Integrated 16×16 Photonic Analog Vector-Matrix Multiplier with Task-Specific Tuning after Deterministic Calibration

Kohei Ikeda[1,2], Shota Kita[1,2], Kengo Nozaki[1,2], Kenta Takata[1,2], Kazuo Aoyama[3], Keijiro Suzuki[4], Yuriko Maegami[4], Morifumi Ohno[4], Guanwei Cong[4], Noritsugu Yamamoto[4], Koji Yamada[4], Akihiko Shinya[1,2], Hiroshi Sawada[3], Masaya Notomi[1,2]

[1] NTT Nanophotonics Center, Nippon Telegraph and Telephone Corp., kohei.ikeda@ntt.com
[2] NTT Basic Research Labs., Nippon Telegraph and Telephone Corp.
[3] NTT Communication Science Labs., Nippon Telegraph and Telephone Corp.
[4] National Institute of Advanced Industrial Science and Technology (AIST)

Abstract *We demonstrate a silicon-photonics 16×16 Clements-type vector-matrix multiplier, performing MNIST classification. The imperfection in the largest-scale photonic analog circuit was corrected using our proposed novel machine learning-based tuning method with a high fidelity of 0.904, significantly improving the experimental MNIST classification accuracy.*

Introduction

Over the past several years, the role of photonics has changed rapidly because of the rise of silicon photonics. Now, photonics is considered not only for optical communication systems but for computing platforms, so the boundary between electronics and photonics needs to be clarified. Although the primary focus on computing has been achieved through CMOS processors, the great potential of photonics has been shown for computing, especially in artificial neural networks (ANNs) [1-4]. Photonic interferometers efficiently implement analog linear algebra operations, which is dominantly used for deep learning [5], at the speed of light. Therefore, photonic analog vector-matrix multipliers (PVMMs) can be significant building blocks for energy-saving and low-latency ANN architectures.

To counter the exponential growth in demand for deep learning, scaling to a large-scale network is a crucial issue to be addressed [6, 7]. However, it is challenging to demonstrate the operation of large-scale photonic analog circuits, including PVMMs, because a more effective calibration method still needs to be used to overcome practical imperfections (fabrication errors and crosstalk originating from both thermal and electric effects). Although several on-chip learning algorithms are being developed to overcome this issue [8, 9], they are challenging to scale up to a more extensive network and require considerable time to gain convergence. Recently, physical implementation of three-layer deep neural networks including nonlinear activation functions has been demonstrated [10]. Despite this success, previously reported demonstrations of PVMMs based on silicon photonics that perform universal unitary transformation are limited in their circuit size (e.g., 4×4 [1, 11, 12], 6×6 [10], and 8×8 [13]) and practical applications. We note that MNIST classification [14] is demonstrated in Ref. [12] and Ref. [13]; however, they limit the MNIST digit data to be inferred from “0” to “3” [12] or strongly rely on digital post-processing [13] due to the constraints of PVMM scaling. 10-digit MNIST classification has not been implemented using only a Mach-Zehnder interferometer (MZI)-based PVMM.

This paper demonstrates a 16×16 Clements-type [15] PVMM operated with an efficient tuning method based on machine learning. MNIST classification is implemented using only a single PVMM that directly classifies the input data into 10 optical output ports without any digital post-processing. We obtained a significant improvement in the classification accuracy. Although our matrix processing comprises 240 tunable phase shifters to be specifically tuned, the largest size of silicon photonics-based PVMM demonstrated to date, our presented tuning method could successfully suppress the crosstalk effect. This method opens the door to realizing large-scale networks, suggesting the possibility of relaxing the scaling limit of PVMMs.

Design and packaging

Figure 1 (a) shows an optical micrograph of the Clements-type 16×16 PVMM fabricated through a 300-mm wafer process with a CMOS pilot line. Input light coupled into the circuit is first split into 16 channels. Each channel has a dual-arm MZI modulator as an input unit, which encodes a complex value onto the optical field. Complexed valued input has a two times better representation ability than real-valued input. Then, the input signals propagate through a matrix processing unit comprising 120 MZIs (240 phase shifters), which implement a 16×16 unitary transformation operation. The insets show the

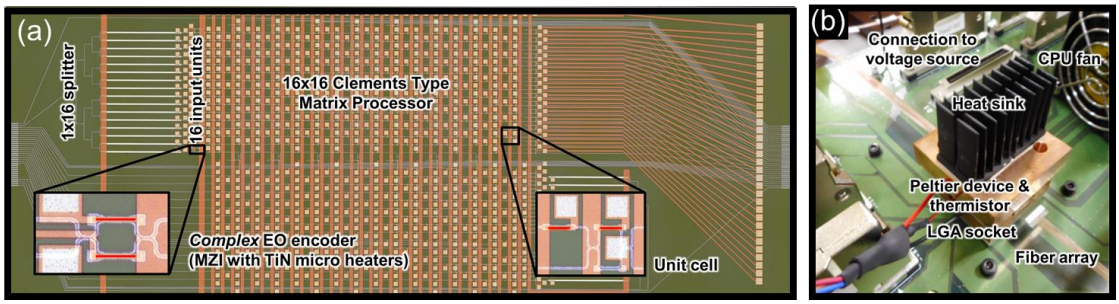

Fig. 1: (a) Optical microscope image of photonic analog vector-matrix multiplier (top half), consisting of input unit and Clements-type 16×16 configuration. Insets represent unit MZI of input and matrix units, respectively. (b) Picture of whole packaging, including printed circuit board with multiple connectors for large-scale electric input system. Silicon die sample was bonded with ceramic interposer.

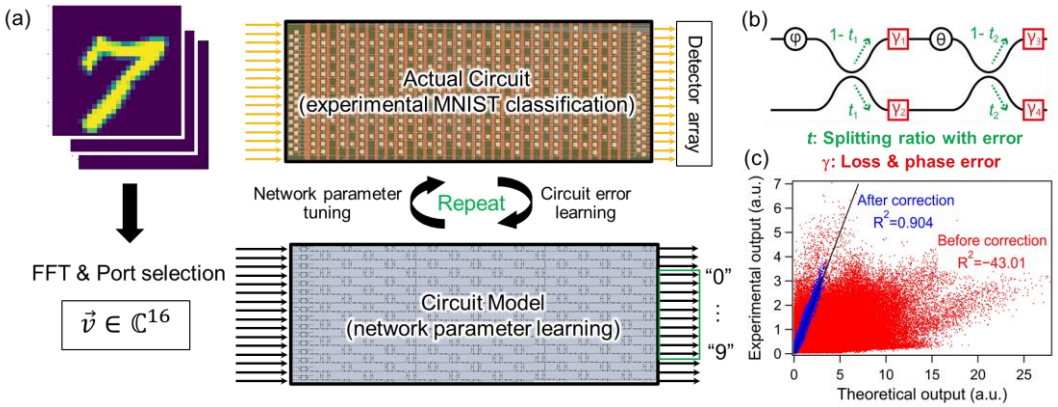

Fig. 2: (a) Demonstration of 16×16 photonic analog vector-matrix multiplier for implementing MNIST classification task. Each MNIST image is reshaped into input vector with 16 elements through pre-processing. Network parameters are gradually tuned toward optimized ones in accordance with offline training. During tuning process, circuit model is repeatedly updated on basis of circuit error learning using relationship between compressed input vectors and output powers from actual circuit. (b) Developed model for each MZI in matrix unit. Six circuit errors (two "t" and four "γ") per each MZI are inferred using machine learning. (c) Correlation between simulated outputs from theoretical model and experimental outputs from actual sample.

unit cells of the input and matrix units. Each MZI comprises two directional couplers (DCs) and two phase shifters based on titanium nitride heaters, which unintentionally cause thermal crosstalk. We developed a packaging technique for a relatively large-scale PVMM, shown in Figure 1(b). A pair of fiber arrays are attached to chip facets. For a proper interface with all phase shifters connected to electric voltage sources, we bonded a ceramic interposer on the chip. We put it in an LGA socket for electric contacts and onto a printed circuit board with many electric connectors for off-chip input modules. We put a temperature control system on the LGA socket for temperature stabilization. Note that we also have electric crosstalk because of the voltage control even with wiring resistance. Therefore, the crosstalk behavior could be complicated and challenging to model in this case.

Task-specific tuning method

Figure 2(a) shows the procedure of the task-specific tuning for MNIST classification. We executed pre-processing, including max-pooling and FFT, to convert original MNIST images (28×28) to 16-dimensional complex input vectors. The input vectors consisted of selected-frequency FFT features from each image. To obtain the optimized network parameters for MNIST classification, we performed offline training using 10,000 MNIST images for a circuit model built on a digital platform. In the model, we assumed circuit errors for each MZI (insertion loss, phase delay, and splitting ratio error) [10], shown in Figure 2(b). Since we cannot directly measure most prior information on the errors in the actual device, the primary model starts from the ideal case (i.e., loss-less, phase error free, and equal splitting ratio). We assigned labels from "0" to "9" at the output port from the 4th to 13th, respectively. The output labels were determined for each input image corresponding to the output port with the highest power among those ten ports. After the training, we mapped the optimized network parameters to the actual circuit where all phase shifters in the MZIs were previously calibrated. Then, CW laser light at 1526 nm was input to the circuit, and 16 outputs were measured. On the basis of the relationship between the input vectors and the output vectors for the 10,000 MNIST images, we updated the model using machine learning, correcting for

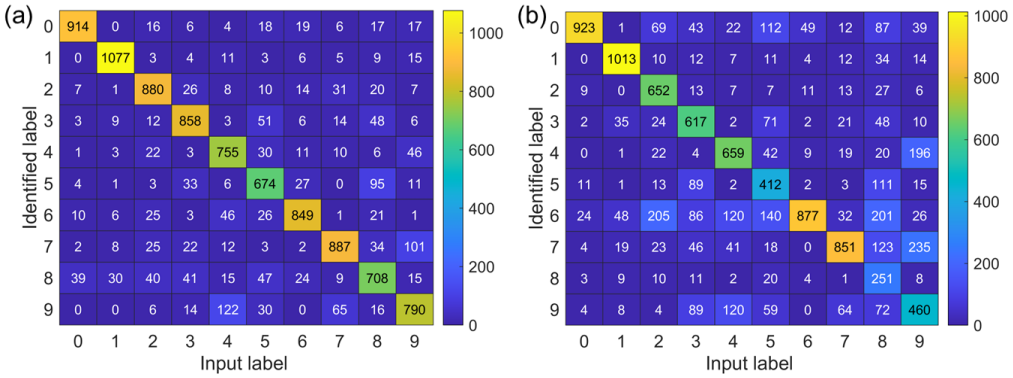

Fig. 3: MNIST classification task for 10,000 MNIST images. (a) Confusion matrix for theoretical model, with accuracy of 83.9%. (b) Confusion matrix for experiment implementation, with accuracy of 67.2%.

circuit errors from the ideal case. Figure 2(c) shows the power relation between the experimental outputs and the theoretical outputs from the updated model. The calculated fidelity (R-squared value) was 0.904, improved from −43.01 (before the update). Since the circuit errors inferred here absorb the crosstalk effect, individual crosstalk correction is not necessary. On the other hand, the distribution of the crosstalk effect depends on the implemented matrix. Thus, when we tune the network parameters for any sort of task, we select a relatively small epoch and learning rate to prevent a drastic change from the initially implemented matrix. Instead of approaching the final matrix once, we repeated the above procedure (circuit error learning, network parameter learning, and experimental MNIST classification) to gradually tune the network parameter distribution.

MNIST classification task

The obtained accuracy for the MNIST classification task was improved gradually along with the iterations. During the network parameter learning, the learning rate was set at 0.001, while the epoch was set at 10-20. Figure 3(a) shows the confusion matrix for the theoretical model after 11 iterations, showing a theoretical accuracy of 83.9%. On the basis of the network parameters obtained in this model, we achieved an experimental accuracy of 67.2%, as shown in Figure 3(b). The gap between theoretical and experimental accuracies was caused by the imperfection of the modeling (R^2 = 0.904) and the residual crosstalk effect. We note that when the network parameters were tuned drastically once, the accuracy was much lower (31.3 %), as plotted in Figure 4, showing the advantage of our proposed method. Finally, we tested the wavelength dependence of the PVMM for MNIST classification, as shown in Figure 4. Each plot was obtained from the classification result of 1,000 MNIST images at each wavelength. Although a nearly linear dependence was observed, a high classification accuracy (> 60%) was maintained within ±5 nm away from the calibrated wavelength. This means that parallel processing using several wavelengths would be feasible, which would enable throughputs to be improved.

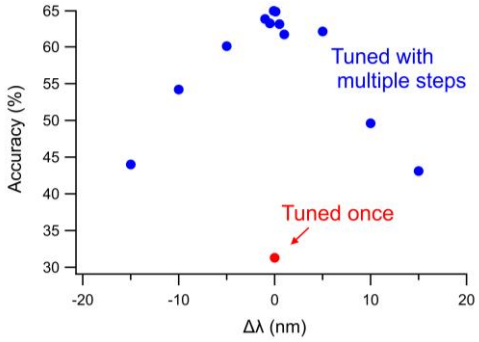

Fig. 4: Wavelength dependence of accuracy for MNIST classification task. Each accuracy (blue plot) was obtained from results of 1,000 images at shifted wavelength from calibration wavelength of 1526 nm. As reference, accuracy obtained when network parameters were tuned once is shown (red plot).

Conclusion

We experimentally demonstrated a 16×16 Clements-type PVMM and achieved the first experimental MNIST classification that defines the classification results directly corresponding to the optical output ports. The obtained accuracy of 67.2% is mainly limited by the compression of input images. Our proposed machine learning tuning method enables quick and accurate deployment of photonic neural network parameters into imperfect circuits, realizing the implementation of a large-scale PVMM. Moreover, we confirmed a low wavelength dependence for MNIST classification for multiple wavelengths. These results provide one avenue for both large-scale and parallel processing.

Acknowledgements

This work was supported by the JST, CREST Grand Number JPMJCR21C3, Japan.

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. Photon., vol. 11, no. 7, pp. 441 – 446, 2017, DOI: 10.1038/NPHOTON.2017.93.

[2] H.-T. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, “Neuromorphic photonic integrated circuits,” IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 6, p. 6101715, 2018, DOI: 10.1109/JSTQE.2018.2840448.

[3] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, “All-optical machine learning using diffractive deep neural networks,” Science, vol. 361, no. 6406, pp. 1004–1008, 2018, DOI: 10.1126/science.aat8084.

[4] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, “Large-scale optical neural networks based on photoelectric multiplication,” Phys. Rev. X, vol. 9, no. 2, p. 021032, 2019, DOI: 10.1103/PhysRevX.9.021032.

[5] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: A tutorial and survey,” Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, 2017, DOI: 10.1109/JPROC.2017.2761740.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015, DOI: 10.1038/nature14539.

[7] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The computational limits of deep learning,” http://arxiv.org/abs/2007.05558, 2020.

[8] H. Zhang, J. Thompson, M. Gu, X. D. Jiang, H. Cai, P. Y. Liu, Y. Shi, Y. Zhang, M. F. Karim, G. Q. Lo, X. Luo, B. Dong, L. C. Kwek, and A Q. Liu, “Efficient On-Chip Training of Optical Neural Networks Using Genetic Algorithm,” ACS Photonics, vol. 8, no. 6, pp. 1662–1672, 2021, DOI: 10.1021/acsphotonics.1c00035.

[9] G. Cong, N. Yamamoto, T. Inoue, Y. Maegami, M. Ohno, S. Kita, S. Namiki, and K. Yamada, “On-chip bacterial foraging training in silicon photonic circuits for projection-enabled nonlinear classification,” Nat. Commun., vol. 13, no. 1, p. 3261, 2022, DOI: 10.1038/s41467-022-30906-3.

[10] S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, “Single chip photonic deep neural network with accelerated training,” https://arxiv.org/abs/2208.01623, 2022.

[11] A. Ribeiro, A. Ruocco, L. Vanacker, and W. Bogaerts, “Demonstration of a 4 × 4-port universal linear circuit,” Optica, vol. 3, no. 12, pp. 1348-1357, 2016, DOI: 10.1364/OPTICA.3.001348.

[12] F. Shokraneh, S. Geoffroy-Gagnon, M. S. Nezami, O. Liboiron-Ladouceur, “A Single Layer Neural Network Implemented by a 4 x 4 MZI-Based Optical Processor,” IEEE Photon. Journal, vol. 11, no. 6, p. 4501612, 2019, DOI: 10.1109/JPHOT.2019.2952562.

[13] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, “An optical neural chip for implementing complex-valued neural network,” Nat. Commun., vol. 12, no. 1, p. 457, 2021, DOI: 10.1038/s41467-020-20719-7.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998, DOI: 10.1109/5.726791.

[15] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, “Optimal design for universal multiport interferometers,” Optica, vol. 3, no. 12, pp. 1460-1465, 2016, DOI: 10.1364/OPTICA.3.001460.