An Optoelectronic Pipelined Convolutional-RNN Architecture for Energy-Efficient AI Accelerator

Chunlu Wang Yutaka Masuda Tohru Ishihara

Graduate School of Informatics
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Abstract— This paper proposes an optoelectronic Convolutional Recurrent Neural Network (C_RNN) architecture, employing RNN layers that replace area-consuming fully connected layers and process image data in a pipelined batch manner. It takes advantage of both the high feature extraction capabilities of CNNs and the compact and power-efficient nature of RNNs. The proposed optoelectronic C_RNN architecture achieves over 97.8% accuracy on the MNIST dataset while maintaining the advantages of power-efficient and high-speed characteristics of photonics. Our proposed optoelectronic C_RNN architecture can reach 240 TOPs/W, which is ten times more efficient than CMOS-based dedicated CNN accelerators.

I. Introduction

With the explosive spread of edge AI in today's advanced information society, energy-efficient AI accelerators are highly demanded [1]. However, according to Moore's Law, the performance improvements produced by the increased integration density of VLSI circuits are approaching the physical limits. Meanwhile, optical communication technology was also rapidly advancing, accelerating the progress of integrated nanophotonics technology. Historically, optical communication technology was primarily used for long-distance communication. However, recent evolutions in nanophotonics have rendered short-distance optical communication feasible [2]. Notably, the multiply-accumulate (MAC) operation inherent in AI accelerators, such as a neural network accelerator, can be effectively executed within optical circuits. Consequently, significant efforts have been made on the development of optical neural networks architecture [3–8].

This paper proposes an optoelectronic pipelined Convolutional Recurrent Neural Network (C_RNN) based on fully-optical vector-matrix multiplication. The main idea of this structure is to replace the fully connected layer in CNN with RNN. Although CNNs have inherent advantages in image processing, the fully connected layers used in CNNs often require hundreds or thousands of nodes, which limits the scalability of the optical implementation of neural networks. Building fully connected layers also incurs increased area and power consumption for custom hardware architectures of neural networks.

Recently, a CNN model utilizing the RNN layer has been proposed [9], which uses the RNN layer in addition to the fully connected layers to enhance the ability to capture spatial dependencies of images. In contrast, our proposed optoelectronic C_RNN architecture replaces the area-consuming fully connected layer with RNN layers to take advantage of the area and power-efficient nature of RNNs. Experimental results obtained using TensorFlow show that the proposed C_RNN achieves inference accuracy of more than 97.8% in MNIST, without sacrificing optics' power efficiency and high-speed nature. Using a commercial optoelectronic circuit simulator, we have also verified that the optoelectronic C_RNN works correctly.

The rest of the paper is organized as follows: Section II summarizes several previous works related to ONN. Section III presents the details of our optoelectronic C_RNN architecture. Section IV shows the experimental results obtained with a virtual environment for machine learning and optoelectronic simulation results using a SPICE-based circuit simulator. Section V concludes this paper.

II. Preliminary and Related Work

A. Basic Building Blocks for Optical Neural Network

Optical MAC units are an integral part of optoelectronic neural networks [3–8]. Figure 1 (a) shows a typical optical multiplication circuit. Suppose we calculate

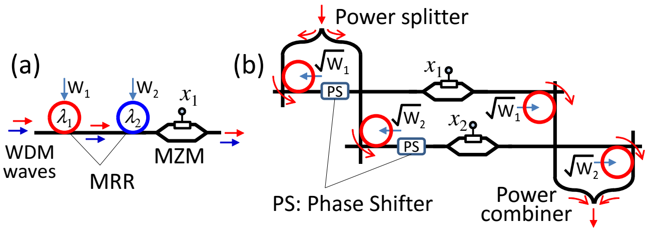

Fig. 1. (a) Optical multiplication circuit and (b) optical MAC unit

0.9×0.5 using carrier wave with a wavelength of λ_1. In this case, we set the micro-ring resonator (MRR) and Mach-Zehnder modulator (MZM) so that those modulators attenuate the carrier wave by 0.9 and 0.5, respectively. When the wave of λ_1 runs through the modulators in series, it takes attenuation of 0.9 in the MRR of labeling λ_1 and another attenuation of 0.5 in the MZM. As a result, the wave attenuated twice will have an attenuation of 0.45, corresponding to the product of 0.9×0.5. If we use multiple wavelengths, we can perform the multiplication in parallel since the waves with different wavelengths do not interfere.

Figure 1 (b) shows an example of an optical MAC unit. It calculates the MAC result of $Y_1 = W_1x_1 + W_2x_2$. First, an optical power splitter splits the carrier wave into two equally divided waves. The waves passing through the MRRs are bent 90 degrees and take a specific attenuation. In the left column, the wave will get into the upper row after taking the weight of $\sqrt{W_1}$, and so on in the right column. The phase shifter is used to represent the sign of the weights. If minus weight needs to be given at a specific MRR, the wave's phase should be shifted by 180 degrees at the corresponding phase shifter. Otherwise, the phase is not shifted. The waves passing through the two rows are summed each other at the power combiner. If a wave with a 180-degree shift and other waves without a shift are combined at the power combiner, the waves are weakened by each other. This corresponds to an arithmetic accumulation of negative and positive values. In this MAC unit, two MRRs are used for each row. This is because the phase of the wave passing through an MRR gets shifted depending on the weight value, and this phase shift will affect the arithmetic accumulation at the power combiner. Therefore, we use two MRRs for representing a specific weight so that the two series MRRs cancel out the phase shifts obtained at the MRRs. This is the basic principle of the optical MAC operation. We fully exploit this optical MAC in our C_RNN architecture.

B. Optoelectronic Hybrid RNN using WDM

An optoelectronic RNN architecture based on dynamic latches has been proposed in [8]. The key to the architecture is to implement the vector-matrix multiplication optically to exploit the speed of light and implement the activation and feedback mechanisms electronically to exploit the controllability of electronics. The photonics part uses a wavelength division multiplexing (WDM) technology to exploit the parallelism of the light, as shown in the lower part of Fig. 2. Since optical waves with different wavelengths do not interfere, parallel calculations can be done in optical circuits using the WDM waves, as explained in Fig. 1. The electronics part comprises electrical feedback with a dynamic latch to synchronize the recurrent loops with a clock signal, as shown in the upper part of Fig. 2. The dynamic latches directly store analog values in the parasitic capacitances of the MZM inputs. The optical-to-electrical signal conversion delay is much smaller than the analog-to-digital conversion thanks to the nanophotonic O-E-O converter proposed in [10]. We fully exploit this electro-optic hybrid RNN circuit in our proposed C_RNN architecture.

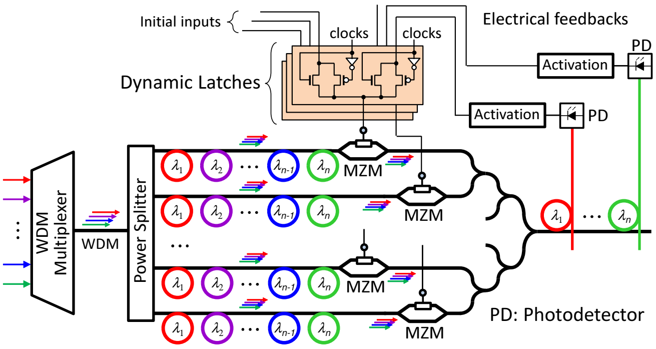

Fig. 2. Optoelectronic Hybrid RNN Exploiting WDM.

C. Optical Convolutional Processing Circuit

An integrated photonic hardware accelerator for convolution operations has been proposed in [4]. It can run at 10^{12} MAC operations per second and achieves 96.1% accuracy on MNIST. An electronic digital signal processor (DSP) implements the fully connected layer and activation function. This implementation involves analog-to-digital (AD) conversion between analog optical and digital electrical signals that cause substantial power consumption and latency.

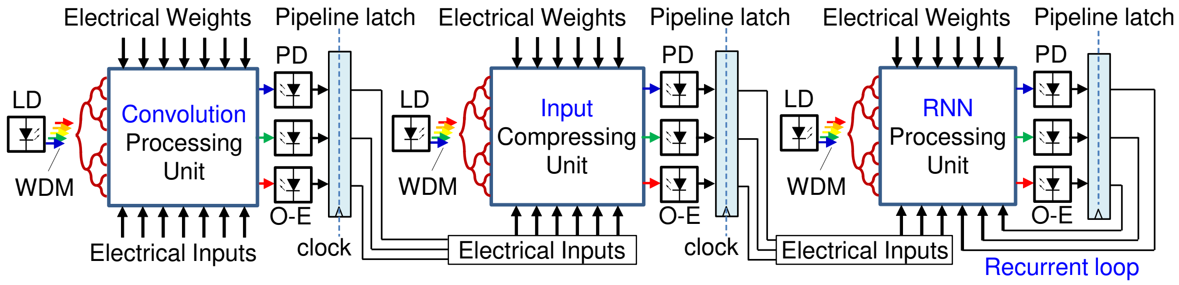

Fig. 3. Pipelined Optoelectronic C_RNN Architecture using Wavelength Division Multiplexing (WDM).

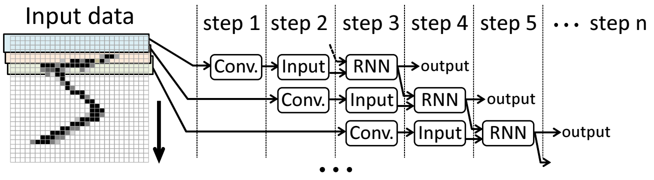

Fig. 4. Pipelined C$_R$*NNprocessing*.

III. Optoelectronic C_RNN Architecture

A. Pipelined Optoelectronic C_RNN Architecture

Figure 3 shows the overall structure of the proposed C_RNN. It consists of three pipeline stages. The convolution processing unit is responsible for image feature extraction, the input compressing unit performs input dimension compression for the RNN processing unit, and the RNN processing unit works as a replacement for the fully connected layer in traditional CNNs. The key to this architecture is to divide large input data into small pieces and process them by a compact optical circuit in an iterative pipeline manner, as shown in Fig. 4. In a typical optical circuit, a large portion of the energy is consumed by laser light (LD in Fig. 3). The loop-based architecture that iteratively utilizes this laser power thus leads to a reduction of energy consumption. The target frequency for the loop in this circuit is 20 GHz. Although the high clock frequency in CMOS circuits causes an increase in power consumption, the power consumed in the optical circuits only weakly depends on the frequency since the power consumed by the laser light is dominant and independent of the frequency. Therefore, the higher the frequency in the optical circuits, the smaller the energy and latency per inference.

B. Optical Convolution Processing Unit

The convolution processing unit consists of a convolution layer, an average pooling layer, and an activation function. To make the optical circuit simple, we configure the circuit to process convolution, pooling, and activation functions in that order. Note that the normal order is convolution, activation, and pooling. Preliminary experimental results using TensorFlow show that this reordering has a negligible impact on inference accuracy.

Figure 5 (a) shows an optoelectronic circuit implementation for the convolution processing unit. As explained in subsection II.A, our optical circuits are designed with the optical MAC unit. First, we use the wavelength division multiplexing (WDM) technology to multiplex the two waves into the same waveguide. Then, we equally divide the WDM waves with the power splitter and pass them into the MRRs. As explained with Fig. 1, waves having a wavelength of λ_1 are bent 90 degrees at the MRRs labeling λ_1 and take a specific attenuation programmed. We use two series MRRs to give a single weight value to the wave as explained with Fig. 1 (b). The phase shifters (PSs) are used for representing the sign of weights. The MZMs are used for multiplication with input values (x_i). Next, we show an architecture that performs convolution and average pooling using only a single-step MAC operation. The convolution is performed with four 2 × 2 kernels, namely kernel 1 to kernel 4 as follows:

kernel 1: $[[W_{1,1}, W_{1,2}], [W_{1,3}, W_{1,4}]]$,
kernel 2: $[[W_{2,1}, W_{2,2}], [W_{2,3}, W_{2,4}]]$,
kernel 3: $[[W_{3,1}, W_{3,2}], [W_{3,3}, W_{3,4}]]$,
kernel 4: $[[W_{4,1}, W_{4,2}], [W_{4,3}, W_{4,4}]]$.

We use precomputed weights to simultaneously perform the convolution and average pooling in the optical circuit. For the combination of 2 × 2 convolution with a stride of 1 × 1 and 2 × 2 pooling with a stride of 2 × 2, every 3 × 3 region which consists of nine pixels is processed by the combined convolution and average-pooling. As shown in Fig. 6, our circuit takes the image's top-left 3 × 3 pixel region as the first target region. When we process the region with kenel 1, equation (1) should be calculated, where the weights are aggregated to nine weights which

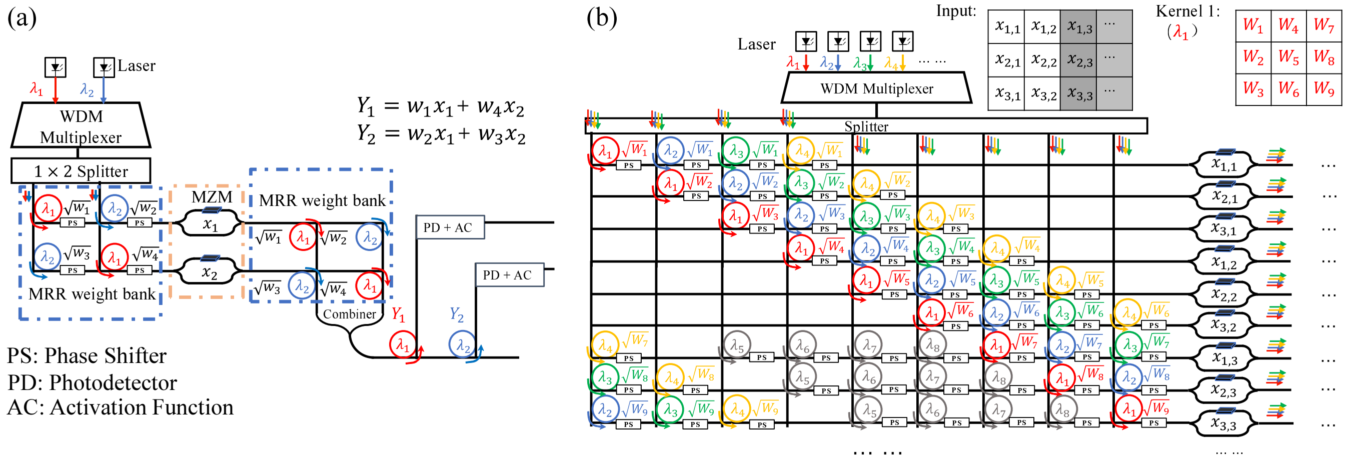

Fig. 5. Convolution operations with four 2 × 2 convolution kernels on a 2 × 2 image region.

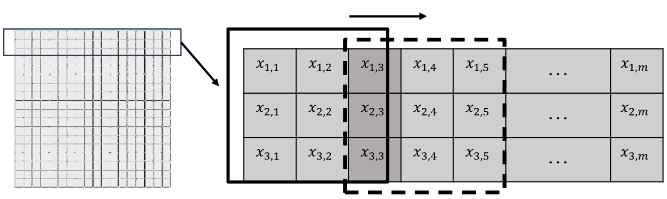
$x_{1,1}$ $x_{1,2}$ $x_{1,3}$ $x_{1,4}$ $x_{1,5}$ … $x_{1,m}$
$x_{2,1}$ $x_{2,2}$ $x_{2,3}$ $x_{2,4}$ $x_{2,5}$ … $x_{2,m}$
$x_{3,1}$ $x_{3,2}$ $x_{3,3}$ $x_{3,4}$ $x_{3,5}$ … $x_{3,m}$

Fig. 6. Example of pipelined convolution processing.

correspond to the input pixels ($x_{i,j}$):

$$\begin{aligned} f_{avepool} &= \frac{1}{4}W_{1,1}x_{1,1} \; + \; \frac{1}{4}(W_{1,1}+W_{1,2})x_{1,2} \; + \; \frac{1}{4}W_{1,2}x_{1,3} \\ &+ \; \frac{1}{4}(W_{1,1}+W_{1,3})x_{2,1} \; + \; \frac{1}{4}(W_{1,1}+W_{1,2}+W_{1,3}+W_{1,4})x_{2,2} \\ &+ \; \frac{1}{4}(W_{1,2}+W_{1,4})x_{2,3} \; + \; \frac{1}{4}W_{1,3}x_{3,1} \; + \; \frac{1}{4}(W_{1,3}+W_{1,4})x_{3,2} \\ &+ \; \frac{1}{4}W_{1,4}x_{3,3} \end{aligned} \tag{1}$$

Using the nine pre-aggregated weights, the combined convolution and average pooling can be performed in a single step, as shown in equation (1). This largely simplifies the optical circuit for this operation. Figure 5(b) shows the partial optical circuit diagram corresponding to this operation. The next computation region is considered after shifting the pooling window by two pixels to the right. The overlapping third column shown in Fig. 6 will be involved in the subsequent convolution again. We use different wavelengths of light and corresponding MRRs to distinguish the multiplication results for the overlapping pixels as shown in the lower three rows in Fig. 5 (b). On the row in the optical circuit corresponding to $x_{i,j}$ of the overlapping pixels shown in Fig. 6, the number of MRRs with different resonant wavelengths in the left and right MRR weight bank arrays will both become 8.

C. *Dynamic Latch and Activation Function*

Since our circuit shown in Fig. 5 calculates the combined convolution and average-pooling as the electric field strength value, it is necessary to extract the electric field strength value from the optical signal. Note that the signal power P is the square of the electric field strength E (i.e., $P = E^2$). If we simply use a photodetector to convert an optical signal into a photocurrent, the photocurrent is proportional to the signal power rather than the electric field strength. Homodyne detection is commonly used for extracting the electrical field strength value as an electrical signal from the optical signal [3]. The electrical signal is then processed by an activation function implemented as a dedicated electrical circuit, which is explained in Fig. 2. After each activation operation, we utilize dynamic latches [8] to temporarily store an electrical signal as an analog value. The dynamic latch was initially proposed for temporarily storing analog values in electrical analog neural networks [11]. We use the dynamic latches between all pipeline stages in our circuit, as shown in Fig. 3. Since the leakage current drawn from an off-state dynamic latch is on the order of picoamperes, the charge lost due to leakage current within a few tens of picoseconds is about 0.001% of the charge stored in the dynamic latch, which is negligible. The dynamic latches can also be used for energy-efficient recurrent loops within the RNN Processing Unit [8].

An activation function used in our circuit is a hyperbolic tangent function (tanh). The dynamic latch also works to saturate voltages below 0V or above the supply voltage (that is, 1.2V) into 0V to 1.2V, which corresponds to the tanh function.

D. *Optical Input Compressing Unit and RNN Processing Unit*

The overall structure of the RNN Processing Unit is depicted in Fig. 2. The Input Compressing Unit is similar to this structure. However, the output signal of the Input Compressing Unit is not feedback to itself but directly input to the RNN Processing Unit as shown in Fig. 3. Both the Input Compressing Unit and the RNN Processing Unit employ optical VMM composed of MRR weight banks and MZMs to compute in the optics domain, while the activation and feedback circuit work in the electronics domain.

The Input Compressing Unit is designed to reduce the input dimension of the RNN Processing Unit, thereby reducing the number of MZMs used in the RNN Processing Unit. Our preliminary experiments showed that the Input Compressing Unit could compress the input dimension by more than one-fourth without sacrificing the inference accuracy.

IV. Experimental Evaluation

A. *Optoelectronic Circuit Simulation*

As a test circuit, we designed a circuit of the Convolution Processing Unit with a 3 × 5 input. In the experiment, we set the clock period to 200 ps, and we prepared the input sequence as follows: 0s: [[0,0,0,0,-1], [0,1,1,1,1], [0,0,0,0,1]], 2E-10s: [[0,0,0,0,0], [0,0,0,0,0], [0,0,0,0,0]], 4E-10s: [[1,0,-1,0,1], [1,0,-1,0,1], [1,0,-1,0,1]], 6E-10s: [[1,1,1,1,1], [1,1,1,1,1], [1,1,1,1,1]], 8E-10s: [[0,0,0,1,1], [0,0,0,1,1], [0,0,0,1,1]]. $V_0 - V_7$ are the outputs of the optical circuit and are connected to the next layer. The weights of the four convolution kernels are set to kernel 1: [[1, 0], [0, -1]], kernel 2: [[0,1], [-1, 0]], kernel 3: [[-1, -1], [1, 1]] and kernel 4: [[1, -1], [1, -1]].

The final simulation results are shown in Fig. 7. The output voltages correspond to the ideal values. The consistency between output voltages and expected values indicates that optoelectronic functions work accurately through the Convolution Processing Unit.

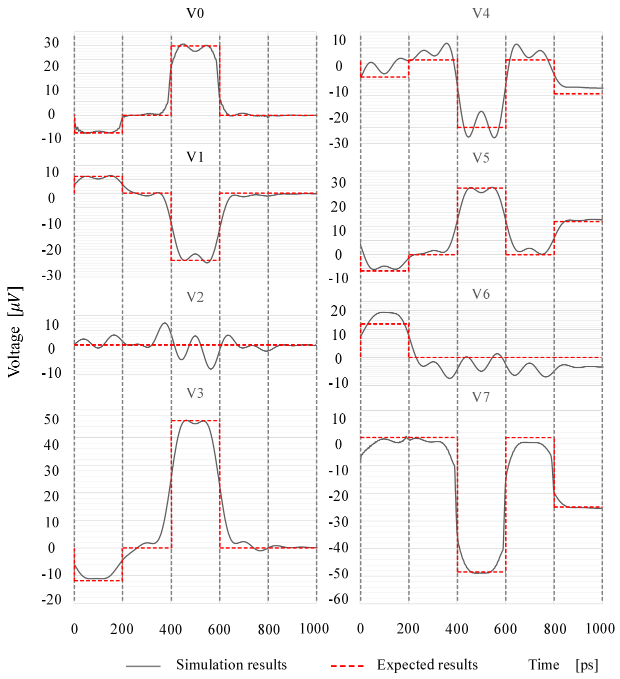

Fig. 7. Optoelectronic circuit simulation results. Solid lines correspond to V0 to V7 simulation results, and dotted lines are expected results. 6μ represents the output value $\frac{1}{4}$, 12μ represents the output value $\frac{1}{2}$, and so on.

B. *Accuracy, Power, and Area Estimation*

B.1. Evaluation Setup

As a benchmark, we use MNIST for image classification. We designed test circuits based on the proposed C_RNN architecture and evaluated its accuracy using TensorFlow.

After the input image is processed by convolution and average-pooling layers, the original input image of 28 × 28 × 1 is converted to 13 × 13 × 4. The convolution layer consists of four 2×2 kernels, and the average-pooling layer uses a window size of 2 × 2 with a stride of 2 × 2 for the average pooling operation.

B.2. Pipelined Input

The input image is divided into small image fragments to make the optical circuit compact and power efficient. Here, patch size refers to the amount of input data processed per single step in the model, as explained in Fig. 4. We evaluated the inference accuracy for different patch sizes. Experimental results show that for the original image of 28×28, patch size= 3×28 can get a better accuracy.

The result obtained by Convolutional Processing Unit consists of four channels that correspond to four kernels. RNN was originally designed to handle sequential data, and the input each step can only be one-dimensional information. To achieve higher accuracy, we tested multiple input scenarios. Ultimately, we found that the method shown in Fig. 8 achieves the highest accuracy.

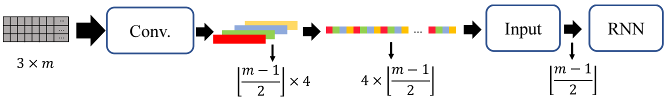
3 × m
Conv.
Input
RNN
$\left\lfloor \frac{m-1}{2} \right\rfloor \times 4$
$4 \times \left\lfloor \frac{m-1}{2} \right\rfloor$
...
$\left\lfloor \frac{m-1}{2} \right\rfloor$

Fig. 8. Convert multi-dimensional tensor into one-dimensional tensor for input into RNN.

The Input Compressing Unit has 13 output nodes; therefore, the RNN Processing Unit has 13 input nodes. The number of nodes in the hidden layer of the RNN in this experiment is set to 54.

B.3. Evaluation Result

For comparison, we conduct experiments under the same conditions on several models . The experimental results are presented in Fig. 9. Conv_2FC means a typical CNN model consisting of a convolution layer, an average pooling layer, a fully connected layer, a layer with 98 nodes, and an output layer. The parameters of the convolution layer and the pooling layer are consistent with the proposed C_RNN architecture.

RNN means recurrent neural network with 98 nodes.

In this experiment, we assume that the maximum signal power of the light given to the photodetector (PD in short) is 50 μW. This is because the maximum allowable voltage of the electrical circuit used for the activation function is 1.2 V and the voltage is produced as the product of the 24 kΩ resistance and the photocurrent (i.e., 50 μA) generated by the PD with a responsivity of 1 AW^{-1} [10]. Therefore, the signal power of the laser diode (LD in short) is determined so that the signal power at the PD is no more than the maximum allowable power of the PD (i.e., 50 μW) when the signal power is attenuated through the MRRs and MZMs in the optical VMM. It should be noted that in our proposed optical circuit, light passing through a set of MRRs, an MZM, and a phase shifter will attenuate to 0.075 times of the original signal power in total, and the emission efficiency of the LD

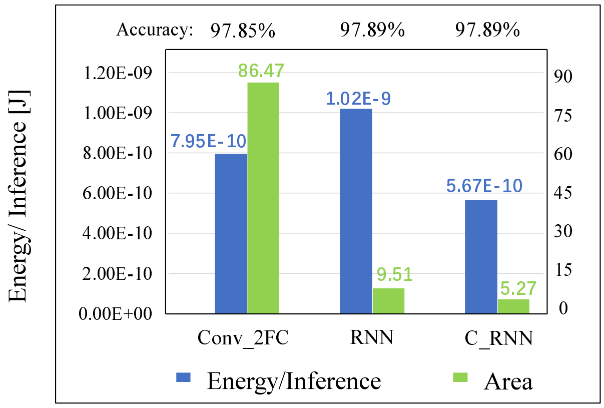

Fig. 9. Energy, area and accuracy comparison between Conv_2FC, RNN architecture and proposed C_RNN architecture.

is roughly 20%. Based on the power consumption of the LD, the total power consumption of all the optoelectronic circuits is calculated.

The bar graph in Fig. 9 shows the comparison of the proposed C_RNN model with other models in terms of energy consumption per inference and area on the MNIST dataset. The area of the Mach-Zehnder modulator and micro-ring resonator used in the comparison is determined based on the values used in [12] for fair comparison.

In our architecture, the running time of each unit can be guaranteed to not exceed 50 ps. We run the whole architecture at 20 GHz. When we set the number of nodes in the RNN hidden layer as 54, the proposed C_RNN consumes 5.67×10^2 pJ per inference, and its area is 5.27 mm^2. It can be seen from Fig. 9 that our proposed structure has the smallest energy consumption per inference and area compared to other structures. Compared with Conv_2FC, our proposed model uses the RNN layer to replace the fully connected layer and reduces each input to process images, significantly reducing the area consumption. Because of the pipeline processing mode, the energy consumption of the circuit per cycle is reduced. Compared with the RNN model, the import of the Input Layer Processing Unit greatly reduces the number of nodes in the RNN Unit, and the proposed C_RNN architecture's circuit area is only 55.4% of RNN architecture. Moreover, thanks to the addition of the average-pooling layer, which reduces the number of cycles required to process an image, the proposed C_RNN architecture only requires 55.59% of the energy consumption of the RNN structure to process an image.

Lightspeeur 2803s [13] achieves 24 TOPs/W, which is one of the highest energy efficiency among CMOS-based CNN accelerators. Our proposed optoelectronic C_RNN architecture can reach 240 TOPs/W, which is ten times more energy efficient.

V. Conclusion

In this paper, we propose a novel optoelectronic C_RNN architecture, which replaces the fully connected layers in the traditional Convolutional Neural Network (CNN) model with Recurrent Neural Network (RNN) layers. The vector-matrix multiplication is implemented optically, and activation and feedback mechanisms are realized electronically while interconnecting layers using electrical signals. This electro-optic hybrid implementation takes advantage of both the ultra-high-speed characteristics of light and the controllability of electronics. Experimental results obtained using TensorFlow demonstrate that the proposed optoelectronic C_RNN architecture achieves favorable performance in small-scale grayscale image classification tasks with a highly compact and low-power circuit structure without sacrificing the high-speed nature of light. We also verify the correct operation of the optoelectronic C_RNN using a commercial optoelectronic circuit simulator. Our future work will focus on extending the Convolutional processing unit to achieve better accuracy while balancing power efficiency and circuit area.

Acknowledgement

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

References

[1] M. M. H. Shuvo *et al.*, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," in *Proc. IEEE*, vol. 111, no. 1, Jan. 2023, pp. 42–91.

[2] X. Wu *et al.*, "Suor: Sectioned undirectional optical ring for chip multiprocessor," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 10, no. 4, pp. 1–25, 2014.

[3] N. Hattori *et al.*, "Optical-electronic implementation of artificial neural network for ultrafast and accurate inference processing," in *Proc. SPIE, AI and Optical Data Sciences II*, vol. 11703, 2021, p. 117031E.

[4] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[5] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific reports*, vol. 7, no. 1, p. 7430, 2017.

[6] Y. Shen *et al.*, "Deep Learning with Coherent Nanophotonic Circuits," *Nature*, vol. 11, no. 7, p. 441–446, June 2017.

[7] 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.

[8] T. Ichikawa *et al.*, "Optoelectronic implementation of compact and power-efficient recurrent neural networks," in *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2022, pp. 390–393.

[9] Z. Zuo *et al.*, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 18–26.

[10] K. Nozaki *et al.*, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," *Nature Photonics*, vol. 13, no. 7, pp. 454–459, 2019.

[11] Y. Arima *et al.*, "A 336-neuron, 28 k-synapse, self-learning neural network chip with branch-neuron-unit architecture," *IEEE journal of solid-state circuits*, vol. 26, no. 11, pp. 1637–1644, 1991.

[12] J. Gu *et al.*, "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," in *Proc. DATE*, February 2021, pp. 238–243.

[13] M. P. Véstias *et al.*, "Moving deep learning to the edge," *Algorithms*, vol. 13, no. 5, p. 125, 2020.