Optoelectronic Implementation of Compact and Power-efficient Recurrent Neural Networks

Taisei Ichikawa†, Yutaka Masuda†, Tohru Ishihara†, Akihiko Shinya‡, and Masaya Notomi‡
† Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan
‡ NTT Nanophotonics Center / Basic Research Laboratories, 3-1 Morinosato Wakamiya, Atsugi, Japan

***Abstract*—Optoelectronic implementation of artificial neural networks (ANNs) has been attracting significant attention due to its potential for low-power computation at the speed of light. Among the ANNs, adopting recurrent neural network (RNN) is a promising solution since it provides sufficient inference accuracy with a more compact structure than other ANNs. This paper proposes a novel optoelectronic architecture of RNN. The key idea is to implement the vector-matrix multiplication optically to exploit the speed of light and implement the activation and feedback electronically to exploit the controllability of electronics. The electronics part is composed of an electrical feedback circuit with a dynamic latch to synchronize the recurrent loops with a clock signal. Using a commercial optoelectronic circuit simulator, we confirm the correct behavior of the optoelectronic RNN. Experimental results obtained using TensorFlow show that the proposed optoelectronic RNN achieves more than 98% inference accuracy in image classification with a minimal footprint without sacrificing low-power and high-speed nature of light.**

***Index Terms*—optical computing, neuromorphic computing, recurrent neural network**

I. Introduction

Today's highly developed information society would not have been realized without optical communication and CMOS LSI technologies. Traditionally, optical communication technology has been developed for long-distance communication. However, recent innovations in nanophotonics have moved it to shorter distances and are now migrated onto silicon chips as optical networks-on-a-chip. Concurrently, the rapid progress of research on neural network (NN) has stimulated research on optical neural network (ONN) [1]–[9]. Whereas multi-layer perceptrons (MLPs) generally provide higher expressiveness by exploiting the deep-layer structure, such a structure requires an extremely large footprint. On the other hand, recurrent neural network (RNN) has a loop structure which results in a smaller implementation area than MLP. Especially for image classification, RNN can achieve outstanding inference accuracy with a much smaller area than the other DNN counterparts. For example, Visin et al. proposed an efficient RNN model [10] that replaced the fully connected layers with RNN that swiped through the image, attaining accuracy of 99.55% with MNIST dataset, a widely used image database of handwritten digits.

To incorporate the ultra-fast nature of light into the RNN model, several previous works [11], [12] proposed optical implementation of RNN. Although those RNNs perform the inference processing at the speed of light, they involve a large footprint since all neurons and edges in the model are based on a large photonics device called Mach-Zehnder Interferometer (MZI), which limits the scalability of the ONN. Tait et al. proposed a more compact optoelectronic RNN [1] based on micro-ring modulator array. However, this RNN in turn has a structure with higher delays in optical-to-electrical signal conversion, limiting the overall performance of the RNN.

In this paper, we propose for the first time an optoelectronic RNN architecture that performs the inference processing at the speed of light without sacrificing the compact nature of RNN. The key idea of the proposed architecture is to implement the vector-matrix multiplication part optically and implement the activation and feedback part electronically. Thanks to the electro-optic hybrid implementation, the proposed architecture fully takes advantage of the speed of light and the controllability of the electronics. Experimental results obtained using TensorFlow show that the proposed RNN achieves inference accuracy of more than 98% in MNIST with a minimal footprint, without sacrificing optics' low-power consumption and high-speed nature. Using a commercial optoelectronic circuit simulator, we have also verified that the optoelectronic RNN works correctly. The rest of the paper is organized as follows. Section II summarizes several previous works on the ONNs. Section III proposes a detailed architecture of our optoelectronic RNN. Section IV shows experimental results obtained with the optoelectronic circuit simulator and TensorFlow. Section V concludes this paper.

II. Related Works

A. Coherent RNN with Mach-Zehnder Interferometer Array

In [5], a fully optical neural network (ONN) architecture is presented. The core part of the ONN architecture is a matrix multiplication unit composed of a reconfigurable Mach-Zehnder Interferometer (MZI) array. Although this ONN architecture is much faster than the electronic counterparts, one significant drawback in the architecture is high photonic component utilization and area cost. The area for the implementation is quadratically proportional to the number of neurons in the network. In [6], a more compact ONN architecture based on Fast Fourier Transform (FFT) is proposed. However, the area required for the implementation is massive as the number of MZIs required is still quadratic to the number of neurons.

All-optical recurrent neural network (RNN) based on the MZI array described above is proposed in [11] [12]. This RNN architecture performs low-power and high-speed sequence processing using the MZI array and looped waveguides. Although

the architecture is more compact than ONN proposed in [5] [6] thanks to the recurrent structure, the circuit is inherently massive as it is based on the MZI array presented in [5].

B. RNN based on Incoherent Accumulation

An optical circuit structure for vector-matrix multiplication (VMM) based on wavelength division multiplexing (WDM) is proposed in [1]. The incoming WDM carrier waves are weighted by modulators called microring (MRR) weight banks. The weighted optical signals are then summed as photocurrent by a photodetector. Although this architecture is very area-efficient, it has an oscillation issue. If WDM waves are given to a photodetector, an oscillation in photocurrent occurs. This is known as a beat note. One simple approach to eliminate the beat note is low-pass filtering. However, this prevents VMM from exploiting the ultra-high-speed nature of light since the time constant of the low-pass filter is more than two orders of magnitude bigger than that of the fully optical operations. In [2]–[4], [7], [13], [14], similar optical VMM circuits are proposed. These circuits are very small since they are based on photodetector-based accumulation. However, they also have the beat note issue in the photodetector, which prevents the light-speed operation of optical VMM.

The main focus of [1] is on the photonic implementation of a continuous time RNN (CT-RNN). Although this CT-RNN is much more compact than the optical RNNs proposed in [11] [12], it requires low-pass filtered amplifiers to eliminate the beat note, limiting the overall performance of the RNN.

C. ONN based on All-Optical Coherent VMM

A fully coherent VMM based on WDM is proposed in [8], [9]. This VMM employs an optical accumulation circuit using an optical power combiner to solve the beat note issue incurred in the photodetector-based accumulation presented in the previous subsection. The schematic of an ONN is depicted in Fig. 1. Each layer of the ONN is composed of the MRR weight bank [1], parallelized multipliers based on Mach Zehnder Modulator (MZM), an accumulator based on an optical combiner [15], and an optoelectronic activation circuit based on [16] which directly implements a nonlinear function. The incoming WDM carrier waves are first weighted by the MRR weight banks and then the weighted WDM signals are passed to the MZMs for the multiplication. Once the outputs of the MZMs are given to the power combiner as WDM signals, optical accumulations are performed in parallel. Optical waves with the same wavelength are accumulated in the power combiner. This accumulation is performed for every different wavelength in parallel. Since the coherent light is used throughout the multiply and accumulation operation, the light-speed VMM calculation is possible in this architecture. Finally, the accumulated values are extracted by micro-ring resonators and given to the activation circuits separately. The activation circuit is operated in an electronics domain after optical-to-electrical (O-E) conversion by a photodetector [16]. Since this architecture employs the MRR weight bank and the optical combiner for the VMM calculation, it takes advantage of the architectures proposed in [5] and [1], respectively.

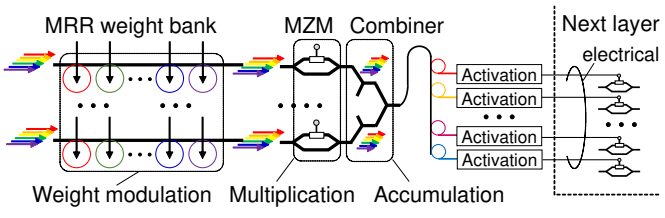

Fig. 1. Optical neural network based on coherent VMM [9].

III. Proposed Optoelectronic RNN Architecture

This section proposes an RNN architecture based on the ONN presented in Fig. 1. To the best of our knowledge, this is the first proposal of an optoelectronic RNN based on this compact and ultra-fast ONN architecture.

A. Electrical Feedback Circuit with Dynamic Latches in RNN

Most of the up-to-date research activities on optical neural networks (ONNs) are focused on a feed-forward type of architecture, such as multi-layer perceptrons (MLPs) and convolutional neural networks (CNNs). The main reason for this trend is that there are no memory devices to store analog optical signals. If a digital optical flip-flop such as proposed in [17] is used instead of storing analog activation results, analog-to-digital conversion (ADC) is needed, which consumes a large amount of energy and area in the recurrent feedback circuitry.

As a remedy to this issue, this section proposes an electrical analog memory based on a dynamic latch. The schematic of the dynamic latch is depicted in the upper left of Fig. 2. Like DRAM cells, the dynamic latch stores electrical charge in the parasitic capacitance on the input of a Mach-Zehnder modulator (MZM). When the transmission gate is ON, the charge moves to the parasitic capacitance, which corresponds to the write operation. Once the transmission gate is turned OFF, the charge is kept stored in the capacitance, which corresponds to the hold operation. This dynamic latch enables energy-efficient feedback operation in the optoelectronic RNN.

B. Optoelectronic RNN Architecture

Figure 2 shows the overall structure of optoelectronic RNN. As presented in the upper part of Fig. 2, the activation circuits and the following dynamic latches are implemented in the electronics domain. The optical VMM is implemented in the optics domain, as shown at the bottom of Fig. 2. For given x_0 and x_1 as electrical inputs, and v_0 and v_1 as electrical feedback signals, the optics domain calculates two VMMs which correspond to the arguments of (1) and (2) where the f represents the activation function. The b_0 and b_1 represent the biases for the activation circuits, respectively.

$$V_0 = f(v_0 W_{0,0} + v_1 W_{0,1} + x_0 W_{0,2} + x_1 W_{0,3} + b_0). \quad (1)$$

$$V_1 = f(v_0 W_{1,0} + v_1 W_{1,1} + x_0 W_{1,2} + x_1 W_{1,3} + b_1). \quad (2)$$

The optical signals with the red and green in Fig. 2 correspond to (1) and (2), respectively. The propagation delay of optical waves traveling through optical waveguides is proportional to

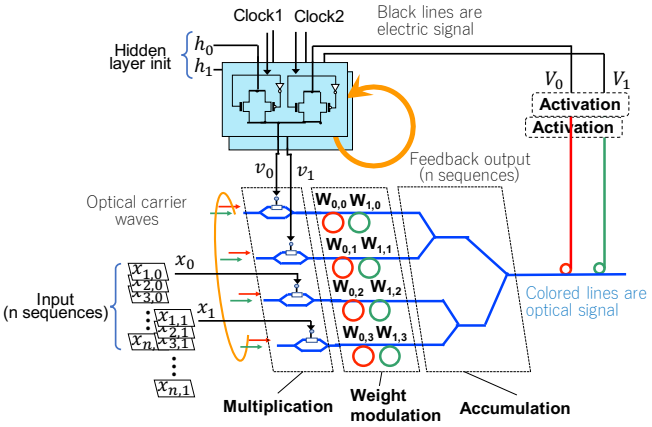

Fig. 2. Proposed optoelectronic RNN architecture with dynamic latches.

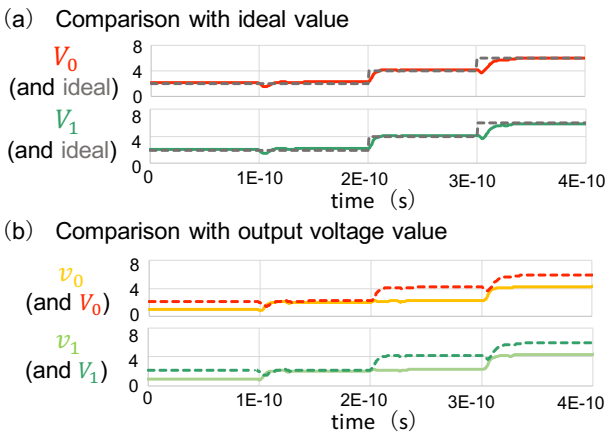

Fig. 3. Optoelectronic circuit simulation results. (a) Comparison with expected values. Solid lines correspond to V_0 and V_1, and doted lines are expected results. (b) Comparison between the output in the previous cycle and current input.

the total path length. The total length of the path starting from the MZM to the input of the activation circuit can be designed to be less than 1 mm. Since the speed of light traveling in the semiconductor waveguides is about 100μm/ps, the delay of the optical VMM is less than 10 ps. The worst case delay of the activation circuit based on the O-E converter proposed in [16] is about 30 ps. According to circuit simulation for the dynamic latch designed with 16 nm CMOS process technology [18], the write delay of the latch is less than 10 ps. As a result, the recurrent cycle time can be less than 50 ps, corresponding to the recurrent clock frequency of 20 GHz.

The dynamic latch explained in section III-A can be accurately functioned as a temporal analog memory in the feedback circuit of the optoelectronic RNN. Since the leakage current drawn through the transmission gate of the dynamic latch is an order of picoampere, the amount of charge lost by the leakage current within several tens of picoseconds is an order of 0.01 attocoulomb. This charge loss corresponds to less than 0.01% of the value stored in the dynamic latch since the parasitic capacitance of the MZMs is around 1.0 femtofarad. Therefore, the dynamic latch accurately works as a temporal analog memory in the optoelectronic RNN which operates with a clock cycle time of fewer than 100 picoseconds.

IV. Experimental Evaluation

Section IV-A examines the functional behavior of the proposed RNN via the optoelectronic circuit simulator. Then, using TensorFlow, Section IV-B demonstrates the area and power-saving effects thanks to the proposed RNN.

A. Optoelectronic Circuit Simulation

1) Evaluation Setup: As a test circuit, we design the optoelectronic RNN circuit with dynamic latches based on the circuit depicted in Fig. 2. The electrical signals x_0 and x_1 are inputs to the input layer, and v_0 and v_1 are inputs to the hidden layer. Only in the first clock cycle, h_0 and h_1 are used as the inputs to the hidden layer. V_0 and V_1 are the outputs from the activation function, which are recurrently connected to v_0 and v_1. Remind that the red and green allows are the optical waves. In the experiment, we prepare four sequences. More specifically, x_0 and x_1 are set to $[1, -1, 1, -1]$ and $[1, 1, 1, 1]$, respectively. The weights W_{0i} and W_{1i} are set to $[1, 1, 1, -1]$ and $[1, 1, 1, -1]$, as noted to red and green circles, respectively. Both h_0 and h_1 are set to 1. For the above setting, we preliminary calculate the ideal v_0 and v_1, which are both $[2, 2, 4, 6]$. In the experiment, we use an ELU function as the activation function, set the clock period to 100 ps, and simulate the optoelectronic circuit using a commercial optoelectronic circuit simulator in order to verify the behavior of the designed circuit.

2) Evaluation Results: Figure 3 shows the simulation results with the test circuit. In Fig. 3(a), solid lines correspond to V_0 and V_1, and donated gray lines are ideal results. From the figure, we can see that the output voltage converges to the expected value. For example, in the third clock cycle, V_0 converges to "4", which is an ideal result. This consistency indicates that optoelectronic functions work accurately through the dynamic latches and optical VMM. Figure 3(b) shows the comparison between the output and input values, e.g., V_0 and v_0. We can see that the output value in the previous clock cycle is successfully used in the next clock cycle. This means that the hold operation and synchronization in the dynamic latches work correctly, enabling an accurate control mechanism. From the above, we experimentally confirm that the proposed optoelectronic RNN circuit works correctly in terms of functional operation.

B. Accuracy, Power, and Area Estimation

1) Evaluation Setup: As a benchmark, we use MNIST, a widely used handwritten digits collection. To classify the MNIST images by RNN, we handle 28 iterations of 28 pixels for every image sample of MNIST. We design test circuits based on the proposed RNN architecture and evaluate their accuracy using TensorFlow, an open-source library for numerical computation and large-scale machine learning. The input and output layers have 28 and 10 nodes, respectively. The number of nodes in the hidden layer is swept to 16, 32, 64, and 98. We assume that the minimum detectable power of the

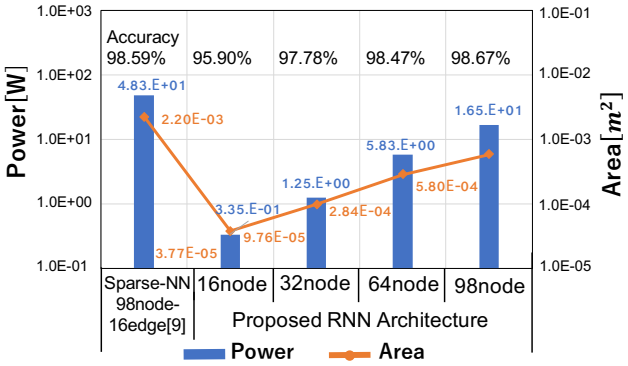

Fig. 4. Accuracy, power, and area comparison between MLP architecture [9] and proposed RNN architecture.

photodetector is 10 μW in this evaluation. The signal power of the laser source is determined so that the signal power is no less than the minimum detectable power of the photodetector (i.e., 10 μW), even if the signal power is attenuated through the power splitters and the combiners in the optical VMM. Based on the signal power of the laser sources, the total power consumption of the optoelectronic RNN circuit is calculated.

2) Evaluation Results: Bar charts in Fig. 4 show the power-saving effects thanks to the proposed RNN architecture. As shown in Fig. 4, the proposed RNN can dramatically reduce the power dissipation compared to the WDM-based optical MLP [9]. For example, when we set the number of nodes in the hidden layer to 98, the proposed RNN consumes the power dissipation of 1.65 $\times 10^1$ W while the MLP requires 4.83 $\times 10^1$ W. Namely, the proposed RNN reduces the power dissipation to 34%, which is a significant power saving effect. Moreover, by mitigating the accuracy constraint, the power dissipation of the proposed RNN can be further reduced. For example, when the accuracy constraint is set to 97% and the number of nodes in the hidden layer is set to 32, the power dissipation can be further reduced by one order of magnitude from 1.65 $\times 10^1$ W to 1.25 $\times 10^0$ W. From the above, we experimentally confirm that the proposed optoelectronic RNN can provide good inference accuracy while dramatically reducing power dissipation. The results of area comparison between the optical MLP proposed in [9] and our RNN architecture are shown in the line chart of Fig. 4. The area of Mach-Zehnder modulator and micro-ring resonator used in the comparison is determined based on the value used in [14] for fair comparison. As can be seen from the line chart in Fig. 4, the area of our RNN with 98-node is more than 3X smaller than the MLP based ONN [9] without sacrificing the inference accuracy. In summary, we experimentally confirm that the proposed optoelectronic RNN can provide good inference accuracy in image classification with a minimal area without sacrificing low-power and high-speed nature of light.

V. Conclusion

This paper proposed a novel architecture of optoelectronic RNN. The key idea behind the proposed architecture is that the vector-matrix multiplication part is implemented optically and the activation and feedback part is implemented electronically. Thanks to this electro-optic hybrid implementation, the proposed architecture fully takes advantage of the ultra high-speed nature of light and the controllability of the electronics. Experimental results obtained using TensorFlow showed that the proposed optoelectronic RNN architecture achieves more than 98% inference accuracy in image classification with a very compact and low-power circuit structure without sacrificing the high-speed nature of light. We also confirm the correct operation of the optoelectronic RNN using a commercial optoelectronic circuit simulator. Our future work will be devoted to developing edge and node pruning algorithms for achieving better power- and area-efficiency without compromising the accuracy of the classification.

Acknowledgement

This work is partly supported by JST CREST Grant Number JPMJCR21C3 and MEXT/JSPS KAKENHI Grant Number 20H04155.

References

[1] A. N. Tait et al., "Neuromorphic Photonic Networks using Silicon Photonic Weight Banks," *Sci. Rep.*, vol. 7, no. 1, Aug. 2017.
[2] W. Liu et al., "HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers," in *Proc. DATE*, March 2019, pp. 1483–1488.
[3] N. Janosik et al., "High-resolution Silicon Microring based Architecture for Optical Matrix Multiplication," in *Proc. CLEO*, no. SM2J.3, May 2019.
[4] M. B. On et al., "Wavelength-Space Domain High-Throughput Artificial Neural Networks by Parallel Photoelectric Matrix Multiplier," in *Proc. IEEE OFC*, Mar. 2020, pp. 1–3.
[5] Y. Shen et al., "Deep Learning with Coherent Nanophotonic Circuits," *Nature*, vol. 11, no. 7, p. 441–446, June 2017.
[6] J. Gu et al., "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE TCAD*, vol. 40, no. 9, pp. 1796–1809, Sep. 2021.
[7] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.
[8] T. Ishihara et al., "An Optical Neural Network Architecture based on Highly Parallelized WDM-Multiplier-Accumulator," in *Proc. IEEE/ACM Workshop on Photonics-Optics Technology Oriented Networking, Information and Computing Systems*, Nov. 2019, pp. 15–21.
[9] N. Hattori et al., "Optical-electronic implementation of artificial neural network for ultrafast and accurate inference processing," in *Proc. SPIE OPTO, AI and Optical Data Sciences II*, vol. 11703, 2021.
[10] F. Visin et al., "ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks," July 2015. [Online]. Available: arXiv:1505.00393
[11] G. Mourgias-Alexandris et al., "All-Optical WDM Recurrent Neural Networks With Gating," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 5, pp. 1–7, 2020.
[12] C. Feng et al., "Compact Design of On-Chip Elman Optical Recurrent Neural Network," in *Proc. CLEO*, no. JTh2B, May 2020.
[13] M. Miscuglio et al., "Photonic Tensor Cores for Machine Learning," *Applied Physics Reviews*, vol. 7, no. 031404, 2020.
[14] J. Gu et al., "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," in *Proc. DATE*, Feb. 2021, pp. 238–243.
[15] S. Kita et al., "Ultrashort Low-Loss ψ Gates for Linear Optical Logic on Si Photonics Platform," *Communications Physics*, vol. 3, no. 33, Mar. 2020.
[16] K. Nozaki et al., "Femtofarad Optoelectronic Integration Demonstrating Energy-Saving Signal Conversion and Nonlinear Functions," *Nature Photonics*, vol. 13, p. 454–459, July 2019.
[17] L. Liu et al., "An Ultra-Small, Low-Power, All-Optical Flip-Flop Memory on a Silicon Chip," *Nature Photonics*, vol. 4, p. 182–187, Jan. 2010.
[18] S. Sinha et al., "Design Benchmarking to 7nm with FinFET Predictive Technology Models," in *Proc. ISLPED*, July 2012, pp. 15–20.