O^{2}NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands

Jiaqi Gu[1], Zheng Zhao[2], Chenghao Feng[1], Zhoufeng Ying[3], 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, zhengzhao, fengchenghao1996, zfying}@utexas.edu, chenrt@austin.utexas.edu, dpan@ece.utexas.edu

***Abstract*— Optical neuromorphic computing has demonstrated promising performance with ultra-high computation speed, high bandwidth, and low energy consumption. The traditional optical neural network (ONN) architectures realize neuromorphic computing via electrical weight encoding. However, previous ONN design methodologies can only handle static linear projection with stationary synaptic weights, thus fail to support efficient and flexible computing when both operands are dynamically-encoded light signals. In this work, we propose a novel ONN engine O^{2}NN based on wavelength-division multiplexing and differential detection to enable high-performance, robust, and versatile photonic neural computing with both light operands. Balanced optical weights and augmented quantization are introduced to enhance the representability and efficiency of our architecture. Static and dynamic variations are discussed in detail with a knowledge-distillation-based solution given for robustness improvement. Discussions on hardware cost and efficiency are provided for a comprehensive comparison with prior work. Simulation and experimental results show that the proposed ONN architecture provides flexible, efficient, and robust support for high-performance photonic neural computing with fully-optical operands under low-bit quantization and practical variations.**

I. Introduction

Deep neural networks (DNNs) demonstrate record-breaking performance on various applications in recent years. However, their escalating model scales and computation demands cast substantial technical challenges to traditional electrical digital computing platforms. As a compelling alternative, optical neural networks (ONNs) have attracted increasing attention with ultra-high speed, ultra-low latency, and low energy consumption, which provide a promising next-generation artificial intelligence (AI) acceleration platform. Previous research successfully demonstrated ONNs with silicon-based integrated photonic circuits. Shen *et al.* [1] proposed to use singular value decomposition to decompose weight matrices and map them onto cascaded Mach-Zehnder interferometer (MZI) arrays to achieve matrix-vector multiplication. This coherent ONN demonstrates ultra-low inference latency with an over 100 GHz photo-detection rate. However, this architecture depends on complicated matrix decomposition with a large area cost and high control complexity [1], [2]. A slimmed ONN [2] was proposed to improve the hardware efficiency through a software-hardware co-design methodology. Fast-Fourier-transform-based ONNs [3], [4] were proposed to further reduce the ONN area cost by mapping neurocomputing onto the optical frequency-domain, achieving a smaller footprint but lacks the flexibility to support general matrix multiplication. Apart from these coherent photonic NN designs, ONN architectures based on photonic adders [5], wavelength-division multiplexing (WDM) techniques [6], and optical microring (MR) resonators were proposed to implement incoherent ONNs [7]–[9]. Those ONNs focus on dot-product computation and encode trained weights by tuning the configurable MR resonators (MRRs) to modulate the magnitude of optical signals with different wavelengths. Though MRR-ONNs [7], [8] have an advantage in circuit footprint and power consumption, they are unable to support linear dot-product between two optically-encoded matrices and are also noise-prone due to MRR weight bank sensitivity issues.

However, the previous electrical-weight-based design methodology potentially limits the application range of ONNs to accelerate modern advanced DNNs. Prior ONN designs do not have the capability to support robust and efficient computing with both operands being dynamically-encoded optical signals, which includes essential operations in attention-based models [10] and advanced NNs with dynamically-generated weights [11]. Moreover, fully-optical operands can potentially benefit ONN on-chip training and online learning applications with frequent and high-speed weight updating [12], [13]. In terms of robustness, previous MZI-based ONN architectures encounter nontrivial accuracy degradation under low-bit signal quantization and practical device variation [14], [15], lacking compatibility with modern neural compression techniques.

In this paper, we propose a new ONN architecture O^{2}NN to enable high-performance and versatile photonic neuromorphic computing. We present a WDM-based differential dot-product unit with augmented and balanced optical weights as the core engine. The main contributions and key features are as follows,

- Flexibility: we propose a novel ONN architecture based on WDM and differential detection to enable dynamic neural computing between two fully-optical operands.
- Expressivity: we introduce extended optical weights and augmented quantization to improve the model expressivity.
- Robustness: we given a comprehensive analysis on the variation-robustness of our photonic core and provide an effective solution to improve the computational fidelity with knowledge-distillation-based noise-aware training.

978-3-9819263-5-4/DATE21/©2021 EDAA

II. Preliminaries

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

A. Neural Networks with General Matrix Multiplication

Modern neural networks extensively adopt fully-connected layers and convolutional layers to achieve linear projection and feature extraction. Those linear operators can ultimately be implemented by general matrix multiplication (GEMM). For example, a 2-dimensional $K \times K$ convolution can be described as $\boldsymbol{y} = \boldsymbol{W} * \boldsymbol{x}$, $\boldsymbol{W} \in \mathbb{R}^{C_{out} \times C_{in} \times K \times K}, \boldsymbol{x} \in \mathbb{R}^{C_{in} \times H \times W}$, where C_{in}, C_{out}, k, H, W are input channel, output channel, kernel size, input height and width. To efficiently implement this algorithm, an *im2col* algorithm is widely adopted to unroll each convolution patch as a $(C_{in} \times K \times K)$-length vector. Therefore, the convolution is transformed to a GEMM $\boldsymbol{y}^{C_{out} \times (H'W')} = \boldsymbol{W}^{C_{out} \times N} \cdot \boldsymbol{x}^{N \times (H'W')}$, where H',W' are spatial height and width of $\boldsymbol{y}$, and N represents the unrolled vector length $(C_{in} \times K \times K)$. This *im2col* algorithm lays the foundation for modern high-performance CNN accelerator designs. Besides GEMM with static weights, advanced DNN architectures, e.g., attention-based natural language processing models [10] and dynamic CNNs with real-time-generated weights [11], require dynamic tensor-product-based operations to achieve better representability. Such essential and computationally-expensive modules require high-performance accelerators to support both operands to be dynamic signals.

B. Optical Neural Network Architecture

Here we give a short ONN literature review. Shen *et al.* [1] proposed to map decomposed unitary matrices to cascaded Mach-Zehnder interferometer (MZI) arrays to achieve neural network acceleration. Later, a recurrent ONN architecture was proposed based on MZI arrays [16]. This MZI-based ONN has a relatively high area cost and unsatisfactory noise-robustness [14], [15]. Zhao *et al.* [2] proposed a slimmed architecture, achieving 15%-38% area reduction and better robustness. Fast-Fourier-transform-based ONNs [3], [4] were proposed to map neural computations in the frequency-domain, which further reduces the ONN area cost by approximately $3\times$. Incoherent ONNs were proposed [7], [8] to achieve matrix multiplication with a small footprint. Previous ONNs only focus on inner-product with one operand being electrically-encoded in device configurations.

III. Proposed O^{2}NN Architecture

In this section, we introduce the architecture and features of the proposed O^{2}NN, including expressivity, efficiency, and robustness.

A. Dot-Product Engine with Both Optical Operands

Our proposed architecture is designed with a WDM-based differential structure to support flexible *fully-optical vector dot-product* computations. It allows both operands to be dynamically-encoded optical signals, which is inherently different from previous electro-optic neural architectures that are limited to stationary electrical weights [1]–[3], [8], [17].

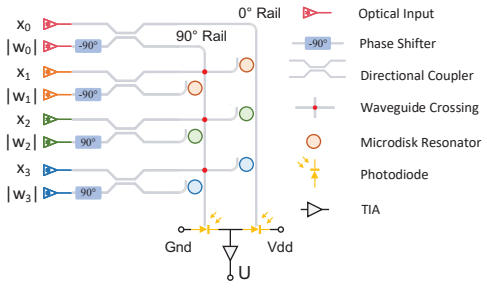


Fig. 1: Schematic of proposed WDM-based differential dot-product architecture with optical-weight extension.

Figure 1 demonstrates the structure of the engine to achieve dot-product between two optical vectors. In this architecture, the optical input vectors are denoted as $\boldsymbol{x} \in \mathbb{R}_+^N$ and $\boldsymbol{w} \in \mathbb{R}_+^N$, which are encoded into the light magnitude with a non-negative range of $[0, 1]$. Each pair of elements x_i and w_i is encoded in a unique wavelength λ_i. Interestingly, by putting a -$\pi/2$ degree phase shifter (PS) on the lower input port of a 2×2 optical directional coupler (DC), we can achieve an orthogonal addition/subtraction pair in the complex domain,

$$\begin{pmatrix} z_i^0 \\ z_i^1 \end{pmatrix} = \underbrace{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix}}_{\text{directional coupler}} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & e^{-j\pi/2} \end{pmatrix}}_{\text{phase shifter}} \begin{pmatrix} x_i \\ w_i \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} x_i + w_i \\ j(x_i - w_i) \end{pmatrix}, \quad (1)$$

where z_i^0, z_i^1 represent upper and lower output port of the directional coupler, respectively. Different z_i with different wavelengths λ_i will be re-directed by the resonated MR onto their corresponding rails, i.e., z_i^0 onto 0° rail and z_i^1 onto 90° rail. According to the WDM technique, different optical wavelengths can propagate on the same waveguide without mutual interference, which enables highly parallel signal processing. At the end of the rail, photodiodes (PDs) are used to accumulate the energy of the WDM optical signals, proportional to the square of magnitude, and generate photocurrent I^0 and I^1,

$$\begin{pmatrix} I^0 \\ I^1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \|\boldsymbol{x} + \boldsymbol{w}\|_2^2 \\ \|j(\boldsymbol{x} - \boldsymbol{w})\|_2^2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sum_{i=0}^{N-1} (x_i + w_i)^2 \\ \sum_{i=0}^{N-1} (x_i - w_i)^2 \end{pmatrix}. \quad (2)$$

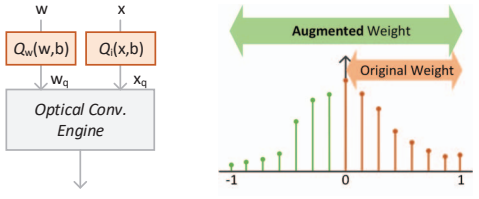
To calculate the optical dot product, we adopt a differential structure to transfer the two rails of photocurrent to an electrical voltage signal U which carries the dot product result,

$$U = G(I_0 - I_1) = 2G \sum_{i=0}^{N-1} x_i w_i \propto \sum_{i=0}^{N-1} x_i w_i, \quad (3)$$

where G is the gain of the on-chip transimpedance amplifiers (TIA). The superiority of the proposed architecture is that both operands are high-speed optical signals that allow dynamic encoding. Also, all components in this computing core are of fixed configuration, which can be fully passive with near-zero energy consumption, no external control overhead, and no potential thermal crosstalk, especially when both operands are dynamically generated from other optical circuits.

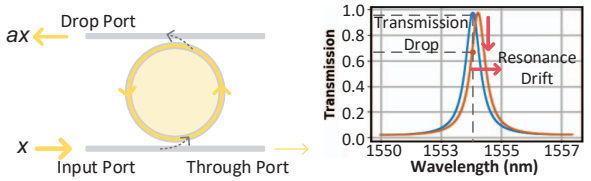
B. Expressivity Boost with Optical-Weight Extension

As analyzed in the previous section, both operands are constrained to be non-negative values as they are encoded into



(a) (b)

Fig. 2: (a) Distribution of weights with 3-bit augmented quantization. (b) Augmented optical quantization flow.



(a) (b)

Fig. 3: (a) Add-drop MR resonator structure with non-ideal transmission factor. (b) Drop port transmission decay caused by resonation wavelength shift.

the light magnitude. However, if one operand is weight, then it will cause trainability issues since non-negative weights inevitably limit the model expressivity due to abnormal activation distribution and pruned solution space. To solve this weight range limitation problem, we apply a static weight extension technique to augment the proposed architecture with better model expressivity and minimum hardware cost. By simply changing half of the passive phase shifters from $-90°$ to $90°$ and encoding $|w| \in [0, 1]$ into the light magnitude, shown in Fig. 1, we can statically allow half of the weights to be negative. The advantage is that the sign bit is offloaded to the extra π phase shift in the passive phase shifter without changing the input optical signal range. With static weight extension, our engine is able to generate a balanced output distribution with negligible hardware cost, which is the key feature that guarantees our superior model expressivity.

C. Performance Boost with Augmented Optical Quantization

For efficient optical neuromorphic computing, low-bitwidth inputs and weights are highly preferable. [1], [8], [15]. In this section, we introduce how augmented optical quantization empowers our proposed architecture with superior compatibility with low-bit quantization shown in Fig. 2a. Given a b-bit quantized signal within $[0, 1]$, all possible quantized values can be expressed as $\{\frac{k}{2^b-1}\}_{k=0}^{2^b-1}$ using a uniform quantizer,

$$\mathcal{Q}(x, b) = \frac{1}{2^b - 1}\texttt{Round}\left(\frac{x}{1/(2^b - 1)}\right) \tag{4}$$

With the extra π phase shift on the negative optical path mentioned in Section III-B, the engine is able to equivalently express negative weights $\{-\frac{k}{2^b-1}\}_{k=0}^{2^b-1}$, thus the number of implementable quantized weights $\boldsymbol{w}_q$ is almost doubled for free with a zero-centered symmetric distribution shown in Fig. 2b. Even with binarized weights $|w| \in \{0, 1\}$, augmented optical quantization will boost our architecture to a ternary ONN $w \in \{-1, 0, 1\}$ with a higher model expressivity and representability but still maintain high performance from binarized laser modulation and potential ADC/DAC elimination. Moreover, our proposed engine can naturally implement scaled quantized weights $w \in \{-\mathbb{E}[|w|], 0, \mathbb{E}[|w|]\}$ [11], [18], [19], where $\mathbb{E}[|w|]$ calculates the layer-wise average of absolute weights, to achieve better trainability by setting the laser input intensity corresponding to those scaled values at no hardware cost. A quantization-aware training procedure [20] is adopted to train our proposed ONN. We denote the b-bit quantized weights and input as $\mathcal{Q}_w(w, b)$ and $\mathcal{Q}_i(x, b)$ respectively.

D. Robustness Analysis and Solution

In this section, we analyze the variation-robustness of the proposed $\texttt{O}^2\texttt{NN}$ and present a solution to maximize its fidelity.

1) Dynamic Variation Analysis: Considering there is stochastic dynamic drift in the analog optical signals, we have $\hat{x}_i = (x_i + \delta x_i)e^{j\delta\phi_i^d}$ and $\hat{w}_i = (w_i + \delta w_i)e^{j\delta\phi_i^d}$, where $\delta\phi_i^d$ is the dynamic phase drift. For a given input signal speed $\mathcal{B}$, the signal-to-noise ratio (SNR) is,

$$SNR = \frac{\bar{P}(\boldsymbol{x})}{\bar{P}(\delta\boldsymbol{x})} = \frac{\mathbb{E}[\boldsymbol{x}^2]}{\sigma^2} \approx \frac{C}{\mathcal{B}}, \quad \delta\boldsymbol{x} \sim \mathcal{N}(0, \sigma_x^2), \tag{5}$$

where the SNR is empirically to be inversely proportional to the input signal rate, e.g., 40 Gb/s signal rate corresponds to an SNR of 10 [21], thus the constant C is approximately set to 40. We extract the relative phase drift between two operands to an equivalent dynamic phase perturbation on the phase shifter, i.e., $\hat{x}_i = (x_i + \delta x_i)$ and $\hat{w}_i = (w_i + \delta w_i)$, and $\phi_i = \pm\pi/2 + \delta\phi_i^d$, where $\delta\phi^d \sim \mathcal{N}(0, \sigma_\phi^2)$ is the dynamic input phase drift.

2) Static Variation Analysis: Considering the phase shifter produces an extra phase drift $\delta\phi^s \sim \mathcal{N}(0, \sigma_\phi^2)$. Though this drift is deterministic, it is expensive to evaluate each device drift individually for a large accelerator, hence we assume the static phase error is also a Gaussian random variable. Hence we have $\phi_i = \pm\pi/2 + \delta\phi_i^d + \delta\phi_i^s \sim \mathcal{N}(\pm\pi/2, 2\sigma_\phi^2)$, then the output of the directional coupler can be derived as,

$$\begin{aligned}\begin{pmatrix}\hat{z}_i^0\\ \hat{z}_i^1\end{pmatrix} &= \frac{1}{\sqrt{2}}\begin{pmatrix}1 & j\\ j & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & e^{j\phi_i}\end{pmatrix}\begin{pmatrix}\hat{x}_i\\ \hat{w}_i\end{pmatrix}\\ &= \frac{1}{\sqrt{2}}\begin{pmatrix}\hat{x}_i - \hat{w}_i\sin\phi_i + j\hat{w}_i\cos\phi_i\\ \hat{w}_i\cos\phi_i + j(\hat{x}_i + \hat{w}_i\sin\phi_i)\end{pmatrix}.\end{aligned} \tag{6}$$

Then we further consider the non-ideal transmission factor of the MR resonator that only transmits $\alpha \in [0, 1]$ of the light energy to the rail due to resonance spectrum drift and insertion loss, shown in Fig. 3a and 3b. α is estimated by a unilateral normally distributed variable $\alpha \sim \max(0, 1 - |\mathcal{N}(0, \sigma_\alpha^2)|)$. Thus, the photocurrent can be given by,

$$\begin{pmatrix}\hat{I}_0\\ \hat{I}_1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}\sum_{i=0}^{N-1}\alpha_i^0(\hat{x}_i^2 - 2\hat{x}_i\hat{w}_i\sin\phi_i + \hat{w}_i^2)\\ \sum_{i=0}^{N-1}\alpha_i^1(\hat{x}_i^2 + 2\hat{x}_i\hat{w}_i\sin\phi_i + \hat{w}_i^2)\end{pmatrix}. \tag{7}$$

Therefore, the differential output of the engine becomes,

$$\hat{U} \propto \sum_{i=0}^{N-1}\left(\frac{\alpha_i^0 - \alpha_i^1}{4}(\hat{x}_i^2 + \hat{w}_i^2) - \frac{\alpha_i^0 + \alpha_i^1}{2}\hat{x}_i\hat{w}_i\sin\phi_i\right). \tag{8}$$

Our engine is highly robust to static device noises since the design point $\phi = \pm\frac{\pi}{2}$ and $\alpha = 1$ are the local optima of *sin* and MR resonance curve with the minimum sensitivity.

3) Address Static and Dynamic Noises via Variation-Aware Knowledge Distillation: We handle the above static and dynamic variations by training ONNs with the non-ideality modeling, shown in Eq. (8). We apply a knowledge distillation training strategy to improve the noise-tolerance of our architecture. First, we pre-train an ideal ONN model without noise injection. as the teacher model $f_t(\cdot; \boldsymbol{W})$. Then we inject both static and dynamic variations to a noisy student model $f_s(\cdot; \boldsymbol{W}, \sigma_x, \sigma_\phi, \sigma_\alpha)$. We train the student model with a combined objective of hard target and soft target,

$$\begin{aligned} \mathcal{L} &= \beta T^2 \mathcal{D}_{KL}(q, p) + (1-\beta) H(y, \texttt{softmax}(f_s)), \\ p &= \frac{\exp(f_s/T)}{\sum \exp(f_s/T)}, \quad q = \frac{\exp(f_t/T)}{\sum \exp(f_t/T)}, \end{aligned} \tag{9}$$

where $\mathcal{D}_{KL}$ is the KL divergence, T is a temperature to control the smoothness, $H(y, \texttt{softmax}(f_s))$ is the cross-entropy loss, and β is a weighting factor to balance the soft and hard targets. Though this method introduces marginal training time overhead, it can effectively improve the ONN robustness to both static and dynamic errors. A noise source cooling strategy that gradually reduces the noise intensity is leveraged in low-bit (e.g., <3 bit) quantized training for better convergence.

E. Discussion: Hardware Cost and Features

In this section, we analyze and compare the hardware cost and features of our proposed ONN with previous ONN designs.

1) Optical Input Encoding Cost: The optical inputs are driven by coherent sources with phase shifters to control their phases. The weight encoding cost can be amortized by broadcasting to multiple processing units [23]. Moreover, since the weights are relatively stationary in ONNs, they can be directly modulated by phase change materials [17] or efficient laser modulation, which has near-zero area cost and power overhead. Since our architecture supports both operands to be optical signals, our architecture is the first integrated ONN that can achieve multiplication beyond static synaptic weights. Dynamic optical signals can be directly fed into our engine to support fully-optical attention-like operations [10] and NNs with dynamically-generated weights [11], where no extra energy is required due to its fully-passive design.

2) Area Cost, Latency, and Energy Consumption: Now we give a theoretical analysis of the hardware cost. Figure 4 shows how we assign multiple engines to a GEMM task with the *im2col* algorithm. Without losing generality, we only consider the most area-consuming directional couplers (DCs) and phase shifters (PSs) and assume they share the same size and aspect ratio of $w_{dc}/h_{dc} = 2$. We partition the GEMM task into $P \times Q$ sub-tasks to balance hardware cost and parallelism. For a matrix multiplication $\boldsymbol{A}_{M\times N} \cdot \boldsymbol{B}_{N\times L}$, the proposed architecture costs PQN PSs and PQN DCs. This partitioned engine assignment has an estimated latency of $\tau_{ours} = \frac{ML(2w_{dc}+Nh_{dc})}{PQc} = \frac{ML(N+4)h_{dc}}{PQc}$, where c is the speed of light. The previous MZI-based ONN architecture costs

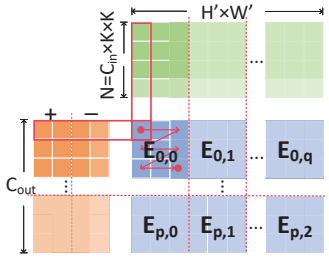


Fig. 4: Tiling-based engine assignment for parallel GEMM.

$M(M-1) + N(N-1) + 2\max(M,N)$ DCs and the same number of PSs to implement an $M \times N$ matrix-vector multiplication with latency $\tau_{mzi} = \frac{4(M+N)Lw_{dc}}{c} = \frac{8(M+N)Lh_{dc}}{c}$ [1]. We compare their latency-area product (LAP),

$$\begin{aligned} A_{mzi} \cdot \tau_{mzi} &= 8(M+N)(M^2+N^2)Lh_{dc}/c \\ A_{ours} \cdot \tau_{ours} &= MNL(N+4)h_{dc}/c. \end{aligned} \tag{10}$$

For fully-connected layers, we assume $M = N$, then we have $\frac{32N}{N+4}$ times smaller LAP than MZI-ONN. For a typical convolutional layer, we assume $N = K^2M$. Then the LAP improvement is around $8(K^2+1)$ times. If the MZI-based ONN adopts $P \times Q$ MZI sub-arrays, it costs around $\frac{8LNh_{dc}}{Qc}$ latency and PQN^2 components, which is still $8P$ times less efficient than our architecture.

Our architecture is also more energy-efficient than prior ONNs. For the photonics part, the only optical device tuning power is the phase control and modulation. As mentioned before, the power of the weight modulation can be amortized by weight sharing and even reduced by direct laser modulation. In attention-like operations and layers with dynamically-generated weights, since both operands are directly from the previous layer and already in the optical domain, our architecture potentially consumes near-zero energy. Hence we have comparable or better energy efficiency than previous coherent ONNs in different application scenarios. For the electrical part, since our engine supports binarized inputs, our architecture is compatible with a DAC/ADC-less design, enabling potentially-ultra-low power as ADCs/DACs take most power [8], [15].

3) Differences from Prior Work: We compare with previous ONNs in Table I. Though larger than MRR-ONN, compared with other coherent ONNs [1]–[3], our architecture has a smaller area cost. No previous ONN can directly perform linear inner-product between two optical signals. Our proposed architecture is the first integrated ONN that supports both operands to be optical signals, making it possible to realize direct layer cascading and *optical-optical product* that is necessary in attention-based neural architectures and NNs with dynamically-generated weights [11]. Compared with noise-sensitive MRR-ONN and unscalable, error-prone MZI-ONN [1], [14], [15] , our architecture achieves a relatively-low hardware cost, good model expressivity, and much better variation-tolerance. Furthermore, our architecture can well-support a wide spectrum of modern DNN architectures across CNN, MLP, and AdderNet, etc. MZI-ONN has low compatibility with network compression given its complicated principle [3], [15], and

TABLE I: Comparison among ONNs. Area cost is normalized to O^2NN on a size-N matrix-vector multiplication based on real device sizes [1], [7], [8], [22], i.e., one MZI ≈240×40 μm^2, one DC≈60×40 μm^2, one PS≈60×40 μm^2, and one MRR≈20×20 μm^2. Note that our area is not a simple accumulation of device sizes but is estimated with real layout information as a reference. Power is normalized to ours with the same statistics from the PDK [22], i.e., one PS≈20 mW and one MRR≈4 mW. The block size is set to k=4 for FFT-ONN [3].

	MZI-ONN [1]	Slim-ONN [2]	FFT-ONN [3]	MRR-ONN [7]	O^2NN
Norm. Area Cost	∼1.71×	∼0.86×	∼0.86×	∼0.1×	1×
Norm. Power	∼2×	∼1×	∼1.25×	∼0.2×	1×
General Matrix Multiplication	Yes	No	No	Yes	Yes
Optical Operand Support	Only One	Only One	Only One	Only One	Both
Robustness	Medium	Medium	Medium	Low	High
Control Complexity	Medium-High	Medium	Medium-Low	High	Medium
CNN Support	Yes	No	No	Yes	Yes
Quantization Compatibility	Low	Low	Medium	Medium-High	High
Output Range	Positive	Positive	Positive	Positive&Negative	Positive&Negative

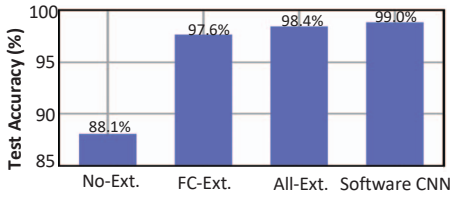


Fig. 5: Evaluation of 8-bit optical-weight extension on MNIST. *Ext.* is short for extension.

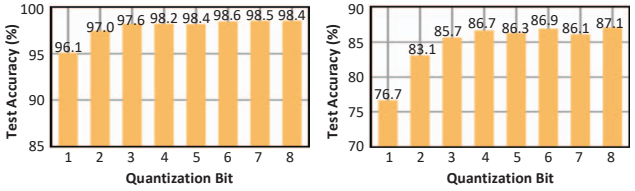


Fig. 6: Low-bit quantization on (a) MNIST and (b) FMNIST.

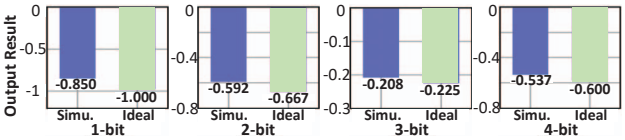


Fig. 7: Optical simulation results with 1- to 4-bit precision. 1-bit: $\boldsymbol{x}$=(1,0,1,1), $\boldsymbol{w}$=(1,0,-1,-1). 2-bit: $\boldsymbol{x}$=$(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3})$, $\boldsymbol{w}$=$(\frac{1}{3}, 0, -\frac{2}{3}, -1)$. 3-bit: $\boldsymbol{x}$=$(0, \frac{1}{7}, \frac{1}{7}, \frac{6}{7})$, $\boldsymbol{w}$=$(\frac{1}{7}, \frac{6}{7}, -\frac{5}{7}, -\frac{2}{7})$. 4-bit: $\boldsymbol{x}$=$(\frac{1}{15}, \frac{1}{5}, \frac{11}{15}, \frac{7}{15})$, $\boldsymbol{w}$=$(\frac{8}{15}, \frac{2}{15}, -\frac{11}{15}, -\frac{4}{15})$.

MRR-ONN only scales its weight range without increasing the valid quantization levels. In contrast, our ONN can seamlessly support extremely-low-bit quantization with better expressivity.

IV. Experimental Results

We conduct experiments on the MNIST and FashionMNIST (FMNIST) dataset. We use a CNN setting C16-C16-P5-F32-F10, where C16 is a 3×3 convolutional (Conv) layer with 16 kernels, P5 means average pooling with output size 5×5, and F32 is a fully-connected (FC) layer with 32 neurons. We implement ONNs with PyTorch and train all models for 50 epochs with the Adam optimizer. and a mini-batch size of 32. We use Lumerical INTERCONNECT to do optical simulation with real devices from the AIM PDK [22], which should already model comprehensive and practical non-ideal factors. In knowledge distillation, we set T=6 and β=0.9. We gradually cool down the noise intensity by 20% in lower than 3-bit cases.

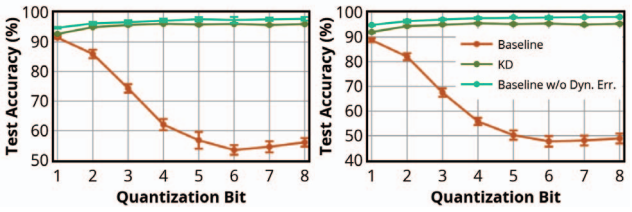
A. Comparison Experiments

We first validate the effectiveness of optical-weight extension and augmented optical quantization, then evaluate the robustness via optical simulation and comparison experiments.

1) Optical-Weight Extension: We compare four configurations of an 8-bit quantized optical CNN: 1) no optical-weight extension, 2) only extend FC layers, 3) apply weight extension to both FC and Conv layers, and 4) ideal software CNN. Figure. 5 shows that weight extension for fully-connected layers is essential for model expressivity. With balanced convolutions and fully-connected layers, the ONN model can recover its full modeling capacity with the highest inference accuracy. Therefore, the proposed ONN architecture can be used to accelerate modern CNN models with negligible accuracy degradation (∼0.5%) compared with the original software CNNs.

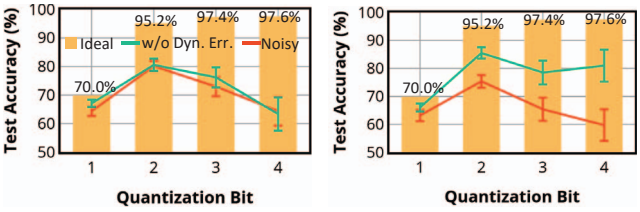
2) Augmented Optical Quantization: In analog DNN accelerators, the maximum precision is 8-bit or even 4-bit considering control complexity [1], [8], [15] In Fig. 6, our augmented optical quantization enlarges the solution space and achieves high accuracy even under low-bit quantization on both dataset. Even for binarized optical inputs, we can still maintain >96% accuracy on MNIST and 76% on FashionMNIST, which enables the coexistence of hardware-efficient optical computing and improved inference accuracy. In contrast, the state-of-the-art quantized MZI-ONN still suffers from > 10% accuracy drop on MNIST under extremely-low-bit quantization [15].

3) Variation-Robustness Evaluation: We use Lumerical INTERCONNECT tools with the AMF process design kit (PDK) [24] to validate the fidelity of our architecture under static device variations. The simulation results in Fig. 7 show that phase shifter drift and MRR non-ideality lead to 10-15% dot-product error. Then, we further consider dynamic variations in our accuracy evaluation with our PyTorch-based ONN simulator on different setups, 1) noise-unaware training



(a) (b)

Fig. 8: Robustness evaluation on MNIST. Error bars show the $\pm 1\sigma$ variance. (a) σ_ϕ=0.04, σ_α=0.04, SNR=39.81 (16 dB) (b) σ_ϕ=0.05, σ_α=0.05, SNR=31.62 (15 dB).



(a) (b)

Fig. 9: Robustness of MRR-ONN [7] on MNIST. (a) σ_α=0.04, SNR=39.81 (16 dB). (b) σ_α=0.05, SNR=31.62 (15 dB).

(`Baseline`), 2) noise-unaware training w/o dynamic variations (`Baseline w/o Dyn. Err.`), and 3) variation-aware knowledge distillation (`KD`). Figure 8 shows that our `O`^{2}`NN` is extremely robust to large static device error [14], [15], consistent with the analysis in Section III-D2, but sensitive to dynamic variations. Our knowledge-distillation-based training method can help recover the majority of the accuracy with $\sim$3% degradation under both static and dynamic noises. We also observe higher robustness to dynamic noise with lower bitwidth, which is beneficial for binary or ternary ONNs and thus further shows our superior compatibility with low-bit quantization. As a comparison, we show the robustness of MRR-ONN in Fig. 9. We observe that it has high sensitivity to both static and dynamic errors and suffers from a larger accuracy degradation than our `O`^{2}`NN` under low-bit quantization. MZI-ONN typically suffers from even larger accuracy loss due to severe phase error accumulation effects [1], [15], thus we do not show its accuracy here for brevity.

V. Conclusion

In this work, we propose a new optical neural network architecture `O`^{2}`NN` to enable efficient and noise-robust photonic neuromorphic computing, which is the first one that supports tensor product with both operands to be dynamically-encoded light signals. A novel WDM-based differential dot-product engine is presented with extended optical weights and augmented quantization techniques, demonstrating enhanced model expressivity and performance under low-bit quantization. We give an analysis of static and dynamic variations and present a knowledge-distillation-based training method to enable variation-tolerant optical neurocomputing under practical noises. A thorough comparison with prior work is provided to show our advantages in hardware cost, efficiency, and features. Experimental results demonstrate that our `O`^{2}`NN` can support flexible, robust, and efficient optical neural computing with both operands being optical signals even when low-bit optical quantization and practical variations exist.

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] Z. Zhao, D. Liu, M. Li *et al.*, "Hardware-software co-design of slimmed optical neural networks," in *Proc. ASPDAC*, 2019.
[3] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards area-efficient optical neural networks: an FFT-based architecture," in *Proc. ASPDAC*, 2020.
[4] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE TCAD*, 2020.
[5] Z. Ying, C. Feng, Z. Zhao *et al.*, "Electronic-photonic arithmetic logic unit for high-speed computing," *Nature Communications*, 2020.
[6] C. Sun, M.T. Wade, Y. Lee *et al.*, "Single-chip microprocessor that communicates directly using light," *Nature*, 2015.
[7] A.N. Tait, T.F. de Lima, E. Zhou *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, 2017.
[8] W. Liu, W. Liu, Y. Ye *et al.*, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *Proc. DATE*, 2019.
[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] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017.
[11] Y. Chen, X. Dai, M. Liu *et al.*, "Dynamic convolution: Attention over convolution kernels," in *Proc. CVPR*, 2020.
[12] 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.
[13] T.W. Hughes, M. Minkov, Y. Shi *et al.*, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, 2018.
[14] Z. Zhao, J. Gu, Z. Ying *et al.*, "Design technology for scalable and robust photonic integrated circuits," in *Proc. ICCAD*, 2019.
[15] 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.
[16] C. Feng, Z. Zhao, Z. Ying *et al.*, "Compact design of On-chip Elman Optical Recurrent Neural Network," in *Proc. CLEO*, 2020.
[17] M. Miscuglio and V.J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Review*, 2020.
[18] J. Li, D. Mengu, Y. Luo *et al.*, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," *Advanced Photonics*, pp. 1 – 13, 2019.
[19] C. Zhu, S. Han, H. Mao *et al.*, "Trained ternary quantization," in *Proc. ICLR*, 2017.
[20] 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.
[21] J. Sun, R. Kumar, M. Sakib *et al.*, "A 128 Gb/s PAM4 Silicon Microring Modulator With Integrated Thermo-Optic Resonance Tuning," *Journal of Lightwave Technology*, 2019.
[22] 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.
[23] R. Hamerly, L. Bernstein, A. Sludds *et al.*, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, 2019.
[24] "Advanced micro foundry," http://www.advmf.com/services/.