Compute with Light: Architectures, Technologies and Training Models for Neuromorphic Photonic Circuits

Nikos Pleros, Miltiadis Moralis-Pegios, Angelina Totovic, George Dabos, Apostolos Tsakyridis, George Giamougiannis, George Mourgias-Alexandris, Nikos Passalis, Manos Kirtas, Anastasios Tefas

Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, npleros@csd.auth.gr

Abstract We discuss recent advances in the field of neuromorphic photonics, presenting our recent work and perspective towards optimizing the architecture, the enabling technology and the Deep Learning training models through a hardware/software co-design and co-development framework.

Introduction

The explosive growth of Artificial Intelligence (AI) computing and Deep Learning (DL) applications together with the growing maturity of photonic integration have created a new window of opportunity for the use of optics in computational tasks [1]. Exploiting photons for neural network (NN) hardware implementations expects to utilize the broadband signal carrying credentials of optical technologies together with their low-energy and low-footprint tunability properties. These properties can boost Multiply-Accumulate (MAC) operations within a small energy and area envelope, with computational energy and area efficiency estimations predicted to reach a few fJ/MAC and >TMAC/sec/mm^2, respectively[2-3]. Turning these expectations into a tangible reality requires, however, a synergistic co-design and co-development roadmap among all constituent scientific and technological fields, extending from underlying theory and architectures through co-integrated enabling technology platforms, all properly adapted to DL training models.

In this article, we provide a brief overview of the main neuromorphic photonic building blocks and the associated challenges, reviewing also the progress made along higher on-chip compute rates and low-power and small-size photonic computational elements. Motivated by respective advances in the field of analog electronic in-memory computing, we present how optimized linear optical architectures[4] can be designed for empowering an efficient synergy between WDM and linear optics. We discuss alternative technology roadmaps depending on whether inference or training applications are targeted, presenting recent research attempts to transform plasmonics into suitable computational modules. We provide an overview of recent experimental demonstrations of feed-forward and Recurrent photonic neural networks performing in MNIST[5-8] and time-series classification[10] , respectively, highlighting the importance of hardware-aware training models towards sustaining high-accuracy values at high compute line-rates[11-14].

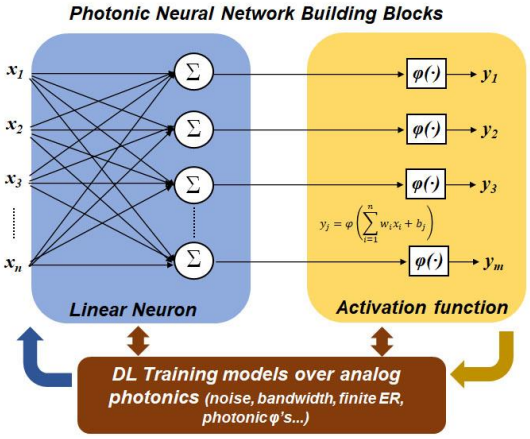

Fig. 1: The basic constituent building blocks towards realizing optical neural networks.

Challenges and state-of-the-art review

Transferring the neural network concepts and principles into a light-enabled platform has to proceed along optimization across all constituent NN pillars, shown in Fig. 1. A neural layer can be broken down into its linear and non-linear part, with the linear neuron stage being responsible for carrying out all necessary multiplication and summation functionalities. Multiplication is usually achieved either by controllable light absorption/amplification[8],[15-18] or by controlling transmittivity within interferometric and resonant modules[6],[19-21] while summation is easily offered through combiner and multiplexing elements. The non-linear segment has to facilitate the realization of the non-linear activation function[8],[22-24], with typical activation functions favored by DL models being the ReLU, PreLU, sigmoid and tanh. Last but not least, migrating from digital to analog NN engines that employ light for all their constituent functions, shapes a new framework for DL training models. Training methods have to incorporate at design-phase the noise[12], bandwidth, processing rate, finite extinction ratio and bit-resolution metrics of the underlying photonic hardware in order to safeguard high-accuracy performance. On top of that, DL training models have often to account for new activations enabled by photonics that are not within the existing portfolio of DL models, like for

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

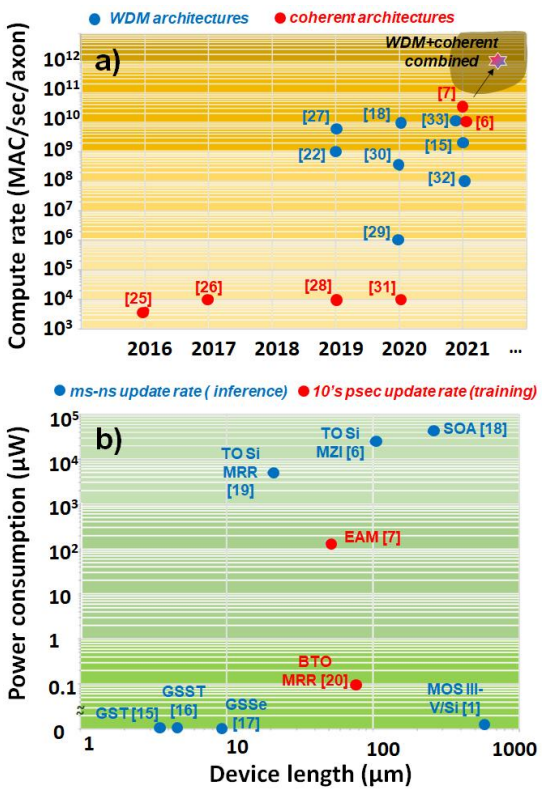

Fig. 2: (a) Compute rate per axon performance of WDM and coherent neuromorphic architectures demonstrated experimentally within the last 5 years, b) power consumption vs device length for the different optical weight enabling technologies, with red dots designating the potential for operating also in training applications.

example the $\sin^2(x)$ activation that can be easily offered in photonic NNs via the o/e/o conversion stage[6],[12].

The development of neuromorphic photonic engines towards meeting the high computational power and area efficiency expectations has to proceed along the challenge of operating at >10 Gs/sec line-rates[1-3] within a challenging up to 8-bit-resolution DL environment. Allowing for high compute rates per axon comprises the main approach to compensate for the higher integration densities supported by electronics, which typically invest in their size benefits for increasing computational density through respective increases in the number of synapses and neurons. Figure 2(a) illustrates the compute rate performance values in MAC/sec/axon reported by the rich variety of optical neural network experimental demonstrations presented within the last five years[6-7],[18],[22],[25-33]. Although different architectural schemes and different constituent integrated photonic technologies have been utilized in all these demonstrations, it can be easily observed that incoherent or WDM architectures[15],[18],[22],[27],[29-30],[32-33] were almost constantly within the GHz clock frequency operational area, allowing for 10GMAC/sec/axon compute rates when off-chip data modulation was employed[18],[27],[33]. However, incoherent layouts typically require a different wavelength per single axon within a neuron, necessitating a high amount of wavelength resources for increasing fan-in and total computational power[33]. Single-channel optical neural networks can be accomplished only through coherent photonic interferometric layouts[6-7],[25-26],[28],[31]. This field has been until recently dominated by unitary optical linear matrix designs, where, however, the need for multiple cascaded stages of 2x2 Mach-Zehnder interferometric meshes enforces a tight control over individual device loss uniformity and phase control. This has probably constrained operational line-rates in the sub-MHz regime[25-26],[28],[31]. The on-chip transfer of a novel interferometric scheme that employs dual-IQ-modulator-based computational cells, where weighting is accomplished through a single photonic module, has elevated coherent neurons to 10 GMAC/sec/axon compute rates[6]. More recently, the same architectural scheme was employed in a silicon-chip that uses SiGe Electro-Absorption Modulators (EAMs) for both on-chip data generation and weighting purposes, succeeding to extend on-chip compute rates to 32 GMAC/sec/axon[7]. Allowing for high compute rates in single-channel coherent layouts can pave the inroad towards efficiently synergizing WDM and coherent approaches for boosting performance at >100 GMAC/sec per synaptic element.

Energy efficiency is mainly dictated by the power consumption of the weighting technology, assuming that a typical N-input neuromorphic layout requires a N^2 number of weights for offering N^2 MAC operations. As such, weighting elements have to align to a challenging framework where low power consumption, small footprint and low insertion losses have to be met simultaneously[1,2,3]. An additional critical parameter relates also to their reconfiguration time or update rate: long reconfiguration times suggest a limited capability for updating weight values, implying their use only for inference functionalities[1]. In case the matrix dimensions are higher than the dimensions of the photonic matrix layout or if training applications are targeted, weight values have to be updated at fast time scales, requiring sub-nsec time scales for their reconfiguration times. Fig. 2(b) provides an overview of the main weight technology blocks utilized so far in neuromorphic photonic layouts, revealing the advantages of Phase-Change-Material (PCM) non-volatile memories for use within inference engines[15-17]. Training applications can be, however, sustained only through electro-optic structures that can support

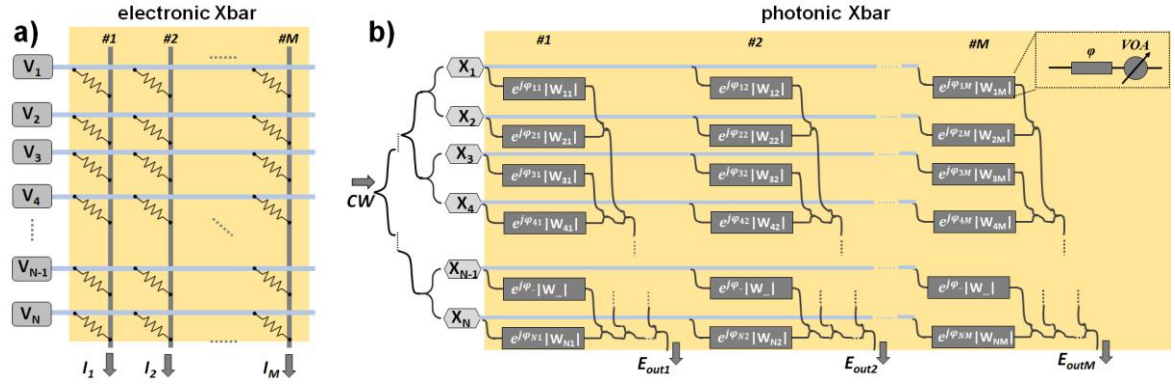

Fig. 3: a) the electronic crossbar layout performing as the linear neural layer stage in analog electronic neural networks, b) the corresponding analogous photonic crossbar[4], with the inset depicting that an optical phase shifter followed by a Variable Optical Attenuator (VOA) serves as the direct analogous of the electronic resistor-based weighting node.

10's of psec response times, like EAM[7] and BTO[20] waveguides, which have been so far utilized only in inference tasks.

PNN architectures, technologies and training

The use of electro-optic weight technologies for supporting fast weight update rates and training applications will lead to higher insertion losses per weighting module. This can't be easily sustained by state-of-the-art coherent neuromorphic schemes, which rely on conventional unitary meshes of 2x2 MZIs[25,26],[28],[31]. Inspired by the electronic crossbar architecture employed in analog electronic neuromorphic circuitry (Fig.3(a)), we have recently demonstrated a novel photonic crossbar design (Fig. 3(b)) that can support any linear transformation in the optical domain, while offering significant insertion loss and fidelity benefits compared to unitary layouts[4]. Fig. 4(a) and (b) depict the first silicon-based implementations of a single-column version of this crossbar[6,7], realizing a 4:1 neuron with thermo-optic weights[6] and a 2:1 neuron using EAM-based weights[7]. The same neuron architecture is targeted also within the H2020 research project PlasmoniAC[34]. PlasmoniAC aims, however, at a neuromorphic plasmo-photonic platform for naturally interfacing the upper electronic memory and control layer with the underlying photonic interconnect layer through the use of 3D co-integrated plasmonic weighting elements[35] (Fig. 4(c)). This research roadmap relies on the well-known size and energy benefits of plasmonics (Fig.4(d)) that may translate to respective advantages in neural layouts (Fig. 4(e)), provided that they can successfully migrate into computational devices.

Realizing high-accuracy neural networks through imperfect photonic elements can only take place through a properly adapted DL training framework[5-7],[11-14]. The benefits of hardware-aware training platform can be clearly highlighted in recent photonic RNN demonstrations[9], where a noisy SOA-based photonic RNN supporting only non-negative weight values turned into a high-accuracy and real-time 10 Gb/s 3- and 4-bit time series classifier[10].

Conclusions

We reviewed our recent work in the field of feed-forward and RNN photonic layouts, discussing the benefits of a hardware-software co-design framework and the roadmap towards energy-area efficient PNNs with high compute rates.

Acknowledgements

This work was supported by the EC through H2020 Projects PLASMONIAC (871391), SiPHO-G (101017194), NEBULA (871658) and by the GSRT through project DeepLight (4233).

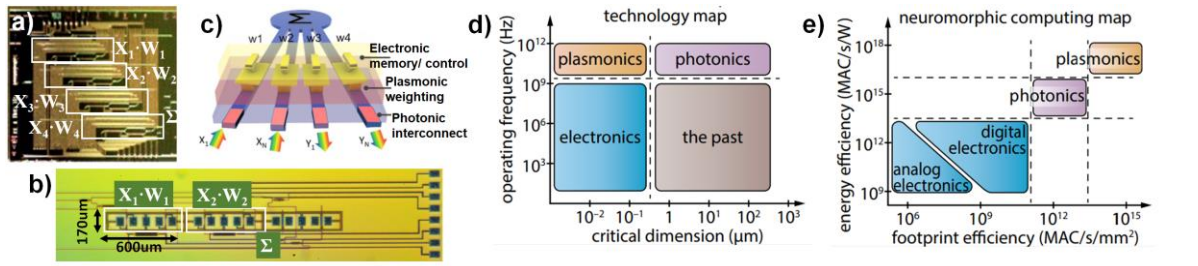

Fig. 4: First single-column photonic crossbar layouts validated experimentally as silicon-chips in a) 4:1 neuron layout with 10GMAC/sec/axon using Si-MZM technology[6], b) 2:1 EAM-based neuron layout with 32GMAC/sec/axon[7]. c) pictorial view of the 3D co-integrated plasmo-photonic platform pursued in H2020 project PlasmoniAC[34,35]. (d) Plasmonics frequency and size positioning within a generic technology map, (e) how plasmonics can be positioned in a neuromorphic computing map when migrating to computational devices.

References

[1] Shastri, B.J., Tait, A.N., Ferreira de Lima, T. et al. Photonics for artificial intelligence and neuromorphic computing. Nat. Photonics 15, 102–114, 2021.

[2] A. R. Totović et al, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," in IEEE J. of Sel. Topics in Quantum Electronics, vol. 26, no. 5, pp. 1-15, Sept.-Oct. 2020.

[3] M. Nahmias, et. al. "Photonic Multiply-Accumulate Operations for Neural Networks," IEEE JSTQE, 26 (1), 2020

[4] G. Giamougiannis et. al., "Coherent photonic crossbar as a universal linear operator", submitted to Physical Review X (2021)

[5] G. Mourgias-Alexandris et al., "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," in J. of Lightwave Technol., vol. 38, no. 4, pp. 811-819, 15 Feb.15, 2020

[6] G. Mourgias-Alexandris, et al., "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate ", Optical Fiber Comm. Conf., Tu5H.1, 2021.

[7] G. Giamougiannis et al, "Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells", submitted at Eur. Conf. on Optical Comm. 2021

[8] G. Mourgias-Alexandris et al, "An all-optical neuron with sigmoid activation function", Opt. Exp., Vol. 27, No. 7, pp. 9620-9630, Mar. 2019

[9] G. Mourgias-Alexandris et al, "All-optical WDM Recurrent Neural Networks with Gating", IEEE J. on Sel. Topics of Quantum Electron., Vol. 26, No. 5, pp. 1-7, Sept. 2020

[10] G. Mourgias-Alexandris et al, "A Photonic Recurrent Neural Network for Time-Series Classification", IEEE J. of Lightwave Technol., Vol. 39, No. 5, pp. 1340-1347, Mar. 2021

[11] N. Passalis et al, "Initializing Photonic Feed-forward Neural Networks using Auxiliary Tasks", Neural Networks, Vol. 129, pp. 103-108, Sept. 2020

[12] N. Passalis et al, "Training Deep Photonic Convolutional Neural Networks with Sinusoidal Activations", IEEE Trans. on Emerging Topics in Comp. Intel., Vol. 5, No. 3, pp 384-393, June 2021

[13] N. Passalis et al, "Sigmoid-based Recurrent Photonic Networks for High Frequency Financial Time Series Analysis Leveraging Noise-Aware Adaptive Initialization", 28th Eur. Signal Proc. Conf. (EUSIPCO) 2020, Amsterdam, Netherlands, Aug. 2020

[14] N. Passalis et al, "Adaptive Initialization for Recurrent Photonic Networks Using Sigmoidal Activations", IEEE Intern. Symp. on Circuits & Systems (ISCAS), Seville, Spain, May 2020

[15] Feldmann, et al. "Parallel convolutional processing using an integrated photonic tensor core", Nature 589, 52–58 (2021).

[16] M. Miscuglio et al, "Artificial Synapse with Mnemonic Functionality using GSST-based Photonic Integrated Memory", Intern. App. Comp. Electromagnetics Society Symp. (ACES), 2020, pp. 1-3

[17] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning",App.Phys.Rev. 7,031404 (2020)

[18] B. Shi et al, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," IEEE J. Sel.Top. Quantum Electron., vol. 26, no. 1, pp. 1–11, Jan. 2020

[19] A. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks", Sci. Rep., 7 (1), 2017

[20] J. Elliott Ortmann et al,"Ultra-Low-Power Tuning in Hybrid Barium Titanate–Silicon Nitride Electro-optic Devices on Silicon", ACS Phot. 2019, 6, 11, 2677–2684

[21] M. Takenaka et al., "III-V/Si Hybrid MOS Optical Phase Shifter for Si Photonic Integrated Circuits," J. Light. Technol., vol. 37, no. 5, pp. 1474–1483, 2019

[22] A. Tait et al., "Silicon Photonic Modulator Neuron", Physical Review Applied, vol. 11, no. 6, 2019

[23] J. K. George, "Neuromorphic photonics with electro-absorption modulators," Opt. Exp. 27, 5181-5191 (2019)

[24] J. Crnjanski et al, "Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron," Opt. Lett. 46, 2003-2006 (2021)

[25] A. Ribeiro et al, "Demonstration of a 4 × 4-port universal linear circuit," Optica 3, 1348-1357 (2016)

[26] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nat. Photonics, vol. 11, no. 7, pp. 441–446, 2017

[27] Y. Huang et al, "Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay," Opt. Express 27, 20456-20467 (2019)

[28] F. Shokraneh et al, "A Single Layer Neural Network Implemented by a 4x4 MZI-Based Optical Processor," in IEEE Photon. J., vol. 11, no. 6, pp. 1-12, Dec. 2019,

[29] C. Huang et al, "Demonstration of photonic neural network for fiber nonlinearity compensation in long-haul transmission systems," in Optical Fiber Comm. Conf. PDP 2020, Th4C.6., 2020

[30] T. F. de Lima et al, "Real-time Operation of Silicon Photonic Neurons," in Optical Fiber Communication Conference (OFC) 2020, M2K.4., 2020

[31] H. Zhang et al. "An optical neural chip for implementing complex-valued neural network", Nat Commun 12, 457 (2021)

[32] W. Zhang et al, "Microring Weight Banks Control beyond 8.5-bits Accuracy", arXiv preprint arXiv:2104.01164, 2021

[33] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks", Nature 589, 44–51 (2021).

[34] http://www.plasmoniac.eu/index.php

[35] R. Stabile et al, "Neuromorphic photonics: 2D or not 2D?", Journal of Applied Physics 129, 200901 (2021)