SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators

Jiaqi Gu[1], Chenghao Feng[1], Zheng Zhao[2], Zhoufeng Ying[3], Mingjie Liu[1], Ray T. Chen[1], and David Z. Pan[1]

[1]ECE Department, The University of Texas at Austin
[2]Synopsys, Inc. [3]Alpine Optoelectronics, Inc.
{jqgu, fengchenghao1996, zhengzhao, zfying, jay_liu}@utexas.edu, chen@ece.utexas.edu, dpan@ece.utexas.edu

***Abstract*— Optical neural networks (ONNs) have demonstrated promising potentials for next-generation artificial intelligence acceleration with ultra-low latency, high bandwidth, and low energy consumption. However, due to high area cost and lack of efficient sparsity exploitation, previous ONN designs fail to provide scalable and efficient neuromorphic computing, which hinders the practical implementation of photonic neural accelerators. In this work, we propose a novel design methodology to enable a more scalable ONN architecture. We propose a nonlinear optical neuron based on multi-operand ring resonators to achieve neuromorphic computing with a compact footprint, low wavelength usage, learnable neuron balancing, and built-in nonlinearity. The structured sparsity is exploited to support more efficient ONN engines via a fine-grained structured pruning technique. A robustness-aware learning method is adopted to guarantee the variation-tolerance of our ONN. Simulation and experimental results show that the proposed ONN achieves one-order-of-magnitude improvement in compactness and efficiency over previous designs with high fidelity and robustness.**

I. Introduction

Deep neural networks (DNNs) have shown record-breaking performance on various artificial intelligence tasks recently. However, traditional electrical digital computing platforms encounter substantial challenges to meet the escalating computation demands of DNNs in the post-Moore's era. As an emerging computing platform, optical neural networks (ONNs) demonstrate compelling potentials for neuromorphic computing with ultra-high speed, ultra-low latency, and low energy consumption. Prior work successfully demonstrated ONNs on silicon-based photonic integrated circuits (PIC). A Mach-Zehnder interferometer (MZI) based coherent ONN has been proposed to accelerate matrix multiplication using singular value decomposition [1], demonstrating ultra-high speed and over 100 GHz photo-detection rate. Based on MZI arrays, a recurrent ONN architecture has been demonstrated using waveguide feedback loops [2]. Since PIC designs currently show less competitive compactness than electrical digitals, recent ONN work mainly focuses on area reduction and scalability improvement. A slimmed ONN [3] was proposed to cut down the MZI usage through a software-hardware co-design methodology. Later, Fast-Fourier-transform-based integrated ONNs [4], [5] demonstrated a more compact design in the optical frequency-domain. Besides, incoherent ONNs have been recently explored to achieve more scalable designs using a smaller device, optical micro-ring resonator (MRR). An MRR-based ONN [6] has been demonstrated with a compact footprint using all-pass MRR weight banks and photonic digital arithmetic units [7]. The wavelength-division multiplexing (WDM) technique is leveraged to implement matrix multiplication in parallel with multiple wavelengths. A variant with a differential structure was proposed using add-drop MRRs to realize full-range weights for better model expressivity [8].

However, previous ONN designs still encounter scalability issues in the practical application. The state-of-the-art coherent ONNs have already demonstrated 2-4× [3], [4] area reduction compared to the original MZI-based ONN, but they generally still have a larger footprint than their incoherent counterparts. Since the physical dimension of one MRR is typically one-order-of-magnitude smaller than an MZI, MRR-based ONN is acknowledged as one of the most compact ONN architectures so far [6], [8]–[10]. MRR-ONNs are close to the current area lower bound of integrated ONN designs [6], [8]–[10], since an MRR is the smallest integrated photonic device that was previously used to achieve one multiplication, which makes it technically challenging for further compactness improvement by using traditional MRRs. Moreover, the high wavelength usage limits the scalability of MRR-ONNs since the maximum wavelengths supported by modern dense WDM (DWDM) techniques are still smaller than the practical matrix dimension, leading to inevitable area increase due to weight bank duplication or latency penalty from weight bank reuse. The robustness concerns and lack of sparsity exploitation also cast practicality issues for state-of-the-art MRR-ONNs.

To break the current compactness record of integrated silicon photonic neural networks, in this work, we propose a novel incoherent ONN architecture `SqueezeLight` that squeezes sparse structured matrices into ultra-compact multi-operand micro-ring resonators to enable scalable, efficient, and robust optical neurocomputing. The main contributions are as follows,

- Scalability: we propose a scalable ONN architecture based on multi-operand ring resonators with built-in nonlinearity and learnable neuron balancing, outperforming prior ONN designs by one order of magnitude in footprint.
- Efficiency: we explore the structured sparsity in our proposed architecture for quadratic efficiency boost through fine-grained structured pruning.
- Robustness: we propose a sensitivity-aware learning technique to overcome device variations and thermal crosstalk in our architecture.

978-3-9819263-5-4/DATE21/©2021 EDAA

II. Preliminaries

In this section, we introduce background knowledge about ONNs and our motivations.

A. Various Neural Network Designs

Classical convolutional neural networks (CNNs) perform discriminative representation and generalization via inner-product-based convolution. Various kernelized NNs [11] have been proposed as substitutions with competitive performance. Extensive linear and nonlinear convolution variants have been proposed for better robustness or computation efficiency, e.g., hyperbolic tangent convolution [12] and *AdderNet* [13]. In this work, we propose a nonlinear neuron that leverages the built-in nonlinearity of multi-operand ring resonators to achieve novel optical CNNs with comparable model expressivity.

B. Optical Neural Architectures

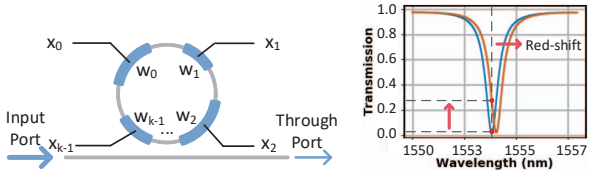
Recently, extensive researches have been done to investigate advanced ONN designs [1], [3]–[6], [8]–[10], [14], [15]. Mainstream linear operators in NNs, e.g., fully-connected layer and convolution, can be unified as general matrix multiplications and mapped to photonic circuits. Several coherent ONNs have been demonstrated for ultra-fast NN inference, e.g., MZI-based ONNs [1], [3] and FFT-based compact ONN [4], [5]. Incoherent ONNs based on MRR weight banks have been demonstrated with a compact footprint [6], [8]. However, the scalability of MRR-based ONNs is limited by their weight bank size and high wavelength usage. In this work, we propose a more compact architecture with a lower device and wavelength usage to break through the ONN scalability bound.

C. Multi-Operand Ring Resonators

Recently, a multi-operand electro-optic logic gate (MOLG) has been proposed to achieve multi-operand boolean functions on a single micro-ring resonator, enabling ultra-compact optical digital computing [16]. The all-pass multi-operand ring resonator (MORR) is shown in Fig. 1a. k independent active thermal actuators along the MORR are simultaneously controlled by k electrical signals x, each inducing a phase shift $\phi_i(x_i)$. The accumulated round-trip phase shift causes a spectrum redshift $\Delta\lambda$, such that the transmitted intensity on the through port changes accordingly. Figure 1b demonstrates the transmission spectrum of an all-pass MORR. Weighted phase shifts can be achieved by different actuator arm lengths, different input ranges, or different materials, etc [16]. The transfer function of a k-operand all-pass MORR is,

$$y = \left| \frac{r - ae^{-j\phi}}{1 - rae^{-j\phi}} \right|^2 d, \quad \phi = \sum_{i=0}^{k-1} \phi_i(x_i), \quad \phi_i(x_i) \propto w_i x_i^2, \tag{1}$$

where x_i is the electrical input voltage, w_i is the corresponding weight for the i-th input, $\phi_i(\cdot)$ is the quadratic phase shift response curve of the thermal actuator, ϕ is the accumulated round-trip phase shift of the MORR, r and a are self-coupling coefficient and single-pass amplitude transmission factor, and $d, y \in [0, 1]$ are the light intensity on the input port and through port, respectively. Instead of using it as a digital logic gate, we introduce a novel ONN architecture that leverages MORRs in analog neuromorphic computing.



(a) (b)

Fig. 1: (a) Structure of an all-pass k-operand MORR. (b) Through port intensity transmission of an all-pass MORR.

III. Proposed Optical Neural Network Architecture

In this section, we discuss details on the proposed ONN architecture shown in Fig. 2 and several essential techniques for scalability and efficiency improvement.

A. MORR-based Nonlinear Neuron

Different from previous matrix multiplication based ONN designs that only focus on linear projection, we introduce a novel ONN architecture based on an ultra-compact MORR to perform efficient analog neural computing. To leverage the nonlinear transmission equation of MORRs in Eq. (1), we first align the idle device to an on-resonance state, where the transmitted light intensity is close to 0. Then we abstract the computational model of the MORR as follows,

$$y = f(\sum_{i=0}^{k-1} \phi_i)d \propto f\Big(\sum_{i=0}^{k-1} w_i x_i^2\Big)d, \quad \text{s.t. } \mathrm{w_i} \geq 0 \tag{2}$$

where $f(\cdot)$ represents the built-in nonlinear $y - \phi$ curve given by Eq. (1). Note that within the practical wavelength range, the shape of the transmission curve keeps almost identical at different wavelengths [17], thus we are justified to assume the same nonlinear curve $f(\cdot)$ for all MORRs. Different from the traditional micro-ring resonator (MRR) that encodes a single weight on its transmission factor with one wavelength, this MORR directly performs length-k vector computation with one device, one wavelength, and direct electrical inputs.

Based on the above MORR neuron, we propose a novel ONN architecture shown in Fig. 2. Our architecture starts with a single laser input and an on-chip frequency comb to generate multiple wavelengths $(\lambda_0, \lambda_1, \cdots)$. Then, a series of narrow-band MRR modulators are used to perform wavelength-wise scaling $\boldsymbol{D} = (d_0, \cdots, d_{Q/2-1}) \in [0, 1]$ to achieve an adaptive dynamic range of MORRs. A WDM multiplexer is used to evenly distribute the WDM light into $2M$ rows. The main part is a $2M \times \frac{N}{2k}$ array that implements an ONN layer with built-in nonlinearity and neuron balancing. For an $M \times N$ weight matrix $\boldsymbol{W}$, there are total $2M$ rows and $\frac{Q}{2} = \frac{N}{2k}$ columns in the MORR array. The q-th MORR in one row will resonate at the wavelength λ_q and perform nonlinear projection on its length-k input as $y_q = f(\sum_{i=0}^{k-1} w_{qi} x_{qi}^2)d_q$. At the end of the m-th row,

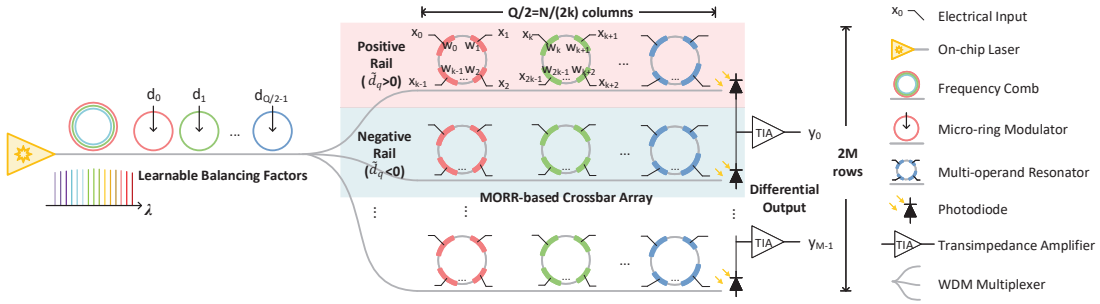


Fig. 2: Proposed MORR-based ONN architecture `SqueezeLight` with learnable neuron balancing.

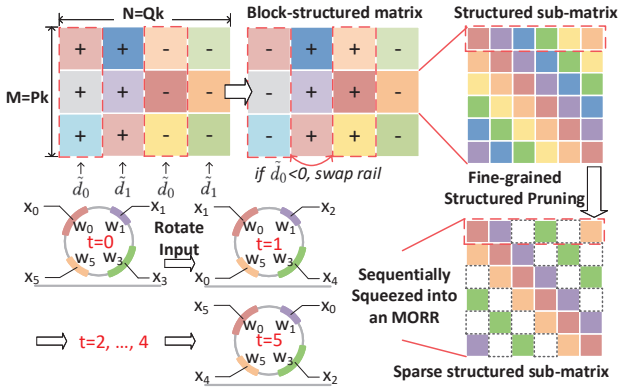


Fig. 3: Block-squeezing based on block-structured matrices. A sparse block can be squeezed into one MORR with fine-grained structured pruning.

a photo-detector is used to generate photo-current based on the accumulated light intensity as $I_m = \sum_{q=0}^{Q/2-1} y_{mq}$.

Note that non-negative weights will limit the solution space, thus we introduce a learnable neuron balancing structure to enhance the crossbar array with a differential structure shown in Fig. 2. Adjacent rows are designed to be the positive rail I^+ and negative rail I^- respectively. The differential photo-current structure at the end is able to generate full-range of outputs,

$$y_m = G(I_m^+ - I_m^-) = G\Big(\sum_{q=0}^{Q/2-1} y_{mq} - \sum_{q=Q/2-1}^{Q-1} y_{mq}\Big), \quad (3)$$

where G is the gain of the transimpedance amplifier (TIA), which can be used to extend the signal range. Note that if $d = 1$, all MORRs will have the same importance, i.e., $y_{mq} \in [0, 1], \forall q$. To further boost the representability of our MORR neuron, we adaptively set the dynamic range of each MORR by learning the balancing factors $\tilde{\boldsymbol{D}} = \{\tilde{d}_q | \tilde{d}_q \in [-G_{max}, G_{max}], \tilde{d}_q = \tilde{d}_{q\ (mod)\ \frac{Q}{2}},\ q \in [0, Q-1]\}$, shared by a column of MORRs,

$$y_m = \sum_{q=0}^{Q-1} f\Big(\sum_{i=0}^{k-1} w_{mqi} x_{qi}^2\Big)\tilde{d}_q. \quad (4)$$

The maximum TIA gain G_{max} expands the implementable range to $\tilde{d} \in$[-G_{max}, G_{max}]. Though all-pass MRR modulators can only implement non-negative scaling, a negative factor $\tilde{d} < 0$ can be realized by simply swapping the rails of corresponding MORRs as shown in Fig. 3. This learnable neuron balancing technique enables adaptive rail assignment and trainable dynamic range for different MORRs, boosting the expressivity of our architecture with balanced output distributions and weighted MORR importance. Note that as a side-effect, we can also save another $2\times$ wavelength usage through splitting one length-Q row into two length-$\frac{Q}{2}$ rails.

B. Peripheral Units

We briefly discuss the entire dataflow and peripheral units, with all system-level details being omitted, but we clarify that advanced system-level innovations should be applicable to our architecture as well [6], [9].

1) Normalization: Normalization layers, e.g., BatchNorm, can be simply implemented by the TIA gain and voltage signal offset without causing further latency penalty.

2) Nonlinear Activation: Since our proposed neuron has built-in nonlinearity, we do not adopt extra activations.

3) Electrical Dataflow: The input signals/weights are loaded from high-bandwidth SRAM or ultra-fast photonic racetrack memory banks [18], and converted to analog signals through electrical digital-to-analog converters (DACs). The detected results are amplified by TIAs. Direct optical-electrical-optical (O-E-O) conversions will be used to cascade multiple ONN layers without voltage-to-transmission encoding.

C. Area Reduction via Block-Squeezing

To achieve a quadratically more compact design, we adopt a block-squeezing method to further improve the scalability. Motivated by structured neural networks that demonstrate comparable model expressivity and better efficiency than classic NNs, we introduce this concept to `SqueezeLight` for higher compactness. Figure 3 visualizes the mapping from 6×6 structured blocks to our MORR array. An $M \times N$ block-structured matrix $\boldsymbol{W}$ consists of $P \times Q$ square sub-matrices $\{w_{pq}\}_{p,q=0}^{P,Q}$, each being a $k \times k$ structured matrix. A $k \times k$ circulant matrix is defined by a length-k primary vector on its first row. Thus each partial projection $\boldsymbol{w}_{pq} \cdot x_q$ can be efficiently implemented by reusing one k-operand MORR for k times with rotated inputs. In this way, we successfully achieve $O(k^2)$ times area reduction and k times wavelength usage reduction.

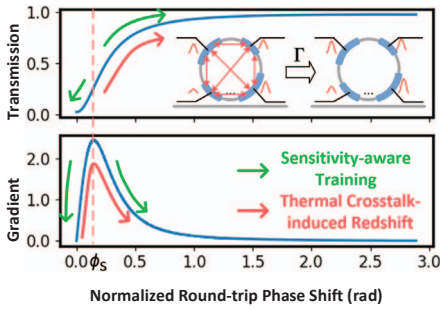


Fig. 4: Transmission curve f and its gradient $\nabla_\phi f$ with thermal crosstalk and sensitivity-aware training.

D. Sparsity Exploration via Fine-Grained Structured Pruning

For an $M \times N$ matrix with $k \times k$ structured blocks, the total components add up to $\frac{N}{2}$ MRRs and $(\frac{MN}{k^2})$ k-operand MORRs. Given fixed M and N, a larger k leads to fewer blocks and less MORR usage. However, it could be challenging for manufacturing such MORRs with too many actuator sections. To overcome this, we explore a fine-grained structured sparsity by pruning each sub-matrix with a uniform sparsity. In Fig. 3, entries that are less important in the primary vector are forced to zero, with k' weights left. The matrix structure will automatically prune corresponding entries in other rows. Once the sub-matrix is larger than the MORR capacity, i.e., $k > k_{max}$, this technique can maintain the highest compactness by sequentially squeezing the pruned block into one MORR with $k' \leq k_{max}$. A two-stage pruning procedure with learning rate rewinding is described in Alg. 1. After pre-training, the weights are pruned with a target sparsity. Then the model will be trained from scratch with a rewound learning rate to achieve better accuracy than traditional post-training fine-tuning.

E. Robustness Boost via Sensitivity-Aware Optimization

The major non-ideal effects of MORRs come from random variations and deterministic intra-MORR thermal crosstalk. The random variations can be modeled as a Gaussian noise on the phase shift $\Delta\phi \in \mathcal{N}(0, \sigma^2)$. The intra-MORR crosstalk among adjacent k thermal actuators can be formulated as $\hat{\mathbf{\Phi}} = \mathbf{\Gamma} \cdot \mathbf{\Phi}$ governed by a coupling matrix $\mathbf{\Gamma}$ as follows,

$$\begin{pmatrix} \hat{\phi}_0 \\ \hat{\phi}_1 \\ \cdots \\ \hat{\phi}_{k-1} \end{pmatrix} = \begin{pmatrix} \gamma_{0,0} & \gamma_{0,1} & \cdots & \gamma_{0,k-1} \\ \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k-1,0} & \gamma_{k-1,1} & \cdots & \gamma_{k-1,k-1} \end{pmatrix} \begin{pmatrix} \phi_0 \\ \phi_1 \\ \cdots \\ \phi_{k-1} \end{pmatrix}, \quad (5)$$

Without loss of generality, we assume the self-coupling factor $\gamma_{i,i} = 1$ and all mutual coupling factors $\gamma_{i,j}$ share the same value γ for efficient estimation. The above crosstalk can be simplified as a constant scaling factor on ϕ as follows,

$$\hat{y}_m = \sum_{q=0}^{Q-1} f\Big((1 + (k'-1)\gamma)\phi_{mq} + \Delta\phi\Big)\tilde{d}_q. \quad (6)$$

Only k' actuators have crosstalk after pruning, thus our pruning can improve the robustness by reducing the noise sources. Figure 4 shows that the transmission curve $f(\cdot)$ has different sensitivity (gradient) on different wavelengths. Crosstalk variations induce an extra redshift in the spectrum, forcing all $\phi < \phi_s$ to have higher sensitivity and $\phi \geq \phi_s$ to have less sensitivity. To improve the robustness, we introduce a sensitivity-aware optimization method, shown in Alg. 1, to model the variations when training an L-layer `SqueezeLight` with the objective,

$$\mathcal{L} = \mathcal{L}_0(x; \boldsymbol{W}, \tilde{\boldsymbol{D}}, \mathbf{\Gamma}, \Delta\phi) + \alpha \sum_{l,m,q=0}^{L-1,M-1,Q-1} \nabla_\phi f(\hat{\phi}_{lmq} + \Delta\phi), \quad (7)$$

where $\mathcal{L}_0(x; \boldsymbol{W}, \tilde{\boldsymbol{D}}, \mathbf{\Gamma}, \Delta\phi)$ is the task-specific loss with noise injection, and the second term, denoted as $\mathcal{L}_S(\mathbf{\Gamma}, \Delta\phi)$, is a sensitivity-aware penalty term weighted by α. This method jointly considers variations and crosstalk with a gradient-based sensitivity penalty, enabling close-to-ideal test accuracy.

Algorithm 1 Training algorithm of `SqueezeLight` with fine-grained structured pruning and sensitivity-aware optimization.

Input: Initial weights $\boldsymbol{W}^0 \in \mathbb{R}^{P\times Q\times k}$ and $\tilde{\boldsymbol{D}}^0 \in \mathcal{R}^{Q/2}$, pruning percentage $T = 1 - \frac{k'}{k}$, pretraining step t_{pre}, initial step size η^0, decay factor β, penalty weight α, variation $\Delta\phi$, and crosstalk coupling matrix $\mathbf{\Gamma}$;
Output: Converged $\boldsymbol{w}^t$, $\boldsymbol{d}^t$, and a pruning mask $\mathcal{M} \in \mathbb{Z}^{P\times Q\times k}$;

for $t \leftarrow 1, \cdots, t_{pre}$ **do** ▷ Stage 1: Pretraining
  $\mathcal{L} \leftarrow \mathcal{L}_0^t(x; \boldsymbol{W}^{t-1}, \tilde{\boldsymbol{D}}^{t-1})$
  $(\boldsymbol{W}^t, \tilde{\boldsymbol{D}}^t) \leftarrow (\boldsymbol{W}^{t-1}, \tilde{\boldsymbol{D}}^{t-1}) - \eta^{t-1}(\nabla_{\boldsymbol{W}}\mathcal{L}, \nabla_{\tilde{\boldsymbol{D}}}\mathcal{L})$
  $\eta^t \leftarrow \eta^{t-1}\beta$ ▷ Learning rate decay
$\eta^t \leftarrow \eta^0, \mathcal{M} \leftarrow 1$ ▷ Learning rate rewinding and initialize mask
for all $\boldsymbol{W}^t_{pqi} \in \boldsymbol{W}^t$ **do**
  if $\boldsymbol{W}^t_{pqi} <$ `percentile`$(\boldsymbol{W}^t_{pq}, T)$ **then**
    $\mathcal{M}_{pqi} \leftarrow 0$ ▷ Compute pruning mask
while not converged **do** ▷ Stage 2: Fine-grained pruning
  $\mathcal{L} \leftarrow \mathcal{L}_0^t(x; \mathcal{M} \odot \boldsymbol{W}^{t-1}, \tilde{\boldsymbol{D}}^{t-1}, \mathbf{\Gamma}, \Delta\phi) + \alpha\mathcal{L}_S(\mathbf{\Gamma}, \Delta\phi)$ ▷ Sensitivity-aware regularization
  $(\boldsymbol{W}^t, \tilde{\boldsymbol{D}}^t) \leftarrow (\boldsymbol{W}^{t-1}, \tilde{\boldsymbol{D}}^{t-1}) - \eta^{t-1}(\nabla_{\boldsymbol{W}}\mathcal{L}, \nabla_{\tilde{\boldsymbol{D}}}\mathcal{L})$
  $\eta^t \leftarrow \eta^{t-1}\beta$ ▷ Learning rate decay

IV. Hardware Feasibility and Efficiency

We give a theoretical analysis of the hardware feasibility and efficiency, and compare essential features with previous ONNs.

A. MORR Physical Feasibility

Our MORR leverages the analog property of a successfully demonstrated digital MOLG [16]. We solve the possible resonator sensitivity issue by using a low-quality-factor (low-Q) MORR filter. Typical MRR filters have a Q value of 5,000 [6], [8], [10], while in our case, we design the MORR with a low Q value of ∼2,000 and a large wavelength tuning range of ∼4 nm. Therefore, the controllability will not be a concern.

B. Area, Latency, and Power

In Table I, our architecture outperforms those three previous coherent ONNs by a large margin [1], [3], [4]. Therefore, we focus on the comparison of area cost $\mathcal{A}$, latency τ, and power $\mathcal{P}$ with the most compact designs MRR-ONN-1 [6] and MRR-ONN-2 [8] in Table I. We assume the current DWDM capacity is B, representing the maximum number of wavelength available [19], [20]. First, the size and power of an MRR and a k-operand MORR can be assumed the

TABLE I: Hardware cost and feature comparison. The matrix is $M \times N$ with size-k blocks. B is the DWDM capacity. For fair comparison, the device counts are converted to #MRRs based on real device sizes [1], [4], [21]. The area ratio β_a and power ratio β_p between one MZI (240×40 μm^2 [1], $\sim$48mW [21]) and one MRR (20×20 μm^2,$\sim$4mW [17]) are β_a=24 and β_p=12.

	MZI-ONN [1]	Slim-ONN [3]	FFT-ONN [4]	MRR-ONN-1 [6]	MRR-ONN-2 [8]	SqueezeLight
#MRRs	$\beta_a MN$	$\sim \frac{\beta_a}{2} MN$	$\sim \frac{\beta_a}{4} MN$	$M \min(N, B)$	$M \min(N, B)$	$\frac{2M}{k} \min(\frac{N}{2k}, B)$
#Wavelength	1	1	1	$\min(N, B)$	$\min(N, B)$	$\min(\frac{N}{2k}, B)$
Latency	1	1	1	$\lceil \frac{N}{B} \rceil$	$\lceil \frac{N}{B} \rceil$	$k\lceil \frac{N}{2kB} \rceil$
Power	$\beta_p MN$	$\sim \frac{\beta_p}{2} MN$	$\sim \beta_p MN$	$M \min(N, B)$	$M \min(N, B)$	$\frac{2M}{k} \min(\frac{N}{2k}, B)$
Nonlinearity	Electrical	Electrical	Electrical	Electrical	Electrical	Built-in
Output range	Non-negative only	Non-negative only	Non-negative only	Non-negative only	Full range	Full range
Control complexity	High	Medium-High	High	High	High	Medium

same since they have the same phase tuning range, i.e., half of the resonance curve. Therefore, we focus on the number of resonators in the discussion. We denote the computation efficiency as $\mathcal{E} = (\mathcal{A}\mathcal{P}\tau)^{-1}$. SqueezeLight achieves the following improvement over two MRR-ONNs when the matrix dimension is smaller than the DWDM capacity, i.e., $N < B$,

$$\frac{\mathcal{A}_{ours}}{\mathcal{A}_{prev}} \approx \frac{\mathcal{P}_{ours}}{\mathcal{P}_{prev}} \approx \frac{1}{k^2}, \ \frac{\tau_{ours}}{\tau_{prev}} = \frac{k\lceil N/B \rceil}{\lceil N/(2kB) \rceil} = k, \ \frac{\mathcal{E}_{ours}}{\mathcal{E}_{prev}} \approx k^3. \quad (8)$$

Once the matrix width is larger than the maximum number of wavelengths available as $\frac{N}{2k} < B < N$, we can achieve,

$$\frac{\mathcal{A}_{ours}}{\mathcal{A}_{prev}} \approx \frac{\mathcal{P}_{ours}}{\mathcal{P}_{prev}} < \frac{2}{k}, \ \frac{\tau_{ours}}{\tau_{prev}} = \frac{k}{\lceil \frac{N}{B} \rceil}, \ \frac{\mathcal{E}_{ours}}{\mathcal{E}_{prev}} \approx \frac{Bk^3}{N} > \frac{k^2}{2}. \quad (9)$$

If the weight matrix is even larger, i.e., $B < \frac{N}{2k}$, we have

$$\frac{\mathcal{A}_{ours}}{\mathcal{A}_{prev}} \approx \frac{\mathcal{P}_{ours}}{\mathcal{P}_{prev}} \approx \frac{2}{k}, \ \frac{\tau_{ours}}{\tau_{prev}} \approx \frac{1}{2}, \ \frac{\mathcal{E}_{ours}}{\mathcal{E}_{prev}} \approx \frac{k^2}{2}, \text{if } B < \frac{N}{2k}. \quad (10)$$

It can be observed that our ONN gains more hardware efficiency advantage as B scales up, thus our scalability grows together with the development of the DWDM technology.

C. Feature Comparison

Several essential features are compared in Table I. Previous ONNs only focus on matrix multiplication, with nonlinearity being offloaded to the electrical domain. In contrast, our proposed neuron leverages the built-in nonlinearity in MORRs to eliminate the overhead from electrical activation, enabling higher speed and efficiency. In terms of model expressivity, MRR-ONN-1 [6] has a limited solution space with only positive weights, while our designs support exploration in an augmented parameter space via learnable neuron balancing. Our ONN also shows lower control complexity and higher efficiency via direct signal encoding $v_x = x$, while previous MRR-ONNs require additional mapping to transform the encoded inputs/weights into voltage signals $v_x = \sqrt{\phi^{-1}(f^{-1}(x))}$.

V. Experimental Results

We conduct optical simulation for functionality validation and comparison experiments on MNIST , FashionMNIST (FMNIST) , and Cifar-10 dataset . We implement all models in PyTorch and evaluate the accuracy on a machine with an Intel Core i7-9700 CPU and an NVIDIA Quadro RTX 6000 GPU. All ONNs are trained for 100 epochs using the Adam optimizer. Quantization-aware training [22] is applied to perform 8-bit weight/input/activation quantization on all models.

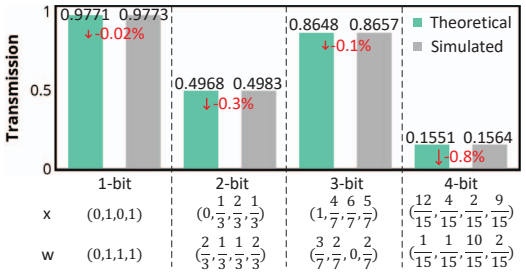


Fig. 5: Comparison between theoretical and simulated results.

A. Fidelity Validation via Optical Simulation

To validate the functionality of the MORR-based neuron, we perform optical simulation using the commercial Lumerical INTERCONNECT tool. Figure 5 plots the theoretical and simulated outputs of a 4-operand MORR under 1- to 4-bit precision. The derived neuron model has a high fidelity with <1% relative error compared with simulation results.

B. Comparison Experiments

We compare the inference accuracy among four ONNs: 1) MRR-ONN-1 with all-pass MRRs, 2) MRR-ONN-2 with add-drop MRRs, and 3) our proposed architecture without pruning (Ours). In all dataset and ONN settings, SqueezeLight achieves comparable test accuracy with 20-30× fewer resonators, 8× lower wavelength usage, and ∼80% fewer parameters. We also evaluate our architecture with low-bit quantization. Even binarized SqueezeLight can achieve >95% accuracy on MNIST with the *large* model, and >98% accuracy can be achieved on 2-8 bit precision.

C. Fine-Grained Structured Pruning

In Table III, the pruned architecture reduces the manufacturing and control complexity as the sparse sub-matrices $k' = 4$ can be implemented only 4-operand MORRs, with no accuracy loss. Moreover, memory efficiency is also improved due to an extra 30% parameter reduction. This enables us to squeeze larger blocks into one MORR with a regular sparsity for better scalability with negligible accuracy loss.

D. Variation-Robustness Evaluation

In Fig. 6, we evaluate the variation-robustness on 1) MRR-ONN-1, 2) MRR-ONN-2, 3) our unpruned architecture (Ours), 4) our pruned architecture (Ours-P), and 5) ours with pruning

TABLE II: Accuracy and hardware cost comparison. *small* model is C32K5S2-BN-C32K5S2-BN-F10, where *C32K5S2* is 5×5 convolution with 32 kernels and stride 2, *BN* is BatchNorm, and *F10* is a linear layer. *large* model is C64K5S2-BN-C64K5S2-BN-F10. We use $k = 8$ in convolutional layers and $k = 4$ in the final classifier. *#Device*, *#λ*, and *#Param* are the number of used resonators, wavelengths, and parameters, respectively. Normalized ratios are shown in the parenthesis. All models are trained with 8-bit weight/input/activation quantization.

Dataset	Model	MRR-ONN-1 [6]				MRR-ONN-2 [8]				Ours			
		Test Acc.	#Device	#λ	#Param	Test Acc.	#Device	#λ	#Param	Test Acc.	#Device	#λ	#Param
MNIST	small	97.81	39.90 K (23.86)	1152(8)	38 K	98.55	39.90 K (23.86)	1152(8)	38 K	98.01	1.67 K (1.00)	144(1)	8 K
MNIST	large	97.89	130.97 K (31.64)	2304(8)	127 K	98.84	130.97 K (31.64)	2304(8)	127 K	98.36	4.14 K (1.00)	288(1)	22 K
FMNIST	small	86.97	39.90 K (23.86)	1152(8)	38 K	89.52	39.90 K (23.86)	1152(8)	38 K	86.65	1.67 K (1.00)	144(1)	8 K
FMNIST	large	87.75	130.97 K (31.64)	2304(8)	127 K	90.30	130.97 K (31.64)	2304(8)	127 K	87.21	4.14 K (1.00)	288(1)	22 K
Cifar-10	large	48.79	143.37 K (28.50)	3136(8)	139 K	61.69	143.37 K (28.50)	3136(8)	139 K	58.29	5.03 K (1.00)	392(1)	26 K

TABLE III: Fine-grained structured pruning evaluation. *#8op* represents the number of 8-operand MORRs. *Ours-P* represents all convolutional layers are pruned from k=8 to k'=4.

Dataset	Model	Ours				Ours-P			
		Acc.	#8op	#4op	#Param	Acc.	#8op	#4op	#Param
MNIST	small	98.01	416	864	8 K	98.02	0	1280	6 K
MNIST	large	98.36	1632	1728	22 K	98.58	0	3360	16 K
FMNIST	small	86.65	416	864	8 K	86.50	0	1280	6 K
FMNIST	large	87.21	1632	1728	22 K	87.36	0	3360	16 K
Cifar-10	large	58.29	1680	2352	26 K	60.52	0	4032	19 K

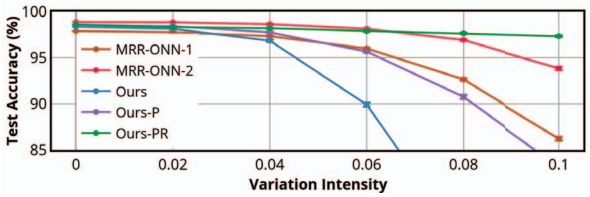


Fig. 6: Robustness evaluation of the *large* model on MNIST. The error bar indicates $\pm 1\sigma$ variance on 20 runs. For example, 0.04 means γ=0.04 and std. $\Delta\phi$=0.04. *Ours-PR* represents pruned model with sensitivity-aware training (α=0.02).

and robustness-aware training (Ours-PR). With the additional intra-MORR crosstalk, our ONN shows lower accuracy than other MRR-ONNs if pruning and nonideality modeling is not performed. When fine-grained structured pruning is applied, the crosstalk sources are cut down from $k = 8$ to $k' = 4$, achieving improved noise-tolerance. With sensitivity-aware training based on Eq. (7), the test accuracy maintains above 97% with a small variance which is reasonably close to the ideal accuracy, while other ONNs show a rapidly-degrading trend as the noise intensity increases. Therefore, our proposed architecture guarantees reliable inference even under practical non-ideal variations by using our lightweight robustness-aware training.

VI. Conclusion

In this work, we propose a novel ONN architecture `SqueezeLight` to break the compactness record of previous designs with higher scalability and efficiency. An MORR-based ultra-compact optical neuron is demonstrated with learnable neuron balancing and built-in nonlinearity. A block-squeezing technique with fine-grained structured pruning is proposed to achieve a quadratically more compact ONN design. Our proposed sensitivity-aware training method enables close-to-ideal neural computing with high robustness. We give a theoretical analysis to show the scalability and efficiency advantage of our ONN design. Experiments show that `SqueezeLight` provides a practical solution for ONNs with high accuracy and 20-30× better scalability and efficiency than prior designs.

Acknowledgment

The authors acknowledge the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR), contract No. FA 9550-17-1-0071, monitored by Dr. Gernot S. Pomrenke.

References

[1] Y. Shen, N.C. Harris, S. Skirlo *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, 2017.

[2] C. Feng, Z. Zhao, Z. Ying *et al.*, "Compact design of On-chip Elman Optical Recurrent Neural Network," in *Proc. CLEO*, 2020.

[3] Z. Zhao, D. Liu, M. Li *et al.*, "Hardware-software co-design of slimmed optical neural networks," in *Proc. ASPDAC*, 2019.

[4] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards area-efficient optical neural networks: an FFT-based architecture," in *Proc. ASPDAC*, 2020.

[5] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE TCAD*, 2020.

[6] W. Liu, W. Liu, Y. Ye *et al.*, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *Proc. DATE*, 2019.

[7] Z. Ying, C. Feng, Z. Zhao *et al.*, "Electronic-photonic arithmetic logic unit for high-speed computing," *Nature Communications*, 2020.

[8] A.N. Tait, T.F. de Lima, E. Zhou *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, 2017.

[9] F. Zokaee, Q. Lou, N. Youngblood *et al.*, "LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks," in *Proc. DATE*, 2020.

[10] M. Miscuglio and V.J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Review*, 2020.

[11] J. Mairal, P. Koniusz, Z. Harchaou *et al.*, "Convolutional Kernel Networks," in *Proc. NeurIPS*, 2014.

[12] W. Liu, Z. Liu, Z. Yu *et al.*, "Decoupled Networks," in *Proc. CVPR*, 2018.

[13] Y. Chen, X. Dai, M. Liu *et al.*, "Dynamic convolution: Attention over convolution kernels," in *Proc. CVPR*, 2020.

[14] J. Gu, Z. Zhao, C. Feng *et al.*, "ROQ: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *Proc. DATE*, 2020.

[15] J. Gu, Z. Zhao, C. Feng *et al.*, "FLOPS: Efficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization," in *Proc. DAC*, 2020.

[16] Z. Ying, C. Feng, Z. Zhao *et al.*, "Integrated multi-operand electro-optic logic gates for optical computing," *Appl. Phys. Lett.*, 2019.

[17] E. Timurdogan, Z. Su, C.V. Poulton *et al.*, "AIM Process Design Kit (AIMPDKv2.0): Silicon Photonics Passive and Active Component Libraries on a 300mm Wafer," in *Proc. IEEE OFC*, 2018.

[18] Z. Sun, W. Wu, and H.H. Li, "Cross-Layer Racetrack Memory Design for Ultra High Density and Low Power Consumption," in *Proc. DAC*, 2013.

[19] D.T.H. Tan, A. Grieco, and Y. Fainman, "Towards 100 channel dense wavelength division multiplexing with 100GHz spacing on silicon," *Opt. Express*, 2014.

[20] J. Yu and X. Zhou, "Ultra-High-Capacity DWDM transmission system for 100G and beyond," *IEEE Communications Magazine*, 2010.

[21] N.C. Harris *et al.*, "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Opt. Express*, 2014.

[22] S. Zhou, Z. Ni, X. Zhou *et al.*, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv:1606.06160*, 2016.