise levels, as depicted in **Figure 4** (The red line indicates the ideal results, whereas the blue line depicts the experimental results). Nonetheless, 3 distinct signal levels were still identifiable, indicating that the ADC has a high tolerance for quantization noise in one-bit operations. This confirms the feasibility of our proposed AWGR wavelength routing optical convolution scheme.

3.3 MNIST dataset recognition

As a proof-of-concept demonstration for the OCU proposed in this work, we constructed a 4×4 convolution operator based on the wavelength routing to implement an optical-electronic hybrid convolutional neural network (CNN) for MNIST dataset recognition. **Figure 5** illustrates the architecture of the neutral network, comprising two convolutional layers and three fully connected layers. The input image size for the first layer is 28×28. This layer is performed by the AWGR-OCU in the optical domain, featuring 6 convolution kernels with a size of 2×2. The resulting output tensor size is 6×27×27. The detailed network structure is provided in **Table 1**.

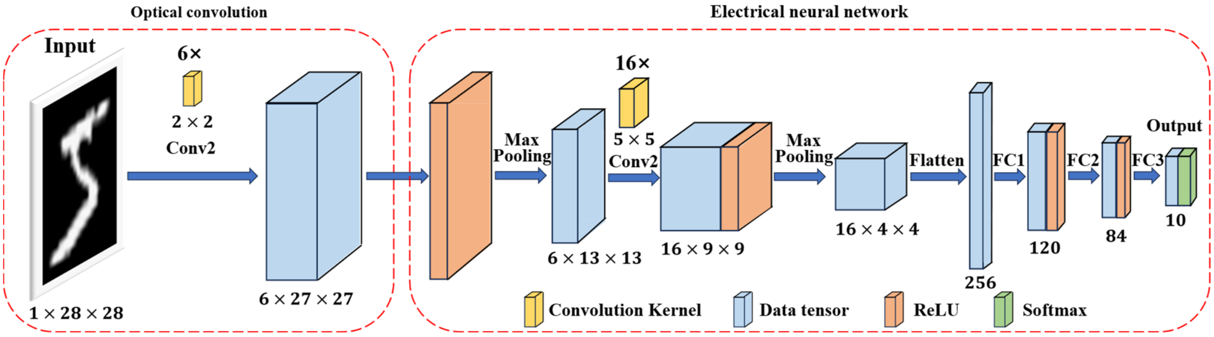


Figure 5. Framework of the optical-electronic hybrid convolutional neural network for MNIST dataset in accordance with LeNet-5.

Table 1. Neural network structure

Layer	Type	Kernel size	Stride	No. of filters
Conv 1	**Conv2d**	**2×2**	**1**	**6**
Pool 1	Maxpool2d	2×2	2	
Conv 2	Conv2d	5×5	1	16
Pool 2	Maxpool2d	2×2	2	
FC1	FC/Linear			120
FC2	FC/Linear			84
FC3	FC/Linear			10

Given that the MNIST dataset comprises images with a size of 28×28 pixels, while the preliminary Optical Convolutional Unit (OCU) system operates at a scale of 4×4, it becomes necessary to partition the data to address this difference in dimensions. New algorithms for data splitting and reorganization were concurrently introduced to facilitate the convolution calculation of two-dimensional image data, as shown in **Figure 6**. For simplicity, the convolution kernel size is set to 2×2, and the image data is 4×2. For two-dimensional data, the convolution kernel and image data are first split by rows to obtain one-dimensional vector data. The OCU then performs convolution on these vectors. The results of these convolutions are directly added together to yield the final convolution result for the two-dimensional data. This splitting method avoids wasting computing resources and requires the minimal number of calculations. For vector lengths greater than 4, column splitting can also be performed similarly. Additionally, scaling up the OCU can accommodate convolution calculations for larger vector lengths. In practical experiments, the first layer of the convolution kernel data is 6×2×2, and the image size is 28×28. Using this splitting method, a total of 392 calculations are required.

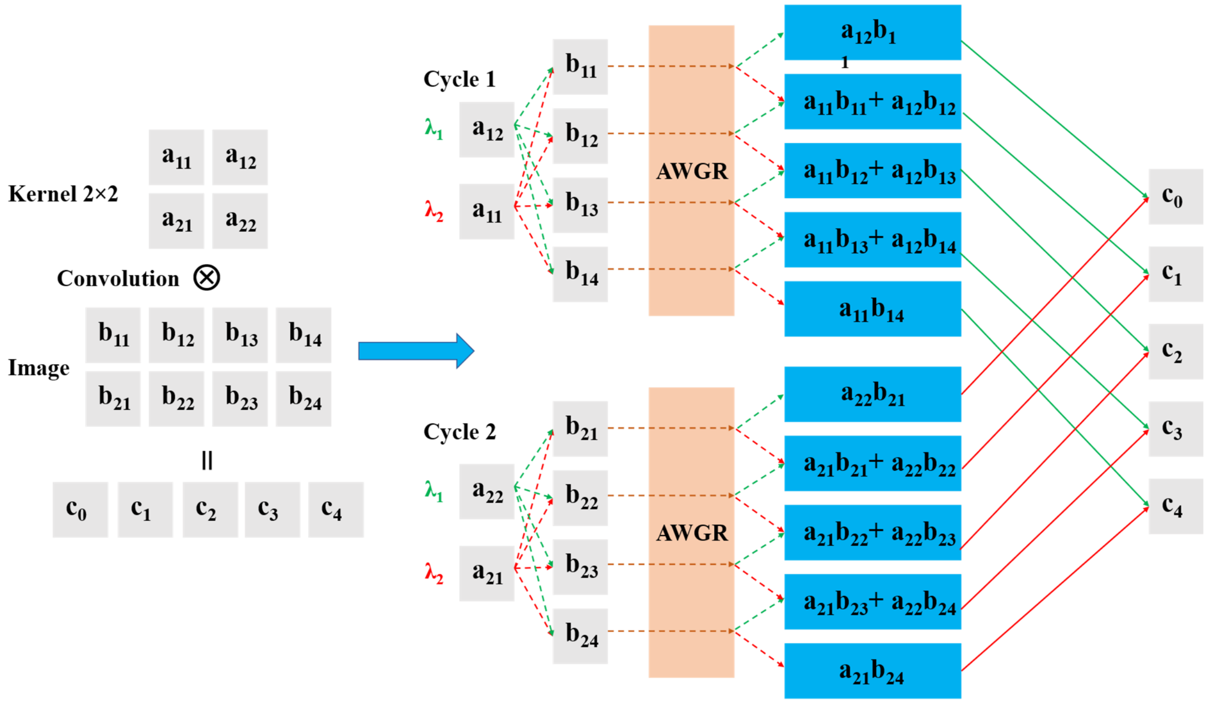
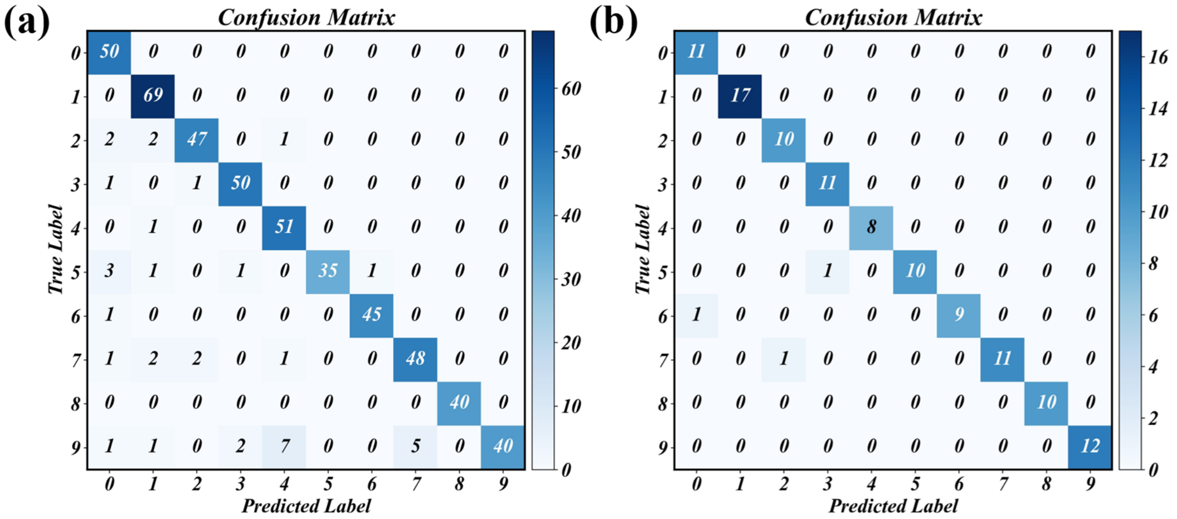


Figure 6. Two-dimension data splitting and reorganization algorithm for 4×4 AWGR-OCU.

The MNIST dataset is split into the training (60,000 images) and testing sets (10,000 images). Our model is trained on the entire training set. To maintain compatibility with the 4-bit precision of the system, the input images utilized during network training undergo quantization to match the same precision. When the FPGA sends data at a speed of 1MHz, a subset of 512 images from the original MNIST test dataset is selected, and the handwritten digit recognition experiments are conducted following the aforementioned procedure. Subsequent to obtaining the computational results from the optical domain convolution layer, the ensuing image processing is conducted directly in the electrical domain, maintaining the parameters of the remaining network layers unaltered. A recognition accuracy of 92.77% is achieved on this randomly selected set of 512 images. The confusion matrix is depicted in **Figure 7(a)**, where out of a total of 512 images, 475 were precisely inferred and correctly identified.

To enhance the performance of the neural network, a fine-tuning is performed in the training process. The feature map of the 512 images is divided into re-training and testing dataset in a ratio of 400:112. Of these, 400 images are designated for re-training within the electrical domain, emphasizing the adjustment of parameters in the remaining network layers to mitigate the influence of noise on convolution calculation accuracy. During the fine-tuning process, the parameters of the Conv 1 layer were kept unchanged, allowing only the parameters of the subsequent layers to undergo fine-tuning retraining. After this fine-tuning process, the network achieved a recognition accuracy of 97.32% on the testing dataset (112 images). The final confusion matrix is presented in **Figure 7(b)**. This level of performance is very close to the inference performance achieved on electronic computers, indicating that the proposed optical convolution paradigm possesses computational capabilities that are on par with those of electronic computers (99.0%).

The entire 4×4 OCU system operates on a clock frequency of 100 MHz. To achieve noise suppression, we initially represent one single user data point with a sequence spanning 100 system clock cycles (equivalent to 1 MHz data transferring speed), and average the 100 cycles at the ADC output. This meticulous approach has allowed us to achieve a network inference accuracy of 97.32% on the MNIST dataset. To study the impact of noises under different data transferring rates, we use a representation scheme equivalent to 5 MHz (where 20 clock cycles represent one data point) and 10 MHz (where 10 clock cycles represent one data point) for transmitting data. As shown in **Figure 8(a)**, increasing the data transmission speed to 5 MHz results in a slight decrease in accuracy, achieving a still commendable 96.25% after fine-tuning. At 10 MHz, a more significant decline in accuracy to 94.50% is observed, as depicted in **Figure 8(b)**.



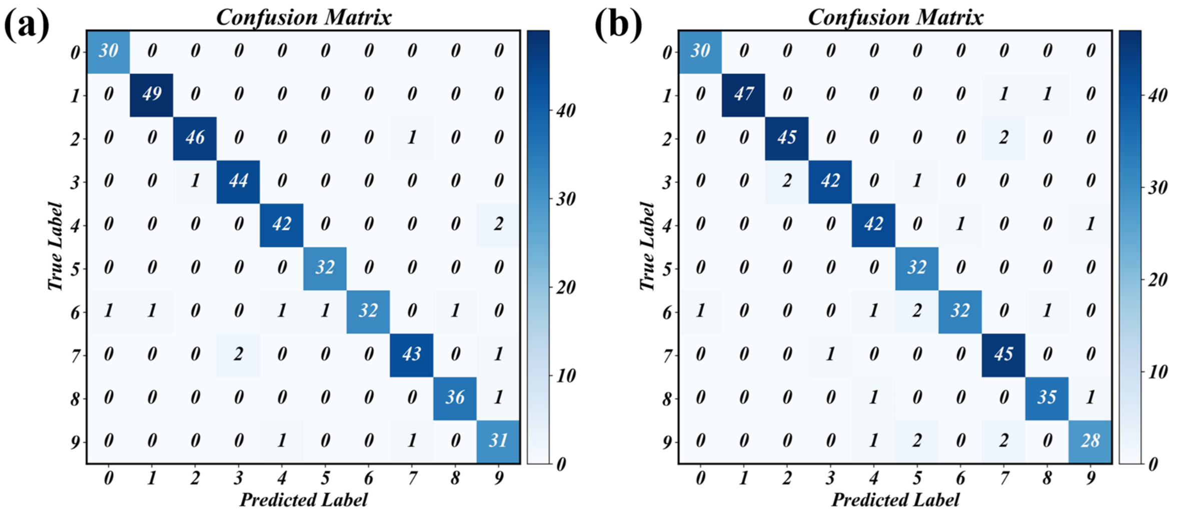
(a)

	0	1	2	3	4	5	6	7	8	9
0	50	0	0	0	0	0	0	0	0	0
1	0	69	0	0	0	0	0	0	0	0
2	2	2	47	0	1	0	0	0	0	0
3	1	0	1	50	0	0	0	0	0	0
4	0	1	0	0	51	0	0	0	0	0
5	3	1	0	1	0	35	1	0	0	0
6	1	0	0	0	0	0	45	0	0	0
7	1	2	2	0	1	0	0	48	0	0
8	0	0	0	0	0	0	0	0	40	0
9	1	1	0	2	7	0	0	5	0	40

(b)

	0	1	2	3	4	5	6	7	8	9
0	11	0	0	0	0	0	0	0	0	0
1	0	17	0	0	0	0	0	0	0	0
2	0	0	10	0	0	0	0	0	0	0
3	0	0	0	11	0	0	0	0	0	0
4	0	0	0	0	8	0	0	0	0	0
5	0	0	0	1	0	10	0	0	0	0
6	1	0	0	0	0	0	9	0	0	0
7	0	0	1	0	0	0	0	11	0	0
8	0	0	0	0	0	0	0	0	10	0
9	0	0	0	0	0	0	0	0	0	12

Figure 7. (a) Confusion matrix corresponding to 92.77% classification inference accuracy at a data transfer speed of 1 MHz. (b) Improved confusion matrix with additional fine-tuning, indicating a 97.32% inference accuracy on the MNIST dataset.



(a)

	0	1	2	3	4	5	6	7	8	9
0	30	0	0	0	0	0	0	0	0	0
1	0	49	0	0	0	0	0	0	0	0
2	0	0	46	0	0	0	0	1	0	0
3	0	0	1	44	0	0	0	0	0	0
4	0	0	0	0	42	0	0	0	0	2
5	0	0	0	0	0	32	0	0	0	0
6	1	1	0	0	1	1	32	0	1	0
7	0	0	0	2	0	0	0	43	0	1
8	0	0	0	0	0	0	0	0	36	1
9	0	0	0	0	1	0	0	1	0	31

(b)

	0	1	2	3	4	5	6	7	8	9
0	30	0	0	0	0	0	0	0	0	0
1	0	47	0	0	0	0	0	1	1	0
2	0	0	45	0	0	0	0	2	0	0
3	0	0	2	42	0	1	0	0	0	0
4	0	0	0	0	42	0	1	0	0	1
5	0	0	0	0	0	32	0	0	0	0
6	1	0	0	0	1	2	32	0	1	0
7	0	0	0	1	0	0	0	45	0	0
8	0	0	0	0	1	0	0	0	35	1
9	0	0	0	0	1	2	0	2	0	28

Figure 8. (a) Confusion matrix corresponding to 96.25% inference accuracy after fine-tuning under 5 MHz. (b) Confusion matrix indicating a 94.50% inference accuracy on the MNIST dataset with fine-tuning at a data transfer speed of 1 MHz.

The accuracy deviations across different data transferring rates highlight the limited bit precision due to system noises, which are caused by numerous factors, including the electrical and optical disturbances, and instability of some optical devices such as the temperature drift and mechanical disturbances, and polarization sensitivity of the EOMs. These noises can be mitigated through high-degree photonic integration[10,11] and enhanced drive circuit, which aligns with the objective of our on-going effort. Additionally, advanced photonic-electronic co-packaging technologies[12] can also reduce high-frequency parasitic crosstalk noises and minimize signal losses at elevated frequencies. These enhancements are crucial for improving the calculation accuracy of AWGR-OCU.

4. CONCLUSION

In this study, we have proposed a direct optical convolution computing architecture based on wavelength routing, which bring forth a range of significant advantages including high scalability, high speed, processing simplicity, minimized device counts, and high efficiency. A proof-of-concept experiment employing standalone optical components is devised to validate optical convolution computing principles with one-bit accuracy. The classification of ten handwritten digit classes sourced from the MNIST database is experimentally demonstrated, achieving a precision of 4-bit. New algorithms for data splitting and reorganization were concurrently introduced to facilitate the convolution calculation of two-dimensional image data. Notably, through Field-Programmable Gate Array (FPGA) across varying data transmission speeds of 1 MHz, 5 MHz, and 10 MHz, inference accuracy rates of 97.32%, 96.25%, and 94.50% are respectively achieved, demonstrating the robustness and versatility of the proposed paradigm. This is the first demonstration of the Waveguide AWGR for direct optical convolution computing, showing potentially superior characteristics compared to other optical computing systems reported in the literature. Our proposed optical convolution computing paradigm shows promising potential for large-scale photonic integration, laying the foundation for the next generation of ultra-high-speed artificial intelligence platforms.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature 521, 436-444 (2015).

[2] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, and D. Englund, "Deep learning with coherent nanophotonic circuits," Nature photonics 11, 441-446 (2017).

[3] B. Bai, Q. Yang, H. Shu, L. Chang, F. Yang, B. Shen, Z. Tao, J. Wang, S. Xu, and W. Xie, "Microcomb-based integrated photonic processing unit," Nature Communications 14, 66 (2023).

[4] C. Li, X. Zhang, J. Li, T. Fang, and X. Dong, "The challenges of modern computing and new opportunities for optics," PhotoniX 2, 1-31 (2021).

[5] Y. Chu, X. Dong, OPTICAL COMPUTING DEVICE AND SYSTEM AND CONVOLUTION COMPUTING METHOD, Chinese patent 202110502270.2, filed on May 8, 2021; International PCT/CN2021/124511 (WO2022/160784), filed on Oct. 18, 2021.

[6] J. Guo, Z. Fan, S. Zhou, B. Liu, J. Meng, J. Zhu, Y. Li, Q. Li, J. Zhao, and J.-J. He, "Monolithic 6× 6 transmitter-router with simultaneous sub-nanosecond port and wavelength switching," Optics Letters 47, 2762-2765 (2022).

[7] Z. Fan, J. Guo, S. Zhang, J. Zhu, J. Meng, Q. Li, Y. Li, J. Zhao, and J.-J. He, "Monolithically Integrated 8× 8 Transmitter-Router Based on Tunable V-Cavity Laser Array and Cyclic Arrayed Waveguide Grating Router," IEEE Photonics Journal 14, 1-8 (2022).

[8] J. Cheng, C. Li, J. Dai, Y. Chu, X. Niu, X. Dong, and J. J. He, "Direct Optical Convolution Computing Based on Arrayed Waveguide Grating Router," Laser & Photonics Reviews, 2301221 (2024).

[9] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," IEEE signal processing magazine 29, 141-142 (2012).

[10] C. Xiang, J. Liu, J. Guo, L. Chang, R. N. Wang, W. Weng, J. Peters, W. Xie, Z. Zhang, and J. Riemensberger, "Laser soliton microcombs heterogeneously integrated on silicon," Science 373, 99-103 (2021).

[11] C. Xiang, W. Jin, O. Terra, B. Dong, H. Wang, L. Wu, J. Guo, T. J. Morin, E. Hughes, and J. Peters, "3D integration enables ultralow-noise isolator-free lasers in silicon photonics," Nature 620, 78-85 (2023).

[12] C. Minkenberg, R. Krishnaswamy, A. Zilkie, and D. Nelson, "Co-packaged datacenter optics: Opportunities and challenges," IET optoelectronics 15, 77-91 (2021).