Microring-Based Multi-Operand Optical Neurons with On-Chip Trainable Nonlinearity

Shupeng Ning,[1] Hanqing Zhu,[1], Ziang Yin,[2], Chenghao Feng,[1] Spencer Denton,[1] David Z. Pan,[1] Jiaqi Gu,[1,2], and Ray T. Chen[1,*]

[1] *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78758, USA*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA*

[*] *chenrt@austin.utexas.edu*

Abstract: We propose an optical neuron based on multi-operand microring resonators, featuring a compact footprint and high scalability with built-in tunable nonlinearity. Experimental results highlight its superior expressivity in machine learning tasks with exceptional hardware efficiency.

1. Introduction

Integrated photonics has emerged as a promising candidate for next-generation artificial intelligence (AI) accelerators, addressing the processing bottlenecks and high power consumption of traditional electronic computing platforms [1]. However, current photonic integrated circuits (PICs) and optical neural networks (ONNs) encounter several practical challenges, including a large footprint, lack of on-chip nonlinearity, and inefficient electrical-optical (E-O) interfaces. These limitations hinder their scalability and applicability to modern AI models [1, 2].

To address these challenges, research efforts have focused on enhancing the scalability of ONNs across the device [2, 3], circuit [4], and architecture levels [5]. In this study, we develop a novel microring-based multi-operand optical neuron (M^2OON) with a compact footprint and on-chip trainable nonlinearity. Experimental results validate the functionality and expressivity of M^2OON, and the M^2OON-based ONN achieves comparable or superior performance in machine learning tasks with fewer learnable parameters compared to traditional approaches.

2. Design and Working Mechanism

2.1. *Multi-operand microring resonators*

Compared with conventional microring resonators (MRRs), the multi-operand MRR has k active actuators controlled simultaneously by independent electrical signals. The transmission function of an add-drop multi-operand MRR at the through port can be expressed as $T_t = f(\phi_{rt}) = f\left(\sum_{i=0}^{k-1} g(w_i \cdot x_i)\right)$. Here, ϕ_{rt} represents the round-trip phase, and $g(w_i \cdot x_i)$ is determined by the weight/signal encoding mechanism (such as thermal tuning, free-carrier effect, etc.). This design squeezes a length-k dot-product within a single device and single wavelength ($k\times$ area/power/wavelength saving than MRR arrays [6]) and introduces on-chip nonlinearity by leveraging the intrinsic transmission characteristic of MRR.

2.2. *ONN architecture for M^2OON*

In this work, we propose an architecture to implement ONN on PICs using M^2OON. In Fig. 1.a, an $M \times N$ weight matrix is partitioned into $P \times Q$ blocks with a size of $k \times k$, which are then mapped onto a multi-operand MRR array. To achieve full-range matrix-vector multiplications (MVMs), the Q multi-operand MRRs for each row of the original matrix are split into positive and negative rails, respectively. A balanced photodetector pair is used to produce a full-range output. Furthermore, we introduce a learnable balancing factor $\boldsymbol{d}$ to scale each MORR's output range adaptively, enhancing the system's expressivity. The balancing factors are achieved via $Q/2$ serial MRRs operating at different wavelengths. Therefore, the output of each row can be expressed as $y_m = \sum f\left(\sum_{i=0}^{k-1} g(w_{mqi} \cdot x_{qi})\right) d_q$, where the sign of d_q depends on the polarity of the corresponding rail.

3. Results and Discussions

We taped out a photonic-electronic neural chip with thermo-optic M^2OON, which incorporates 4 operands and 2 microheaters for calibration and tunable nonlinearity (Fig. 1.b). For testing, we first calibrated each multi-operand MRR to its on-resonance state and then fitted the transmission curve using random length-4 vectors as input (Fig. 1.c). To highlight the unique nonlinear neural processing capability of M^2OON, we implemented a biomedical image processing task that applies nonlinear convolution with a Sobel kernel to a human brain MRI slice. Here, the inputs $\boldsymbol{x}$ and weights $\boldsymbol{w}$ are encoded by a digital-to-analog converter (DAC) and digital potentiometers (POTs), respectively. The inherent tunable nonlinearity of M^2OON allows for more effective feature extraction. As shown in Fig. 2.a, the feature map extracted by M^2OON effectively emphasizes the tumor borders, achieving significantly higher contrast in edge intensities compared to linear convolution, as evidenced by the pixel histogram (Fig. 2.c).

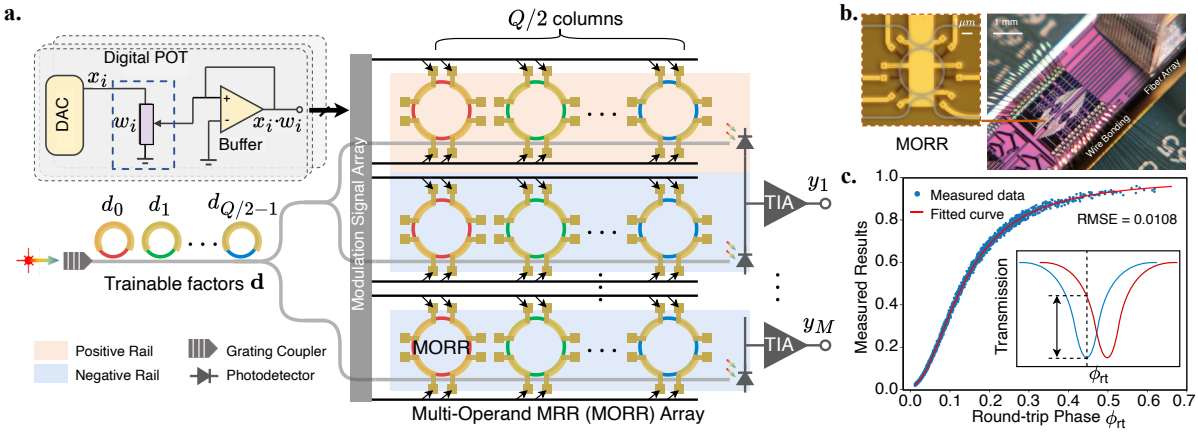

Fig. 1: (a) Schematic of M^{2}OON and proposed ONN architecture. (b) Micrographs of a multi-operand MRR and the M^{2}OON chip with optical and electrical packaging. (c) On-chip measured results and the fitted transmission curve. The inset shows the nonlinear tuning mechanism of multi-operand MRRs.

Additionally, we implemented binary classification tasks on both M^{2}OON and digital computers, using different network configurations and activation functions. Leveraging the enhanced expressivity of our learnable nonlinear neuron, M^{2}OON achieves 100% test accuracy in both tasks, even with a smaller, single-layer network and fewer parameters compared to conventional digital multi-layer perceptrons (Fig. 2.d).

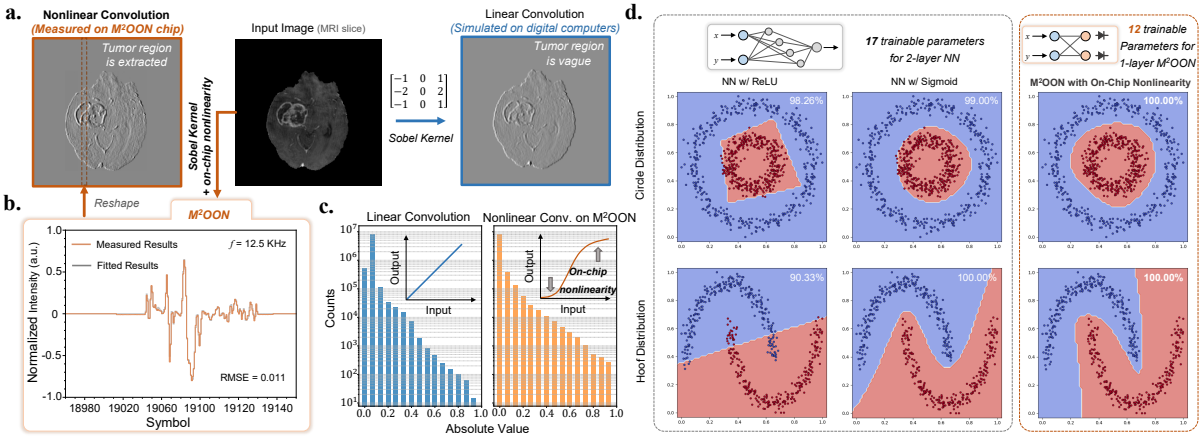

Fig. 2: Experimental results of on-chip image processing and the implementation of a classification task. (a) MRI image and the feature maps extracted using the Sobel kernel.(b) Experimental and ideal waveform of convolution results. (c) Count distribution of the absolute gray values of the feature maps obtained from linear convolution and M^{2}OON. The on-chip nonlinearity suppresses small values while emphasizing larger values for edges. (d) Results of the classification task using different activation functions.

4. Conclusions

This work presents a compact photonic neuron with on-chip trainable nonlinearity and a highly expressive ONN architecture, addressing scalability and nonlinearity challenges in photonic computing. M^{2}OON demonstrates superior performance in AI tasks, paving the way for scalable photonic AI accelerators.

References

1. S. Ning, H. Zhu, C. Feng, J. Gu, Z. Jiang, Z. Ying, J. Midkiff, S. Jain, M. H. Hlaing, D. Z. Pan *et al.*, "Photonic-electronic integrated circuits for high-performance computing and ai accelerators," J. Light. Technol. (2024).
2. C. Feng, J. Gu, H. Zhu, S. Ning, R. Tang, M. Hlaing, J. Midkiff, S. Jain, D. Z. Pan, and R. T. Chen, "Integrated multi-operand optical neurons for scalable and hardware-efficient deep learning," Nanophotonics **13**, 2193–2206 (2024).
3. Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," Nat. communications **13**, 2131 (2022).
4. C. Feng, J. Gu, H. Zhu, Z. Ying, Z. Zhao, D. Z. Pan, and R. T. Chen, "A compact butterfly-style silicon photonic–electronic neural chip for hardware-efficient deep learning," ACS Photonics **9**, 3906–3916 (2022).
5. A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein, D. Bunandar, P. B. Dixon, S. A. Hamilton, M. Streshinsky *et al.*, "Delocalized photonic deep learning on the internet's edge," Science **378**, 270–276 (2022).
6. A. N. Tait, M. A. Nahmias, B. J. Shastri *et al.*, "Broadcast and weight: An integrated network for scalable photonic spike processing," J. Light. Technol. (2014).