Light in AI: Toward Efficient Neurocomputing With Optical Neural Networks—A Tutorial

Jiaqi Gu, *Student Member, IEEE*, Chenghao Feng, *Graduate Student Member, IEEE*, Hanqing Zhu, *Graduate Student Member, IEEE*, Ray T. Chen, *Fellow, IEEE*, and David Z. Pan, *Fellow, IEEE*

***Abstract*—In the post Moore's era, conventional electronic digital computing platforms have encountered escalating challenges to support massively parallel and energy-hungry artificial intelligence (AI) workloads. Intelligent applications in data centers, edge devices, and autonomous vehicles have restricted requirements in throughput, power, and latency, which raises a high demand for a revolutionary neurocomputing solution. Optical neural network (ONN) is a promising hardware platform that could represent a paradigm shift in efficient neurocomputing with its ultra-fast speed, high parallelism, and low energy consumption. In recent years, efforts have been made to facilitate the ONN design stack and push forward the practical application of optical neural accelerators. In this tutorial, we give an overview of state-of-the-art cross-layer co-design methodologies for scalable, robust, and self-learnable ONN designs across the circuit, architecture, and algorithm levels. Besides, we analyze challenges and highlight emerging directions targeting next-generation optics for AI.**

***Index Terms*—Optical neural network, optical computing, scalability, robustness, trainability, software-hardware co-design.**

I. Introduction

DEEP neural networks (DNNs) have received an explosion of interest for their superior performance in artificial intelligent (AI) tasks. The computing capacity is in an arms race with the rapidly escalating model size and data volume. Certain applications, e.g., autonomous vehicles, data centers, and edge devices, have strict efficiency, latency, and bandwidth constraints, raising a surging need to develop more efficient computing solutions. However, as Moore's law is winding down, it becomes increasingly challenging for conventional electrical processors to support such massively parallel and energy-hungry DNN workloads. Limited clock frequency, high latency, high heat density, and large energy consumption of CPUs, FPGAs, and GPUs motivate us to seek an alternative solution using optics.

Optics is a promising medium for ultra-fast matrix-vector multiplication (MVM). The inputs of DNNs can be encoded into optical signals using high-speed optical modulators. Massively-parallel MVM is implemented at the speed of light with near-zero energy consumption by propagating light through an optical system. Recently, the integrated optical neural networks (ONNs) based on silicon photonics have attracted extensive research interest and represented a paradigm shift in efficient AI given their competitive integration density, ultra-high energy efficiency, and good CMOS-compatibility [1]–[3]. With potentially petaFLOPS per mm^2 compute density and attojoule/MAC energy efficiency, fully-optical NNs demonstrate orders-of-magnitude higher performance and efficiency than their electrical counterparts [1]–[5]. Research efforts have been made on reservoir computing [6], spike processing [7], and Ising machines [8], while we focus more on recent advances in photonic artificial NNs.

Besides the above advantages, ONNs currently encounter new challenges in chip area, noise robustness, on-chip trainability, nonlinearity, and weight storage/access. There has been extensive exploration that attempts to address those challenges across the entire ONN design stack. New ONN architectures have been proposed to improve compactness, flexibility, and inference throughput [1], [4], [5], [9], [10]. Non-volatile photonic memory based on phase change materials has been demonstrated to enable photonic in-memory computing [5]. Hardware-software co-design methodologies are investigated to jointly optimize area, power, robustness, endurance, etc. [11], [12]. Various on-chip training protocols have been put forward to facilitate *in-situ* device optimization for self-learnable ONNs [13], [14]. Future photonic AI needs synergistic design technology and a holistic solution to push the limits of the practical deployment of photonic neural accelerators. We released a PyTorch-centric library TorchONN to facilitate the ONN design stack for research exploration.

In this tutorial brief, we cover the following aspects,

- **ONN Circuit-Architecture-Algorithm Co-Design** – We will cover recent ONN architecture innovations, including MZI-ONN [1], [15], Ring-ONN [10], [16], [17], FFT-ONN [9], [18], PCM-based tensor cores [5], [19], etc. We discuss state-of-the-art hardware-software co-design methodologies to facilitate ONN designs on area efficiency [9], [15], [18] and robustness [11], [12], [20].
- **ONN On-Chip Training** – We introduce recent progress in ONN *in-situ* training towards scalable and efficient self-learnable photonic AI engines [14], [21]–[23].
- **Future Trends in *Optics for AI*** – We highlight emerging directions and opportunities to push forward the practical application of next-generation optical neural processors.

Manuscript received January 30, 2022; revised April 20, 2022; accepted April 26, 2022. Date of publication April 28, 2022; date of current version May 27, 2022. This work was supported in part by the Multidisciplinary University Research Initiative Program through the Air Force Office of Scientific Research under Contract FA 9550-17-1-0071, monitored by Dr. Gernot S. Pomrenke. This brief was recommended by Associate Editor Y. Ha. *(Corresponding author: Jiaqi Gu.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: jqgu@utexas.edu; fengchenghao1996@utexas.edu; hqzhu@utexas.edu; chen@ece.utexas.edu; dpan@ece.utexas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2022.3171170.

Digital Object Identifier 10.1109/TCSII.2022.3171170

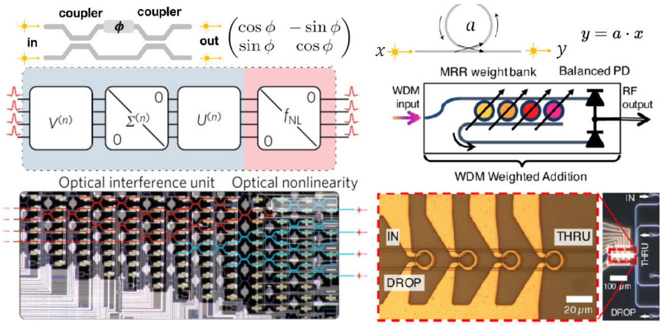

Fig. 1. *Left*: MZI-based ONN architecture [1]. *Right*: MRR-based ONN architecture [16].

II. Optical Neural Network Basics

Constructed by cascaded optical components, photonic integrated circuits (PICs) can be used to realize neurocomputing. Based on the mechanism of information encoding and processing, integrated ONNs are generally categorized into coherent and incoherent ONNs.

Coherent ONNs: Coherent ONNs encode information in the amplitude-phase of optical signals and achieve linear operations with light interference. Basic components in coherent ONNs include phase shifters for phase modulation, directional couplers for interference, and waveguide crossings for signal routing. Mach-Zehnder interferometer (MZI) is a device consisting of two cascaded directional couplers with phase shifters on its inner arm, which can perform arbitrary 2-dimensional rotation $\boldsymbol{R}(\phi) = (\cos\phi, -\sin\phi; \sin\phi, \cos\phi)$. By cascading MZIs into a triangular [1] array, one can construct an arbitrary N-dimensional unitary as $\boldsymbol{U}(\boldsymbol{\Phi}) = \boldsymbol{D}\prod_{i=k}^{2}\prod_{j=1}^{i-1}\boldsymbol{R}_{ij}(\phi_{ij})$ [1]. To implement an $M \times N$ weight matrix $\boldsymbol{W}$, one can decompose $\boldsymbol{W}$ into $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ via singular value decomposition (SVD) and map two unitaries to MZI meshes. Based on this principle, Shen *et al.* demonstrated an MZI-based optical interference unit, shown in Fig. 1, for ultra-fast universal linear projection. Activation can be realized by optical nonlinearity, e.g., saturable absorber. The throughput of coherent ONNs, if using broadband devices, can be significantly improved with wavelength-division multiplex (WDM) techniques. Multiple wavelengths can propagate through the same circuit in parallel. As the first demonstration of coherent ONNs, MZI-based ONN shows potentially order-of-magnitude speedup and higher energy efficiency than electrical GPUs with comparable classification accuracy.

Incoherent ONNs: Incoherent ONNs are multi-wavelength PICs that encode information into light intensity and realize tensor operations using WDM techniques. Multiply-accumulate (MAC) operation is achieved by amplitude modulation and photo-detection of WDM signals. Basic components in incoherent ONNs are used for magnitude modulation, including micro-ring resonators (MRRs) [16], multi-operand rings (MORRs) [10], phase change materials (PCMs) [19], etc. MRR-based ONNs [16], [17] encode a matrix into an MRR weight bank and input modulated WDM signals through it to realize general matrix multiplication (GEMM), shown in Fig. 1. Besides, PCM-based photonic tensor cores (PTCs) [19] were proposed to perform photonic in-memory neurocomputing. PCM devices can temporarily remember the programmed transmission level and achieve light intensity modulation with zero static power consumption.

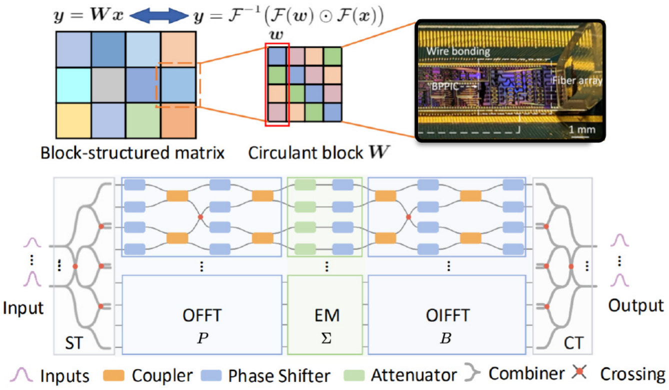

Fig. 2. Butterly-style frequency-domain ONNs [9], [18], [24].

III. ONN Circuit-Architecture-Algorithm Co-Design

After introducing basic knowledge of ONNs, we now discuss how to solve critical issues in the ONN design stack that hinder their practical application. As an emerging technology, integrated ONNs still encounter critical issues in area cost, noise robustness, and trainability. In this section, we introduce state-of-the-art co-design methodologies to improve ONN area efficiency and noise tolerance to push the practical and scalable application of optical AI accelerators.

A. Area-Efficient ONN Architecture Design

Unlike nanometer-level transistors, optical devices are relatively bulky in their physical dimensions, e.g., hundreds or thousands of square micrometers. Hence the large area cost of ONNs becomes a concern to its compute density. We show how to improve the area efficiency for coherent and incoherent ONNs from a software-hardware co-design perspective.

1) FFT-ONN (Butterfly-Style Coherent ONN): Previous MZI-based ONNs [1], [15] consume a large number of bulky MZIs and are area costly. To remedy this, a butterfly-style ONN was proposed, as shown in Fig. 2. Instead of implementing GEMM, FFT-ONN trades universality for area efficiency by adopting a restricted matrix as an efficient substitution. The butterfly-style photonic meshes $\boldsymbol{B}$ and $\boldsymbol{P}$ are constructed by basic optical devices to represent a family of restricted unitary. The diagonal mesh $\boldsymbol{\Sigma}$ can encode complex-valued frequency-domain weights. By setting $\boldsymbol{B} = \mathcal{F}^{-1}$ and $\boldsymbol{P} = \mathcal{F}$, FFT-ONN [9] can achieve circulant matrix multiplication in the Fourier domain, i.e., $Wx \rightarrow \boldsymbol{B\Sigma P}x \rightarrow \mathcal{F}^{-1}(\mathcal{F}(w) \odot \mathcal{F}(x))$. Weights w are encoded into the diagonal photonic mesh $\boldsymbol{\Sigma}$. Without sacrificing much model expressiveness, this butterfly-style ONN [9] demonstrated 3-4× lower area cost compared with previous MZI-based ONNs [1], [15].

To further augment the learnability and compactness of FFT-ONN, a modified version FFT-ONN-v2 [18] was presented to achieve highly-parallel optical convolutional neural networks (CNNs). The key innovation is that $\boldsymbol{B}$ and $\boldsymbol{P}$ matrices are relaxed to trainable butterfly transforms instead of fixed Fourier transform. To boost the throughput, different channels of convolutional kernels are encoded in the $\boldsymbol{\Sigma}$ matrix using narrow-band micro-disk resonators. FFT-ONN-v2 moves beyond the manually-designed Fourier-transform-based design concept and automatically learns the best unitary transform pairs through circuit-algorithm co-design.

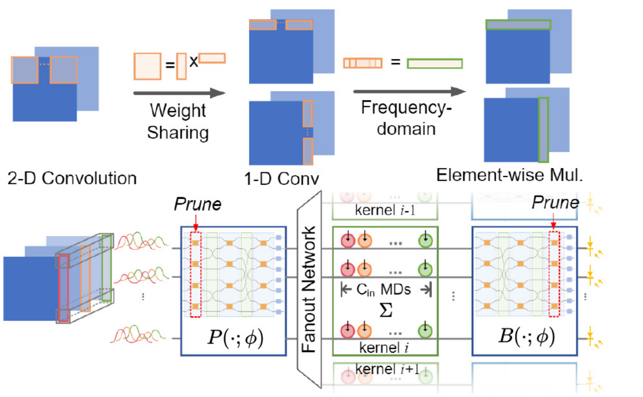

Fig. 3. FFT-ONN-v2 [18] with relaxed butterfly transform and highly-parallel micro-disk-based weight encoding.

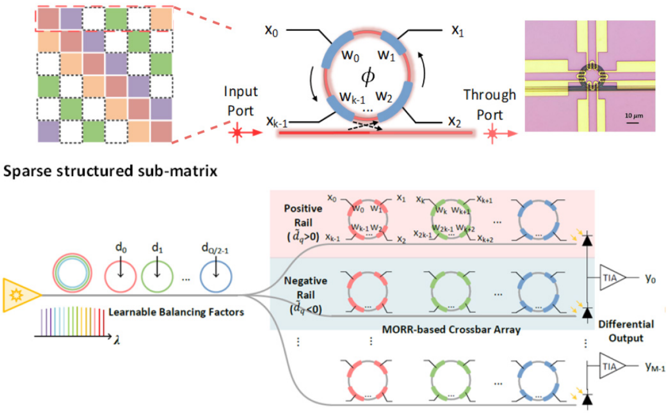

Fig. 4. SqueezeLight architecture [10]. *Top*: Map a sparse structured matrix into one MORR [10] and a 4-operand MORR tapeout [25].

Structural pruning is an effective co-design method to trim redundant phase shifters in $\boldsymbol{B}$ and $\boldsymbol{P}$ during training. By removing phase shifters with a regular pattern and re-training the model to recover the accuracy drop, FFT-ONN-v2 can achieve ~10× area reduction, 10× device-tuning power saving, and significantly higher noise tolerance compared to MZI-based ONNs with negligible model expressiveness degradation.

2) SqueezeLight (Multi-Operand Ring-Based Incoherent ONN): Incoherent ONNs generally have a smaller area than their coherent counterparts due to the compact size of MRRs. How to push the area lower bound set by MRR-ONNs, i.e., one MAC per MRR, implies an opportunity for ONN scalability breakthrough.

A novel ONN architecture SqueezeLight [10] provides a *cross-layer* solution that achieves vector operations within one multi-operand micro-ring (MORR), shown in Fig. 4. At the device level, unlike a normal MRR that only has one phase modulator, each MORR has multiple independent controllers. The phase shifts induced by k control signals will be weighted and accumulated through the round-trip phase shift $\phi = \sum_i^k w_i x_i^2$. The input light will prob the vector dot-product and carry out the result after a nonlinear transmission curve of the ring $y = f(\phi)$.

At the matrix level, by reusing the weights on this MORR-based neuron and rotating the input order, a structured matrix can be mapped onto one MORR with a quadratic footprint reduction. An array of MORR neurons working at different wavelengths can achieve ultra-compact block-structured matrix multiplication with built-in nonlinearity.

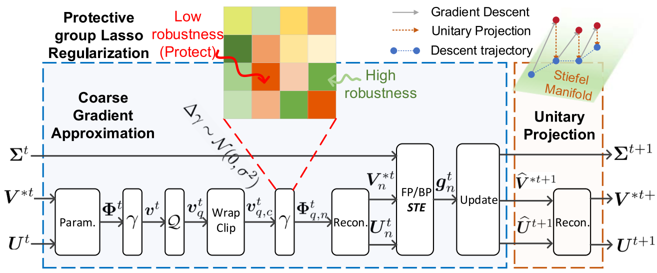

Fig. 5. ROQ [11] for robust ONNs under low-bit controls.

At the algorithm level, structured pruning is applied to explore the matrix sparsity to squeeze a larger block into the MORR. Sensitivity-aware training can further help reduce the impacts from phase noise and crosstalk among k controllers, leading to much higher computing fidelity. SqueezeLight achieves 20-30× less device usage and 8× less wavelength usage than MRR-ONNs, which implies that cross-layer co-design is the key to scalable and robust ONNs.

B. Robustness-Aware ONN Optimization

As analog computing platforms, ONNs inevitably encounter robustness issues due to process variations, device noises, limited control resolution, a non-ideal environment, and limited endurance. Efforts have been made to model the impacts of variations and noises [12], [16], [26]. Here we introduce several techniques to improve robustness for various ONNs.

1) ROQ (Noise-Aware Quantization): Tunable photonic devices are typically controlled by electrical voltage signals. To avoid high control cost, a low-bitwidth control resolution is preferred, which causes non-trivial discretization errors in computing. Different from quantization for classical NNs, ONNs require unique algorithms to handle this discrete optimization problem. ROQ [11] proposes a quantization scheme to optimize MZI control voltages in a discretized unitary subspace and adapt MZI-ONNs to low-bit controls. Unitary matrices U and V^* are decomposed and quantized into low-bitwidth control voltages with dynamic noise injection. The noisy reconstructed matrix is used in the forward propagation. Coarse gradients are approximated using a straight-through estimator and efficiently propagated to the unitary matrices. Unitary projection is iteratively applied to cast the updated U and V^* back to the unitary subspace. ROQ enables differentiable ONN optimization with non-ideality modeling and quantization adaptation, showing much better noise tolerance under low-bit device controls and practical noises.

2) Noise-Aware Training: During software training, noises and variations can be considered to effectively boost the resilience of ONNs. ROQ [11] estimates the noise sensitivity $\frac{\|\boldsymbol{W}(\tilde{\Phi})-\boldsymbol{W}(\Phi)\|}{\|\boldsymbol{W}(\Phi)\|}$ of each weight block and add protective regularization terms into the objective for noise-adaptive optimization. SqueezeLight [10] explicitly models the intra-MORR cross-talk during training and penalize the transmission sensitivity of each MORR device to boost the model robustness. O^{2}NN [27] employs a noise-aware knowledge distillation strategy to guide the optimization of noisy student

ONN models with a noise-free teacher model, which significantly improves robustness against both static process variation and dynamic input signal noises.

3) Noise Reduction via Pruning: Having more tunable devices in the circuits typically means stronger reconfigurability of the PIC but does not necessarily lead to a good ONN design. The reason is that noise-induced errors are generally positively related to the number of noise sources [1]. Pruning redundant devices sacrifices circuit programmability but brings additional robustness benefits [28]. FFT-ONN-family [9], [18], [24] removes unimportant weight blocks and redundant phase shifters in the butterfly-style transform. Pruning circuit chunks or devices in a regular pattern not only saves area cost but leads to significant noise robustness improvement.

4) ELight (Endurance-Enhancement for PCM-ONN): Different from other ONNss, PCM-based PTC [19] encounters a unique endurance issue due to the limited reprogramming times of PCM cells, e.g., ranging from 10^6 to 10^8 times. After being frequently re-written, PCM wires become aged and lose reprogrammability, leading to deviant transmission levels and severe accuracy drop. ELight [29] proposes a synergistic optimization framework to minimize the overall PCM write operations. Write-aware regularization is introduced into training to encourage higher similarity among weight blocks to help redundant PCM re-write elimination. Combined with a weight column reordering approach, ELight can achieve over 20× reduction in the total number of PCM writes. With ELight, PCM-based photonic in-memory neurocomputing will benefit from an order-of-magnitude longer lifetime.

IV. ONN On-Chip Training

Besides inference acceleration, training can also be offloaded to photonic chips. On-chip training has two major advantages. First, it is a promising approach to mitigating ONN noise issues. After deployment of pre-trained ONN models, various non-ideal effects on real PICs can lead to significant performance degradation [12], [22], [26]. Instead of inaccurate and time-consuming software noise simulation, on-chip training naturally handles real physical non-ideality *in-situ* with order-of-magnitude faster speed. Second, on-chip learnability is critical for NN training acceleration and online adaptation directly on optical chips instead of on GPUs.

ONN on-chip training is essentially a stochastic noisy optimization problem with restricted controllability, observability, and resource constraints. Prior work has proposed various algorithms to tackle this challenging problem. The first training protocol is based on brute-force phase tuning [1]. Neuro-evolution was adopted to search device configurations more efficiently [21]. An adjoint variable method [13] was proposed to compute *in-situ* first-order gradients, which requires light field monitoring inside each device. Due to algorithm inefficiency and limited implementation efficiency, the above methods are limited to handle ∼100 MZIs.

1) FLOPS: Zeroth-Order ONN Optimization: To improve the scalability and efficiency, a zeroth-order gradient method FLOPS [22] was proposed to efficiently estimate the gradients of MZI rotation phases and successfully demonstrate training on >1000 MZIs. With *In-situ* noise and crosstalk handling, FLOPS shows 3-5% higher accuracy under practical noises than previous training protocols.

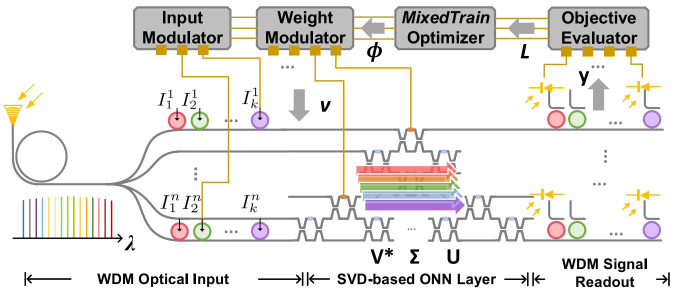

Fig. 6. Zeroth-order ONN on-chip training [14], [22].

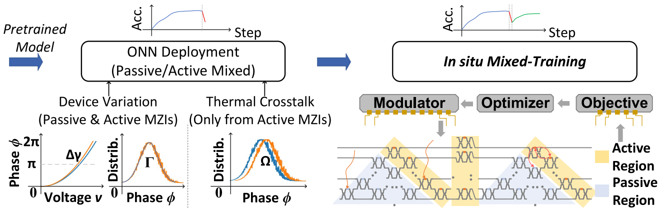

Fig. 7. ONN sparse mixed-training strategy MixedTrain [14].

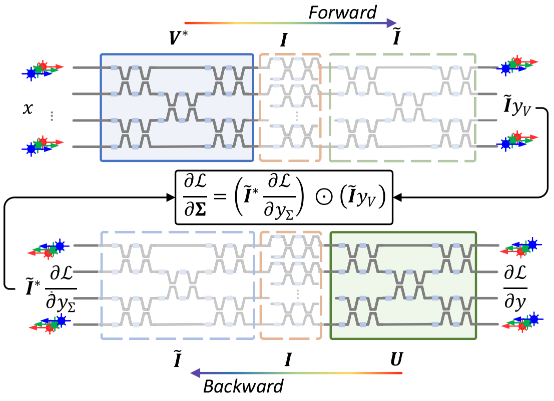
Forward
V^*
I
$\tilde{I}$
x
$\tilde{I}y_V$
$\frac{\partial \mathcal{L}}{\partial \Sigma} = \left(\tilde{I}^* \frac{\partial \mathcal{L}}{\partial y_\Sigma}\right) \odot \left(\tilde{I} y_V\right)$
$\tilde{I}^* \frac{\partial \mathcal{L}}{\partial y_\Sigma}$
$\frac{\partial \mathcal{L}}{\partial y}$
$\tilde{I}$
I
U
Backward

Fig. 8. L^{2}ight [23]: first-order subspace ONN optimization.

2) MixedTrain (Power-Aware Sparse ONN Training): An enhanced solution MixedTrain [14] was proposed to explore sparsity during training by only optimizing a small subset of MZIs, which successfully trains >2500 MZIs. Besides, device programming power is explicitly optimized during on-chip training, leading to >90% training power reduction [14].

3) L^{2}ight (in-situ Subspace First-Order Optimization): L^{2}ight [23] further scales on-chip training to million-parameter MZI-ONNs via *in-situ* subspace optimization. L^{2}ight fixes the unitaries and only trains singular values. In Fig. 8, simply shining light forward and backward through the same MZI meshes will generate the first-order gradients for singular values. By trading full-space trainability, L^{2}ight enjoys the scalability of first-order gradient-based optimization. This brief demonstrates 1000× higher scalability with over 30× energy reduction compared to SoTA ONN training algorithms with superior noise tolerance and on-device transferability.

V. Challenges and Possible Research Directions

Besides the issues in the photonics part, the electronics-photonics integration is the most prominent challenge for optical neural accelerators. Most of the system-level complication still comes from the I/O and control. The overall system performance bottleneck is mainly on the data transaction from memory and analog-to-digital converters (ADCs). In terms of power consumption, nearly 50% of power is from electrical memory, and ADCs/DACs take another 20-30%, while the photonic circuit only consumes less than 10% total power [30]. The speed and power of memory and ADCs determine the overall benefits one could gain from optical neurocomputing.

We point to possible research directions.

Scaling to Larger Models: Promising directions include (1) *trading universality for higher scalability* [9], [24], [31] by restricting the matrix parameter space; (2) *squeezing tensor computations into customized devices*, e.g., map MVM onto MORRs [10], tunable multi-mode interference, or meta-lens [32]; (3) pipelined accelerators with a cluster of PTCs; (4) utilizing wavelength/time/mode-division multiplexing.

Nonlinearity: Though there exist optical nonlinearity, e.g., saturable absorbers, and electrical-optical nonlinear units [33], current activation function is still offloaded to electrical parts. It is of practical usage to design programmable optical nonlinearity with less energy loss and E-O conversion latency.

New Device/Material: Device-level innovation is essential in the ONN design stack, for example, phase shifters with higher tuning efficiency, phase change materials with shorter programming latency and higher lifetime, etc.

Cross-Layer Co-Design: System designs open a new dimension to optimize the performance of ONNs, including optimization on devices, PTC type and size, interconnects, on-chip memory, ADCs/DACs, etc. [17], [34]. Automated architecture search and PTC topology design are promising trends to propel ONN advances with design automation and AI algorithms [28], [35].

VI. Conclusion

Optical neural networks represent a promising computing paradigm for ultra-fast and energy-efficient AI, especially for resource-limited edge devices. Device-circuit-architecture-algorithm co-design is the key to scalable and robust optical NN accelerators. We give a tutorial and overview on SoTA co-design methodologies for area-efficient ONN architectures, robust ONN model optimization, and efficient on-chip training. We also highlight emerging directions from the device and circuit to system-level designs with great research opportunities to explore next-generation photonic AI.

References

[1] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nat. Photon.*, vol. 11, pp. 441–446, Jun. 2017.

[2] G. Wetzstein *et al.*, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, pp. 39–47, Dec. 2020.

[3] B. J. Shastri *et al.*, "Photonics for artificial intelligence and neuromorphic computing," *Nat. Photon.*, vol. 15, pp. 102–114, Jan. 2021.

[4] X. Xu *et al.*, "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, pp. 44–51, Jan. 2021.

[5] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, pp. 52–58, Jan. 2021.

[6] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nat. Commun.*, vol. 4, p. 1364, Jan. 2013.

[7] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightw. Technol.*, vol. 32, no. 21, pp. 4029–4041, Nov. 2014.

[8] C. Roques-Carmes *et al.*, "Heuristic recurrent algorithms for photonic ising machines," *Nat. Commun.*, vol. 11, p. 249, Jan. 2020.

[9] J. Gu, Z. Zhao, C. Feng, M. Liu, R. T. Chen, and D. Z. Pan, "Towards area-efficient optical neural networks: An FFT-based architecture," in *Proc. ASPDAC*, 2020, pp. 476–481.

[10] J. Gu *et al.*, "SqueezeLight: Towards scalable optical neural networks with multi-operand ring resonators," in *Proc. DATE*, Feb. 2021, pp. 238–243.

[11] J. Gu, Z. Zhao, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "ROQ: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *Proc. DATE*, 2020, pp. 1586–1589.

[12] Y. Zhu *et al.*, "Countering variations and thermal effects for accurate optical neural networks," in *Proc. ICCAD*, 2020, p. 152.

[13] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, 2018.

[14] J. Gu *et al.*, "Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization," in *Proc. AAAI*, 2021, pp. 7583–7591.

[15] Z. Zhao *et al.*, "Hardware-software co-design of slimmed optical neural networks," in *Proc. ASPDAC*, 2019, pp. 705–710.

[16] A.N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, p. 7430, Aug. 2017.

[17] W. Liu *et al.*, "HolyLight: A nanophotonic accelerator for deep learning in data centers," in *Proc. DATE*, 2019, pp. 1483–1488.

[18] J. Gu *et al.*, "Toward hardware-efficient optical neural networks: Beyond FFT architecture via joint learnability," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 9, pp. 1796–1809, Sep. 2021.

[19] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, Jan. 2020, Art. no. 31404.

[20] Z. Zhao, J. Gu, Z. Ying, C. Feng, R. T. Chen, and D. Z. Pan, "Design technology for scalable and robust photonic integrated circuits," in *Proc. ICCAD*, 2019, pp. 1–7.

[21] T. Zhang *et al.*, "Efficient training and design of photonic neural network through neuroevolution," *Opt. Exp.*, vol. 27, no. 26, pp. 37150–37163, 2019.

[22] J. Gu, Z. Zhao, C. Feng, W. Li, R. T. Chen, and D. Z. Pan, "FLOPS: Efficient on-chip learning for optical neural networks through stochastic zeroth-order optimization," in *Proc. DAC*, 2020, pp. 1–6.

[23] J. Gu, H. Zhu, C. Feng, Z. Jiang, R. T. Chen, and D. Z. Pan, "L2ight: Enabling on-chip learning for optical neural networks via efficient in-situ subspace optimization," in *Proc. NeurIPS*, 2021, pp. 8649–8661.

[24] C. Feng *et al.*, "Silicon photonic subspace neural chip for hardware-efficient deep learning," 2021, *arXiv:2111.06705*.

[25] Z. Ying, C. Feng, Z. Zhao, R. Soref, D. Pan, and R. T. Chen, "Integrated multi-operand electro-optic logic gates for optical computing," *Appl. Phys. Lett.*, vol. 115, no. 17, 2019, Art. no. 171104.

[26] S. Banerjee, M. Nikdast, and K. Chakrabarty, "Modeling silicon-photonic neural networks under uncertainties," in *Proc. DATE*, 2021, pp. 98–101.

[27] J. Gu, Z. Zhao, C. Feng, Z. Ying, R. T. Chen, and D. Z. Pan, "O2NN: Optical neural networks with differential detection-enabled optical operands," in *Proc. DATE*, 2021, pp. 1062–1067.

[28] J. Gu *et al.*, "ADEPT: Automatic differentiable design of photonic tensor cores," 2021, *arXiv:2112.08703*.

[29] H. Zhu *et al.*, "ELight: Enabling efficient photonic in-memory neurocomputing with life enhancement," in *Proc. ASPDAC*, 2022, pp. 332–338.

[30] C. Ramey, "Silicon photonics for artificial intelligence acceleration," in *Proc. HotChips*, 2020, pp. 1–26.

[31] H. H. Zhu *et al.*, "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.*, vol. 13, p. 1044, Feb. 2022.

[32] Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nat. Commun.*, vol. 13, p. 2131, Apr. 2022.

[33] 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," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7700412.

[34] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," in *Proc. DAC*, 2021, pp. 1069–1074.

[35] M. Li, Z. Yu, Y. Zhang, Y. Fu, and Y. Lin, "O-HAS: Optical hardware accelerator search for boosting both acceleration performance and development speed," in *Proc. ICCAD*, 2021, pp. 1–9.