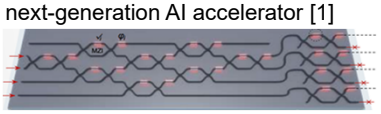
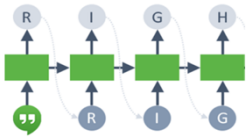
Compact Design of On-chip Elman Optical Recurrent Neural Network

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

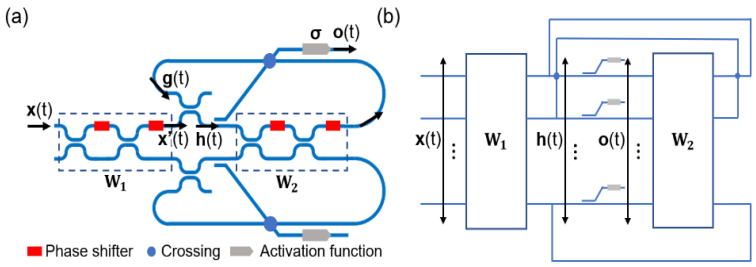
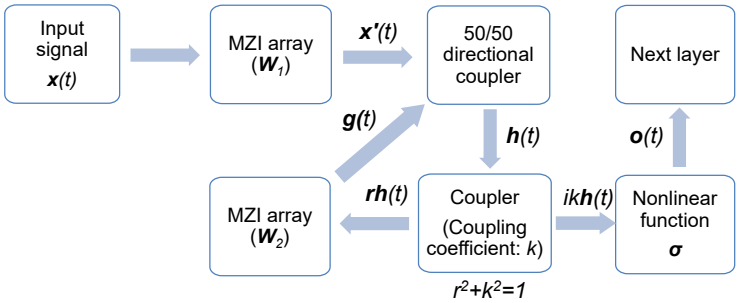


Figure 1: (a) Schematic of a 2x2 photonic RNN and (b) n x n photonic RNN.

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

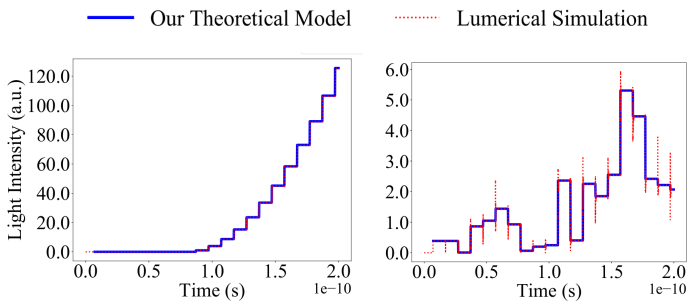


Figure. 2 (a) step response of the RNN, both $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$ are identity matrices (b) Output of the RNN, where $\boldsymbol{W_{hh}}$ and $\boldsymbol{W_{hx}}$ are 2x2 matrices.

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

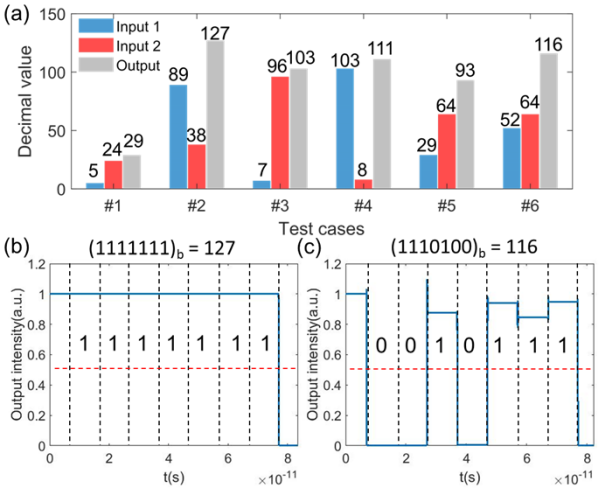


Figure. 3 Tested waveforms of optical RNNs (a) Results of some testing samples (b) Sequential full adder results (89+38=127) (c) Sequential full adder results (52+64=116). Both (a) (b) are simulated on Lumerical Interconnect.

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