Compact Design of On-chip Elman Optical Recurrent Neural Network

Chenghao Feng[1], Zheng Zhao [2], Zhoufeng Ying[1], Jiaqi Gu[2], David Z Pan[2], and Ray T Chen[1]

[1]*Microelectronics Research Center, The University of Texas at Austin, Austin, Texas 78758, USA*

[2]*Computer Engineering Research Center, The University of Texas at Austin, Austin, Texas 78705, USA*

[*] *e-mail address: chenrt@austin.utexas.edu*

Abstract: We propose an on-chip optical Elman recurrent neuron network (RNN) architecture for high-speed sequence processing using Mach-Zehnder interferometers and looped waveguides. The proposed design paves way for future integrated-photonics-based artificial intelligence hardware design.

OCIS codes: (200.4260) Neural networks; (130.3120) Integrated optics devices.

1. Introduction

Artificial neural network (ANN) is developing rapidly and is shaping our lives. Photonic computing has been rekindled as promising for implementing machine learning tasks due to high transmission speed, low power consumption, and advantages in matrix computing compared with electronic architectures [1]. Among ANN architectures, recurrent neuron network (RNN) is specialized in time sequence processing, like speech recognition and sequence prediction.

Using delay systems to realize RNN with integrated photonic circuits is popular in recent studies [2]. Most studies focus on reservoir computing [3], the models of which are not similar to RNN models studied in computer science. As a result, they are not compatible with current training methods in the traditional machine learning area when the model complexity increases.

In this paper, we propose a novel and compact on-chip photonics architecture realizing the classical Elman RNN model, which is well studied in computer science. The mechanism of our architecture is illustrated first and then verified via simulations, which reveals that our architecture is capable of handling problems related to sequence processing. Analysis of the performance as well as the robustness of our architecture will be carried out in the future.

2. Theory and design

Elman RNN defines the following behavior:

$$\mathbf{h}(t) = \boldsymbol{\sigma}_1(\mathbf{W}_{\mathbf{hh}}\mathbf{h}(t-1) + \mathbf{W}_{\mathbf{hx}}\mathbf{x}(t) + \mathbf{b}_{\mathbf{o}}) \tag{1}$$

$$\mathbf{o}(t) = \boldsymbol{\sigma}_2(\mathbf{W}_{\mathbf{ho}}\mathbf{h}(t)). \tag{2}$$

where the hidden state $\mathbf{h}(\mathbf{t})$ is defined by its previous state (t-1) and input signal $\mathbf{x}$(t). $\mathbf{W_{hx}}$, $\mathbf{W_{hh}}$, $\mathbf{W_{ho}}$ are real matrices while $\boldsymbol{\sigma_1}, \boldsymbol{\sigma_2}$ are non-linear activation functions. A great challenge for building optical RNN is that few on-chip silicon photonics devices can memorize. While using electronics components to realize the recurrent relation will slow down the calculation and cause excessive energy consumption.

The schematic architecture of our optical RNN (ORNN) is shown in figure 1, where a 2×2 photonics RNN architecture and its n×n extension are shown. The mechanism of our architecture is shown as follows: The inputs of ORNN $\mathbf{x}$(t) are connected to an MZI array that can implement any real matrix [1] denoted as $\mathbf{W_1}$ producing $\mathbf{x'}$(t). $\mathbf{x'}$(t) will interfere with $\mathbf{g}$ via a 50/50 directional coupler, producing $\mathbf{h}$(t). Then light $\mathbf{h}$(t) will propagate through the second matrix array that can implement matrix $\mathbf{W_2}$ and then be looped back, producing $\mathbf{g}$ to interfere with $\mathbf{x}$(t) in the directional coupler. Finally, some portion of $\mathbf{h}$(t) will be coupled out. The coupled light will pass through a non-linear activation block $\boldsymbol{\sigma}$, producing outputs $\mathbf{o}$(t). The transfer functions can be simplified as:

$$\mathbf{h}(t) = (\mathbf{W}_{\mathbf{hh}}\mathbf{h}(t-\tau-\tau_1)e^{-i\phi-\pi/2} + \mathbf{W}_{\mathbf{hx}}\mathbf{x}(t-\tau_1))e^{-i\phi_1} \tag{3}$$

$$\mathbf{o}(t) = \boldsymbol{\sigma}(\mathbf{W}_{\mathbf{ho}}\mathbf{h}(t)). \tag{4}$$

$$\mathbf{W}_{\mathbf{hh}} = (\sqrt{2}/2)\mathbf{W}_2; \mathbf{W}_{\mathbf{hx}} = (\sqrt{2}/2)r\mathbf{W}_1; \mathbf{W}_{\mathbf{ho}} = ik \tag{5}$$

where τ_1 and ϕ_1 are the signal transmission time and phase change through the MZI array $\mathbf{W_1}$ while τ and ϕ are the signal transmission time and phase delay through $\mathbf{W_2}$ as well as the feedback loop. By selecting appropriate waveguide length, we let $\phi = \frac{\pi}{2} + \phi_1 + 2m\pi$, m is an integer. Other phase changes that will not result in intensity

changes of **o**(t) are not mentioned for simplicity. k is the coupling coefficient between the waveguides carrying signals **h**(t) and **o**(t), while r is the transmission coefficient.

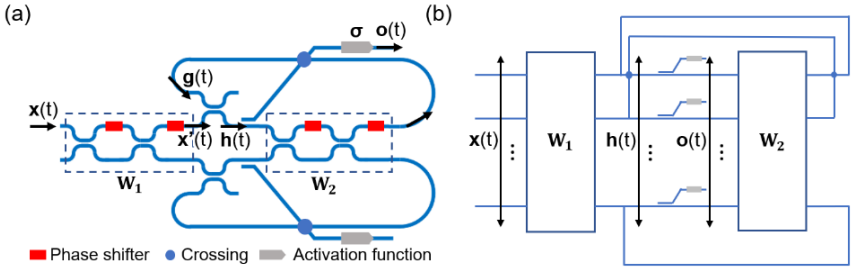


Figure 1: (a) Schematic structure of a 2*2 photonic RNN and (b) n*n photonic RNN.

Comparing the field intensity of **h(t)** and **o(t)** in Eq. (3, 4) with Eq. (1, 2), one can find the transfer function of our architecture is similar to Elman RNN architecture except that $\boldsymbol{\sigma_1}$ is not used to reduce the number of non-linear optical devices. Therefore, our proposed structure can be trained with well-developed training methods.

3. Simulation result

We train the proposed ORNN architecture in PyTorch framework based on Eq. (3-5) and verify our model using Lumerical Interconnect, where the parameters of optical components are set based on previous literatures and process design kit (PDK) models of silicon photonics foundries [4]. The bit rate of the signal is 100 Gb/s. Amplifiers are deployed in the feedback loop to compensate for propagation loss of directional couplers and waveguides. The activation functions of ORNN can be realized by saturable absorbers or electrooptic modulators in previous literatures [5]. Here, we build an ORNN with two inputs, which is similar to the structure in Fig. 1(a).

The training results are shown in Fig. 2. Figure 2(a) measures the step response of the ORNN where $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$ are identical matrices, which reveals that our model matches well with Lumerical Interconnect, a reliable commercial software. Fig. 2(b) and 2(c) show a proof-of-concept application of a sequential binary adder. Note that in the simulation results, the least significant bit (LSB) comes first. The converted decimal sums match well with the correct results. More training results and applications of our ORNN will be shown in the presentation.

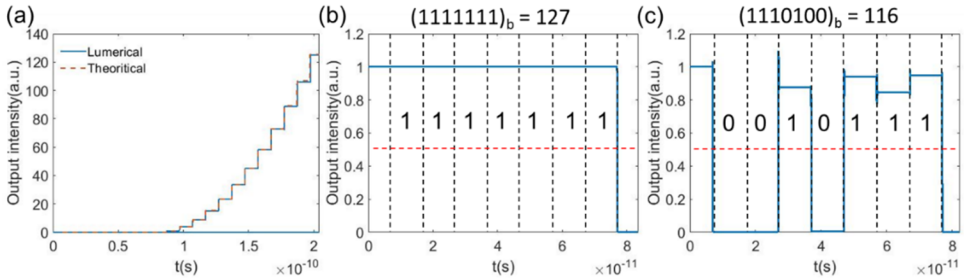


Figure 2 Trained waveforms of optical RNNs (a) step response (b) Sequential full adder results (89+38=127) (c) Sequential full adder results (52+64=116). Both (b) (c) are simulated on Lumerical Interconnect.

In conclusion, we have proposed a compact architecture of Elman recurrent neuron networks and test its performance with Lumerical Interconnect. Our architecture outperforms electronics counterparts in speed and energy consumption. Besides, the architecture is compatible with current machine learning training methodologies. This study paves the way for future large-scale deep learning and neural computing on photonics chips.

[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. Solja, "Deep learning coherent nanophotonic circuits," 1–18 (2017).

[2] D. Brunner, B. Penkovsky, B. A. Marquez, M. Jacquot, I. Fischer, and L. Larger, "Tutorial: Photonic neural networks in delay systems," J. Appl. Phys. **124**(15), (2018).

[3] H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Neuromorphic Photonic Integrated Circuits," IEEE J. Sel. Top. Quantum Electron. (2018).

[4] E. Timurdogan, Z. Su, R.-J. Shiue, C. V. Poulton, M. J. Byrd, S. Xin, and M. R. Watts, "APSUNY Process Design Kit (PDKv3.0): O, C and L Band Silicon Photonics Component Libraries on 300mm Wafers," 2019 Opt. Fiber Commun. Conf. Exhib. Tu2A.1 (2019).

[5] I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks," 1–12 (2019).

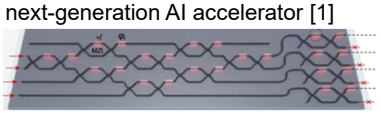
The authors acknowledge support from the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR) (Grant No. FA 955017-1-0071), monitored by Dr. Gernot S. Pomrenke.

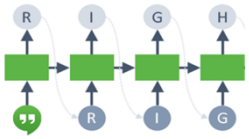
Compact Design of On-chip Elman Optical Recurrent Neural Network

Chenghao Feng[1], Zheng Zhao [2] , Zhoufeng Ying [1], Jiaqi Gu [2], David Z. Pan [2] and Ray T. Chen [1]
[1]*Microelectronics Research Center , The University of Texas at Austin, Austin, TX 78758, USA*
[2]*Computer Engineering Research Center, The University of Texas at Austin, Austin, TX 78705, USA*

Background

- Photonic computing as promising next-generation AI accelerator [1]
  - High transmission speed
  - High bandwidth
  - Low power consumption
- Recurrent neuron network (RNN) has wide application
  - Speech recognition
  - Sequence prediction
- Previous RNN focus on reservoir computing





Need efficient optical RNN architecture

Theory and Architecture

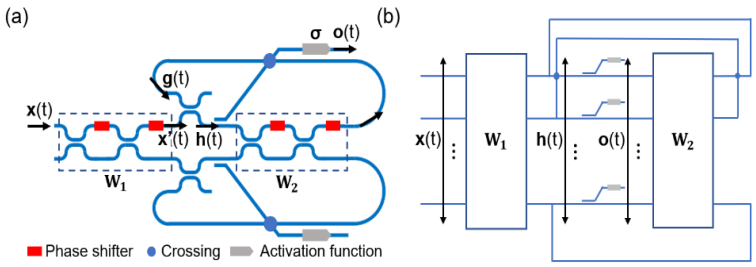
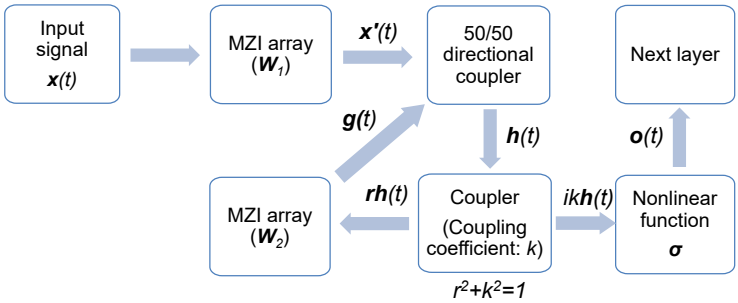


Figure 1: (a) Schematic of a 2x2 photonic RNN and (b) n x n photonic RNN.

Optical path:



Transfer function:

$$\boldsymbol{h}(t) = (\boldsymbol{W_{hh}}\boldsymbol{h}(t-\tau-\tau_1)e^{-i\phi-\pi/2} + \boldsymbol{W_{hx}}\boldsymbol{x}(t-\tau_1))e^{-i\phi_1} \quad (1)$$

$$\mathbf{o}(t) = \boldsymbol{\sigma}(\boldsymbol{W_{ho}}\boldsymbol{h}(t)) \quad (2)$$

$$\boldsymbol{W_{hh}} = (\sqrt{2}/2)\boldsymbol{W}_2; \boldsymbol{W_{hx}} = (\sqrt{2}/2)r\boldsymbol{W}_1; \boldsymbol{W_{ho}} = ik \quad (3)$$

τ_1, ϕ_1: transmission time and phase delay through the MZI array $\boldsymbol{W_1}$

τ, ϕ : transmission time and phase delay from $\boldsymbol{W_2}$ to 50/50 coupler

Let $\phi = \frac{\pi}{2} + \phi_1 + 2m\pi$, m is an integer, the transfer function is similar to RNN function:

$$\boldsymbol{h}(t) = \boldsymbol{\sigma_1}(\boldsymbol{W_{hh}}\boldsymbol{h}(t-1) + \boldsymbol{W_{hx}}\boldsymbol{x}(t) + \boldsymbol{b_o}) \quad (4)$$

$$\boldsymbol{o}(t) = \boldsymbol{\sigma_2}(\boldsymbol{W_{ho}}\boldsymbol{h}(t)) \quad (5)$$

Simulation Methodology

Software platform: PyTorch

Model verification platform: lumerical

Other optical components

- On-chip optical amplifiers
- Silicon photonics devices from the PDK of AIM photonics [4]
- Nonlinear functions based on Optical-electrical-optical (OEO) conversions [5]

Model Verification

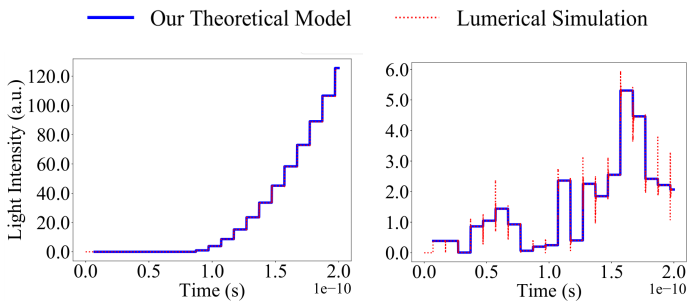


Figure. 2 (a) step response of the RNN, both $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$ are identity matrices (b) Output of the RNN, where $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$ are 2x2 matrices.

Our model fits well with Lumerical Interconnect simulation results.

Machine Learning Task Demo

Application: 7-bit sequential adder:

Input: 7-bit binary number strings, 100 Gb/s

Parameters: $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$: 2x2 matrices, activation function: *tanh(x)*

Output: 7-bit binary number strings, 100% accuracy
Least significant bit (LSB) comes first

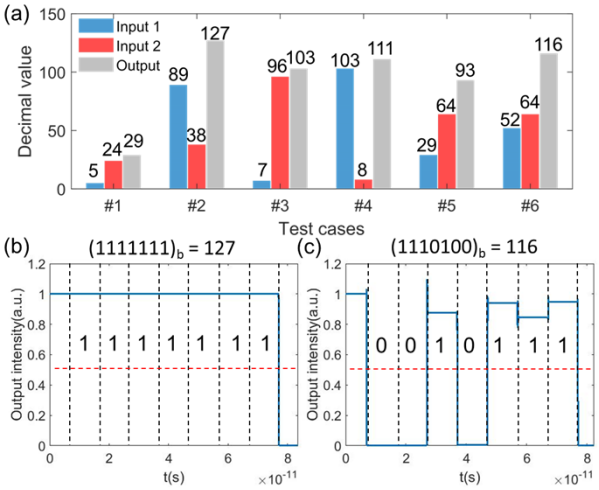


Figure. 3 Tested waveforms of optical RNNs (a) Results of some testing samples (b) Sequential full adder results (89+38=127) (c) Sequential full adder results (52+64=116). Both (a) (b) are simulated on Lumerical Interconnect.

Outlook

- We will investigate more applications of our proposed ORNN.
- We are analyzing and optimizing the robustness of our proposed RNN.
- More advanced RNN architecture, such as Gated Recurrent Unit (GRU) will be realized using optical structures and controlling electrical circuits.

Reference

[1] Y. Shen, et al. Nature Photonics 11.7 (2017): 441.
[2] D. Brunner,et al. J. Appl. Phys. 124(15), (2018).
[3] H. T. Peng, et al. JSTQE, 24.6 (2018): 1-15.
[4] E. Timurdogan, et al. 2019 Opt. Fiber Commun. Conf. Exhib. Tu2A.1 (2019).
[5] I. A. D. Williamson, et al, JSTQE, 26.1 (2019): 1-12.

❖ The work is supported by MURI program through the Air Force Office of Scientific Research (AFOSR), contract No.FA 9550-17-1-0071