An Optical Neural Network Accelerator Based on A Dual Microring Modulator Array

Weiwei Pan
COLLEGE OF INFORMATION SCIENCE & ELECTRONIC ENGINEERING
Zhejiang University
Hangzhou, China
weiweipan@zju.edu.cn

Ruoyun Yao
COLLEGE OF INFORMATION SCIENCE & ELECTRONIC ENGINEERING
Zhejiang University
Hangzhou, China
yaoruoyun@zju.edu.cn

Zhangwan Peng
COLLEGE OF INFORMATION SCIENCE & ELECTRONIC ENGINEERING
Zhejiang University
Hangzhou, China
pengzhangwan@zju.edu.cn

Wanshu Xiong
RESEARCH INSTITUTE OF INTELLIGENT NETWORKS
Zhejiang Lab
Hangzhou, China
xiongwanshu@zhejianglab.com

Jinhua Chen
COLLEGE OF INFORMATION SCIENCE & ELECTRONIC ENGINEERING
Zhejiang University
Hangzhou, China
22231042@zju.edu.cn

Chen Ji
COLLEGE OF INFORMATION SCIENCE & ELECTRONIC ENGINEERING
Zhejiang University
Hangzhou, China
chen.ji@zju.edu.cn

***Abstract*—We propose and demonstrate an integrated scalable optical neural network accelerator based on a dual microring modulator array, which can both implement matrix multiplication and convolutional operation and improve the computation density and speed.**

Keywords—optical neural network accelerator, matrix multiplication, convolutional operation, microring modulators

I. INTRODUCTION

Silicon photonics have become prevalent in accelerating the computation of neural network due to its parallel transmission, large bandwidth and low-power consumption[1-6].The computation of neural network consists of linear operation and nonlinear operation. As it is hard to implement an energy-efficient nonlinear activation using optical devices, several schemes based on optical devices to expedite the linear operation have been proposed. Based on the singular value decomposition, the matrix multiplication can be realized by cascading Mach-Zehnder Interferometer(MZI) array [1]. Nevertheless, the devices required are proportional to the square of the matrix dimension, which makes it is difficult to realize large scale integration. Alternatively, microring modulators (MRMs) can also be used to accelerate the computation because of its small footprint and wavelength-sensitivity [4]. Integrating with an optical frequency comb, large computation speed around 11×10^{12} operation per second is also achievable [7].

However, the optical neural network accelerators are aimed at implementing matrix multiplication or convolutional operations [1,7]. Although the convolutional operation can be converted into the matrix multiplication, there will be additional power consumption for a complex patching scheme [8] and the performance might deteriorate when executing the other operation . Considering the different data feature of the matrix multiplication and the convolutional operation, we propose an optical neural network accelerator based on the dual microring modulator array to perform matrix multiplication and convolutional operation, which helps to improve the computation density and speed.

II. WORKING PRINCIPLE

Our proposal mainly implements the linear operations of the neural network, namely matrix multiplication and the convolutional operation. The convolutional operation mainly contains two inputs, one for the input tensor with a height of H1, width W1 and input channel C1, the other one for the weight tensor, namely kernel in convolutional neural networks which is a four-dimension tensor, input channel C1, output channel C2, height H2, width W2. The output of the convolutional operation is shown in Fig. 1(a), where S denotes the stride that the kernel slides on the input tensor. The convolutional equation is shown in (1). The matrix multiplication is also shown in Fig. 1(b), which is composed of an input vector and a weight matrix.

$$Y_{p,m,n} = \sum_{k=1}^{C1}\sum_{i=1}^{H1}\sum_{j=1}^{W1} X_{k,m+i,n+j}\ W_{p,k,i,j} \qquad (1)$$

In Fig. 2 , our proposal can be divided into five parts, the signal preparation module, the input module, the weighting module, the matrix summation module and the convolutional accumulation module. The signal preparation module offers different kinds of input optical signal with different emission wavelength. The input module is used to map the corresponding input tensor to the electrical waveform of the Mach-Zehnder modulator (MZM). The weighting module is implemented by the dual ring modulator array, which directs the light into north port when performing the convolutional operation or directs the light into the south port while realizing the matrix multiplication.

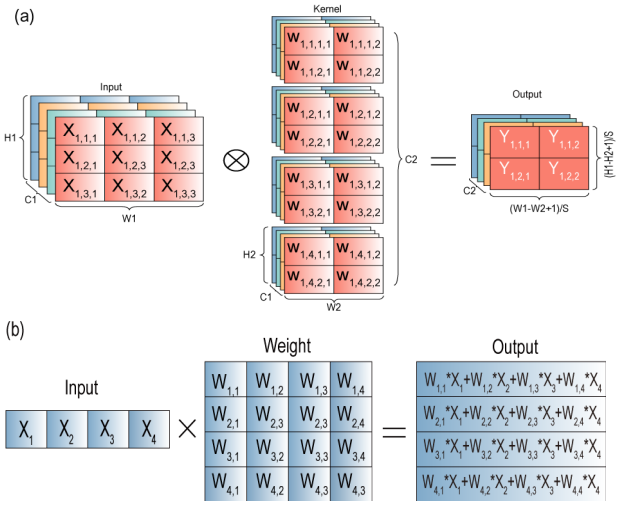

Fig. 1. (a) The convolutional operation wiht a 3-D tensor and a 4-D kernel tensor. (b) The matrix multiplication with a input vector and a weight matrix.

This work was supported in part by the National Natural Science Foundation of China under Grants 61974132, the Zhejiang Lab grant.2020LC0AD01.

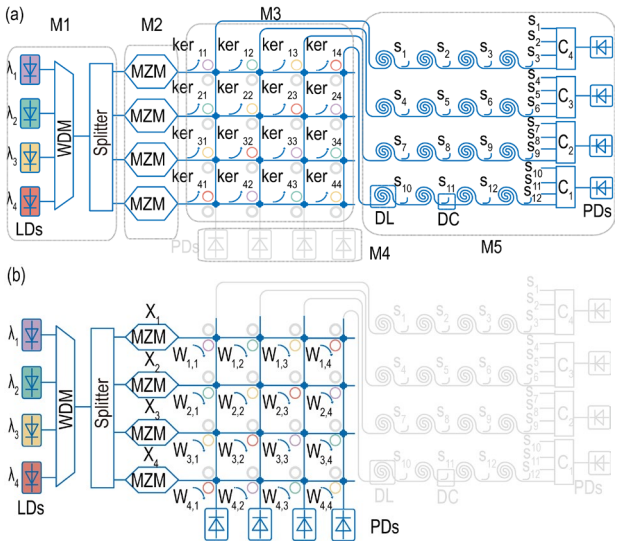

Fig. 2. Our proposed structure that both perform matrix multiplication and the convolutional operation. (a) The structure that performs the convolutional operation. The colored rings represent the upper MRMs, which is around the resonant condition. (b) The structure that performs the matrix multplication. The colored rings represent the lower MRMs, which is around the resonant condition. Devices in the shallower color means that they are not working. LD, laser diode. WDM, wavelength demultiplexing, MZM, mach-zehnder modulator, keri,j denotes the ith input channel, jth output channel of the kernel tensor. Si denotes the ith coupled optical signal after the directional coupler. DC, directional coupler. C_i, the ith combiner. PD, photodiode. M1, input preparation module. M2, input module. M3, weighting module. M4, the matrix summation module. M5, the convolutional accumulation module.

When performing the convolutional operation, the upper microring modulators (MRMs) of the microring module are near the resonant condition to match the corresponding weight while the lower MRMs are off-resonant condition in order not to affect the working condition of the upper MRM as shown in Fig. 2(a). Each element of the input channel of the input tensor is represented by the input waveform of the MZM array. And each row of the MRM array represents the element of the output channel of the kernel tensor while the column of the MRM array express the input channel of the tensor kernel. Therefore, a simple patching scheme is required to transform it into one-dimension data [9-10]. The summation module of the convolutional operation consists of delay lines and directional couplers. These devices are used to couple the signal at the different time to satisfy the convolutional operation.

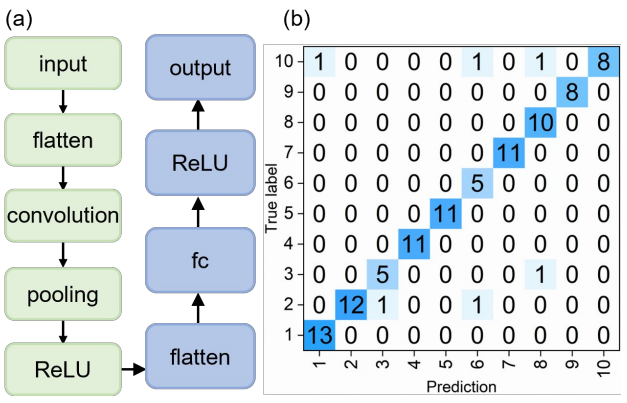

Fig. 3. (a) Our convolutional neural network, consisting of a convolutional layer and a fully connected (fc) layer. (b) Confusion matrix based on our proposal.

In Fig. 2 (b), the matrix multiplication is achievable. The lower MRMs of the dual MRM array are around the resonant condition where the upper MRMs are far away from the resonant wavelength. The input waveform of the MRM array to match the corresponding weight elements. Each row of the dual MRM array represents the row of the weight matrix in Fig. 1(b) and each column of the dual MRM array stands for the column of the weight matrix. And the photodiode array not only converts the optical signal into electrical signal but also perform the summation operation.

III. Simulation

To validate the effectiveness of our proposal, we built a convolutional neural network (CNN) to classify the digital numbers of Modified National Institute of Standard and Technology, where the nonlinear ReLU function defined as $y(x) = max(0, x)$ is performed in the digital computer and the linear operation is finished by our proposal with VPIphotonics. The CNN contains a single convolutional layer and a fully connected layer next to the nonlinear function and are presented in Fig. 3(a). After training, the confusion matrix of classifying the digital number is shown in Fig. 3(b). And the accuracy of classifying the digital number is around 92%, which is comparable to their electronic counterparts.

IV. Conclusion

We offer an optical neural network accelerator for implementing the matrix multiplication and the convolutional operation at the same time based on the dual ring modulator array, which avoids the redundant ports of the microring modulator array and is scalable to realize large scale linear operation. We also simulate the structure in VPIphotonics to verify the validity in performing linear operation in the neural network and it facilitates low-power consumption and high speed operation .

References

[1] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nat. Photonics, vol. 11, no. 7, pp. 441-446. 2017.

[2] S. Xu et al., "Optical coherent dot-product chip for sophisticated deep learning regression," Light: Science & Applications, vol. 10, no. 1, 2021-11-01. 2021.

[3] S. Ohno, R. Tang, K. Toprasertpong, S. Takagi, and M. Takenaka, "Si Microring Resonator Crossbar Array for On-Chip Inference and Training of the Optical Neural Network," ACS Photonics, vol. 9, no. 8, pp. 2614-2622, 2022-08-17. 2022.

[4] A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep.-UK, vol. 7, no. 1, 2017-08-07. 2017.

[5] R. Tang, R. Tanomura, T. Tanemura, and Y. Nakano, "Ten-Port Unitary Optical Processor on a Silicon Photonic Chip," ACS Photonics, vol. 8, no. 7, pp. 2074-2080, 2021-07-21. 2021.

[6] R. Tanomura, R. Tang, S. Ghosh, T. Tanemura, and Y. Nakano, "Robust Integrated Optical Unitary Converter Using Multiport Directional Couplers," J. Lightwave Technol., vol. 38, no. 1, pp. 60-66. 2020.

[7] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks," Nature, vol. 589, no. 7840, pp. 44-51, 2021-01-07. 2021.

[8] Y. Zang, M. Chen, S. Yang, and H. Chen, "Optoelectronic convolutional neural networks based on time-stretch method," Science China Information Sciences, vol. 64, no. 2. 2021.

[9] S. Xu, J. Wang, and W. Zou, "Optical patching scheme for optical convolutional neural networks based on wavelength-division multiplexing and optical delay lines," Opt. Lett., vol. 45, no. 13, pp. 3689-3692, 2020-07-01. 2020.

[10] S. Xu, J. Wang, S. Yi, and W. Zou, "High-order tensor flow processing using integrated photonic circuits," Nat. Commun., vol. 13, no. 1, p. 7970, 2022-12-28. 2022.