Inference and training in deep learning using a symmetric optical crossbar array

Rui Tang[1*], Shuhei Ohno[1], Ken Tanizawa[2], Kazuhiro Ikeda[3], Makoto Okano[3], Kasidit Toprasertpong[1], Shinichi Takagi[1], and Mitsuru Takenaka[1]

[1]*Department of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo 113-8656, Japan*
[2]*Quantum ICT Research Institute, Tamagawa University, Tokyo 194-8610, Japan*
[3]*National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8568, Japan*
[*]*ruitang@mosfet.t.u-tokyo.ac.jp*

Abstract: We propose and demonstrate a symmetric optical crossbar array based on microring resonators (MRRs) to accelerate both the inference and training in deep learning, experimentally achieving a 93.3% classification accuracy in an inference task. © 2024 The Author(s)

1. Introduction

The further development of deep learning systems requires application-specific processors that can simultaneously improve the computation speed and reduce the power consumption. Photonic processors, capable of performing on-chip matrix multiplications, are promising candidates due to the inherent parallel nature of light [1]. So far, various coherent and non-coherent schemes for realizing photonic matrix multipliers have been proposed and demonstrated [2–6]. While the non-coherent schemes do not support complex-valued matrices, they can significantly simplify the control complexity and are usually sufficient for many practical applications. The microring resonator (MRR) crossbar array is a promising non-coherent scheme because of its compact structure and the potential to accelerate both the inference and training in deep learning [6]. The MRR crossbar array can simultaneously implement an arbitrary non-negative real matrix and its transpose matrix without the need to reconfigure the MRRs. This unique property allows for direct on-chip backpropagation, which is essential for in situ training [7]. Previously, we have proposed and demonstrated a 4×4 silicon photonic MRR crossbar array [6]. However, because of its asymmetric structure, the insertion losses are not equal for all optical paths, which resulted in large errors in the matrices experimentally realized on the device.

In this work, we propose and demonstrate a novel MRR crossbar array with a symmetric structure. All optical paths in the new structure have the same lengths and insertion losses, thereby significantly reducing errors caused by imbalanced insertion loss in the realized matrices. Using a 4×4 MRR crossbar array fabricated on a Si-on-insulator (SOI) platform, we experimentally demonstrate the inference and further simulate the on-chip training in a 3-layer neural network for classifying Iris flowers.

2. Principle

The proposed symmetric MRR crossbar array is schematically shown in Fig. 1. For $N\times N$ matrices (N = 4 in Fig. 1), this approach uses N^2 MRRs to represent the matrix and $2N$ Mach-Zehnder interferometers (MZIs) to generate two input vectors: $\boldsymbol{x}$ and $\boldsymbol{\sigma}$ (referred to as the forward and backward signals, respectively). The forward signal $\boldsymbol{x}$ represents the output from the previous layer and is multiplied with the weight matrix $\mathbf{W}$, the backward signal $\boldsymbol{\sigma}$ represents the

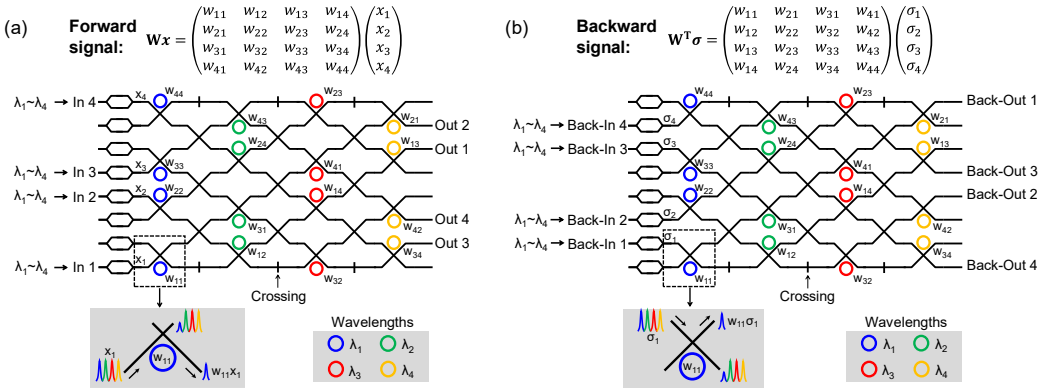

Fig. 1. Proposed optical crossbar array for matrix-vector multiplications. The matrix and vector are generated by microring resonators (MRRs) and Mach-Zehnder interferometers (MZIs), respectively. (a) By injecting a forward signal $\boldsymbol{x}$, the crossbar array performs the multiplication between $\mathbf{W}$ and $\boldsymbol{x}$. (b) By injecting a backward signal $\boldsymbol{\sigma}$, the crossbar array performs the multiplication between $\mathbf{W}^{\mathbf{T}}$ (the transpose of $\mathbf{W}$) and $\boldsymbol{\sigma}$.

error signal backpropagated from the next layer and is multiplied with $\mathbf{W}^{\mathbf{T}}$ without reconfiguring the MRRs. Note that $\boldsymbol{x}$ and $\boldsymbol{\sigma}$ are not injected into the device at the same time. For each direction, N wavelengths are injected into each input port simultaneously. Each MRR is tuned to couple with one wavelength and the associated matrix element is represented by the transmittance of optical power at the drop port. At each output port, N optical signals coupled through N different MRRs are multiplexed into the same waveguide and detected by an on-chip or external photodetector. Therefore, the multiplication and accumulation operations are performed at the MRRs and the photodetectors, respectively. In this new structure, all optical paths by design have the same lengths and insertion losses for the forward/backward signal, respectively.

3. Fabricated device

A 4×4 MRR crossbar array based on the proposed structure is fabricated on the SOI platform, as shown in Fig. 2(a). The propagation loss of the single-mode silicon waveguide is 1.3 dB/cm. The radii of all MRRs are 20 μm, corresponding to a free spectral range of 4.4 nm at a 1550 nm wavelength. For proof-of-concept demonstrations, thermo-optic phase shifters are used to tune the MZIs and MRRs. In future, ultralow-power electro-optic phase shifters can be used to significantly lower the power consumption [8]. The fabricated chip is wire-bonded and then packaged with an optical fiber array for stable characterizations. The MZIs typically have extinction ratios greater than 40 dB, as shown in Fig. 2(b). The extinction ratios measured at the drop ports of the MRRs are typically greater than 25 dB, as shown in Fig. 2(c).

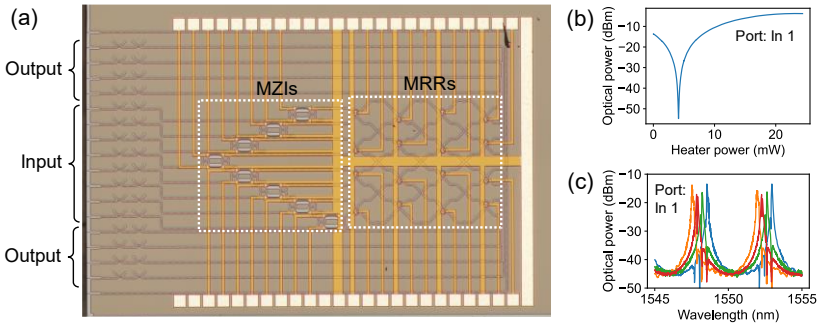

Fig. 2. (a) A 4×4 MRR crossbar array fabricated on the SOI platform. (b) The MZI in port "In 1" is characterized. (c) Transmission spectra measured at the drop ports of four MRRs by sweeping the wavelength of the light injected into port "In 1".

4. Matrix implementation

Four different wavelengths (λ_1=1549.02, λ_2=1549.77, λ_3=1550.52, λ_4=1551.27 nm) are combined externally and then injected into the input ports. The optical power at each output port is measured by a multi-channel optical power meter. By tuning the MRRs via thermo-optic phase shifters, we implement various matrices for the forward and backward directions, as shown in Fig. 3. In contrast to the result in our previous work [6], where the matrices for the forward and backward directions exhibited significant differences, here the desired matrices successfully realized for both directions with negligible distinctions. Undesired noise signals in these matrices are suppressed to the level of approximately -15 dB.

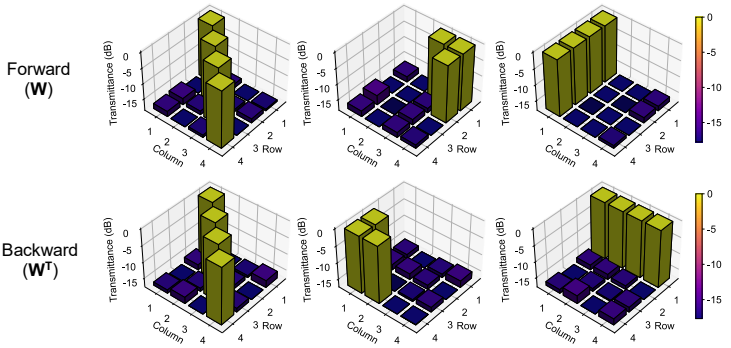

Fig. 3. Experimental implementations of various matrices for forward and backward signals.

5. Inference and training

We construct a 3-layer neural network for classifying Iris flowers, as shown in Fig. 4(a). The network takes a 4-element vector as the input and generates a 3-element vector as the output. The sigmoid function is used as the nonlinear activation function. The data set consists of 150 samples in total: 105 samples are used for training and the remaining 45 samples are used for test. We first trained this network on a computer using the stochastic gradient descent algorithm and achieved 97.8% accuracy on computer using the test data. Then, we use the MRR crossbar array to perform the matrix-vector multiplications and the computer to perform the nonlinear activation functions. Only forward signals are needed in the inference task. For the same test data, a high classification accuracy of 93.3% is obtained, as shown in Fig. 4(b). However, relatively large fluctuations of the output optical powers have been observed, which may be caused by the insufficiently stable temperature control of MRRs. In our experiments, we performed time-averaging measurements to reduce the noise. This issue may be solved by using electro-optic phase shifters that do not generate heat or applying feedback controls on current phase shifters.

Due to the power fluctuation, on-chip training is not performed directly since it will require a significant amount of measurement time. Instead, we characterized all the MZIs and MRRs and created look-up tables that map the settings of MZIs and MRRs to the output power at each output port. Using these look-up tables, we simulate the on-chip training of the 3-layer neural network. Here, both the forward and backward signals are needed. The multiplications between matrix and vector elements are performed by fetching data from the look-up tables, the additions and nonlinear activation functions are performed by the computer. After the training, we perform the same inference tasks using the test data again. The results are shown in Fig. 4(c). While the inference results using the computer and the crossbar array are not exactly the same, relatively high accuracies of 91.1% are obtained in both cases.

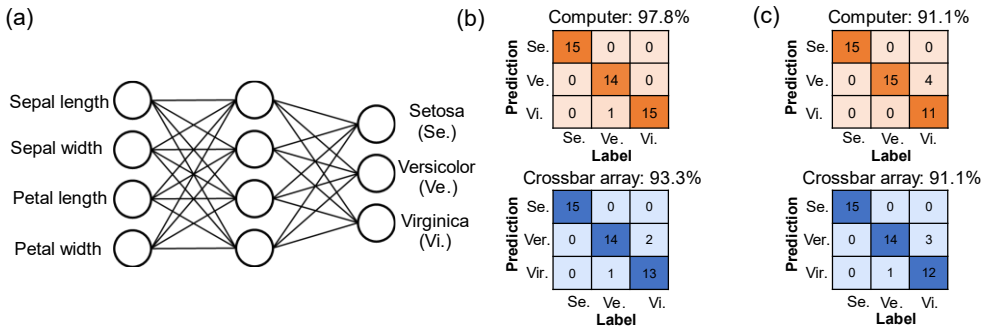

Fig. 4. (a) A 3-layer neural network for classifying Iris flowers. (b) Inference results after the neural network is trained on a computer. A high accuracy of 93.3% is obtained using the MRR crossbar array. (c) Inference results after a simulated on-chip training. An accuracy of 91.1% is obtained using the MRR crossbar array.

6. Conclusion

We have proposed and demonstrated a novel MRR crossbar array for accelerating both the inference and training in deep learning. Using a 4×4 MRR crossbar array to perform matrix-vector multiplications in a 3-layer neural network pre-trained on a computer, we obtained a high classification accuracy of 93.3% in the inference task. After a simulated on-chip training, we obtained an accuracy of 91.1% in the same inference task.

Acknowledgement

This work was partly supported by JST CREST (JPMJCR2004) and JSPS KAKENHI (22K14298).

References

[1] 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**, 441–446 (2017).

[2] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, "Optimal design for universal multiport interferometers," Optica **3**, 1460 (2016).

[3] G. Giamougiannis, A. Tsakyridis, Y. Ma, A. Totović, M. Moralis-Pegios, D. Lazovsky, and N. Pleros, "A Coherent Photonic Crossbar for Scalable Universal Linear Optics," J. Light. Technol. **41**, 2425–2442 (2023).

[4] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**, 7430 (2017).

[5] 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**, 52–58 (2021).

[6] 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 **9**, 2614–2622 (2022).

[7] S. Pai, Z. Sun, T. W. Hughes, T. Park, B. Bartlett, I. A. D. Williamson, M. Minkov, M. Milanizadeh, N. Abebe, F. Morichetti, A. Melloni, S. Fan, O. Solgaard, and D. A. B. Miller, "Experimentally realized in situ backpropagation for deep learning in photonic neural networks," Science **380**, 398–404 (2023).

[8] S. Ohno, Q. Li, N. Sekine, H. Tang, S. Monfray, F. Boeuf, K. Toprasertpong, S. Takagi, and M. Takenaka, "Si microring resonator optical switch based on optical phase shifter with ultrathin-InP/Si hybrid metal-oxide-semiconductor capacitor," Opt. Express **29**, 18502 (2021).