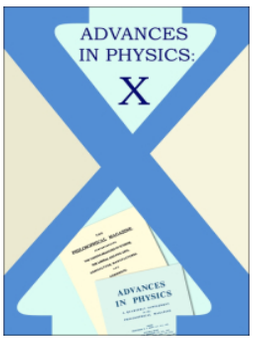

Advances in Physics: X

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tapx20

Prospects and applications of photonic neural networks

Chaoran Huang, Volker J. Sorger, Mario Miscuglio, Mohammed Al-Qadasi, Avilash Mukherjee, Lutz Lampe, Mitchell Nichols, Alexander N. Tait, Thomas Ferreira de Lima, Bicky A. Marquez, Jiahui Wang, Lukas Chrostowski, Mable P. Fok, Daniel Brunner, Shanhui Fan, Sudip Shekhar, Paul R. Prucnal & Bhavin J. Shastri

To cite this article: Chaoran Huang, Volker J. Sorger, Mario Miscuglio, Mohammed Al-Qadasi, Avilash Mukherjee, Lutz Lampe, Mitchell Nichols, Alexander N. Tait, Thomas Ferreira de Lima, Bicky A. Marquez, Jiahui Wang, Lukas Chrostowski, Mable P. Fok, Daniel Brunner, Shanhui Fan, Sudip Shekhar, Paul R. Prucnal & Bhavin J. Shastri (2022) Prospects and applications of photonic neural networks, Advances in Physics: X, 7:1, 1981155, DOI: 10.1080/23746149.2021.1981155

To link to this article: https://doi.org/10.1080/23746149.2021.1981155

Published online: 31 Oct 2021.

ADVANCES IN PHYSICS: X
2022, VOL. 7, NO. 1, 1981155
https://doi.org/10.1080/23746149.2021.1981155

Taylor & Francis
Taylor & Francis Group

REVIEWS

Prospects and applications of photonic neural networks

Chaoran Huang[a,b], Volker J. Sorger[c], Mario Miscuglio[c], Mohammed Al-Qadasi[d], Avilash Mukherjee[d], Lutz Lampe[d], Mitchell Nichols[d], Alexander N. Tait[e], Thomas Ferreira de Lima[a], Bicky A. Marquez[f], Jiahui Wang[g], Lukas Chrostowski[d], Mable P. Fok[h], Daniel Brunner[i], Shanhui Fan[g], Sudip Shekhar[d], Paul R. Prucnal[a] and Bhavin J. Shastri[f,j]

[a]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA; [b]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China; [c]Department of Electrical and Computer Engineering, The George Washington University, Washington DC, DC, USA; [d]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada; [e]National Institute of Standards and Technology, Boulder, Colorado, USA; [f]Department of Electrical Engineering, Queen's University, Kingston, Canada; [g]College of Engineering, Stanford University, Stanford, CA, USA; [h]University of Georgia, Athens, GA, USA; [i]Department of Physics, Engineering Physics & Astronomy, Institut FEMTO-ST, Université Bourgogne-Franche-Comté CNRS UMR, Besancon, France; [j]The Vector Institute, Toronto, Canada

ABSTRACT

Neural networks have enabled applications in artificial intelligence through machine learning, and neuromorphic computing. Software implementations of neural networks on conventional computers that have separate memory and processor (and that operate sequentially) are limited in speed and energy efficiency. Neuromorphic engineering aims to build processors in which hardware mimics neurons and synapses in the brain for distributed and parallel processing. Neuromorphic engineering enabled by photonics (optical physics) can offer sub-nanosecond latencies and high bandwidth with low energies to extend the domain of artificial intelligence and neuromorphic computing applications to machine learning acceleration, nonlinear programming, intelligent signal processing, etc. Photonic neural networks have been demonstrated on integrated platforms and free-space optics depending on the class of applications being targeted. Here, we discuss the prospects and demonstrated applications of these photonic neural networks.

ARTICLE HISTORY
Received 27 April 2021
Accepted 9 September 2021

KEYWORDS
Photonic neural networks; neuromorphic photonics; silicon photonics; neuromorphic computing; machine learning; artificial intelligence

CONTACT Bhavin J. Shastri bhavin.shastri@queensu.ca Queen's University, Kingston ON K7L 3N6, Canada

1. Primer on artificial intelligence, machine learning, neuromorphic computing, and neuromorphic photonics

Creating a machine that can process information like human brains has been a driving force of innovations throughout history. Artificial intelligence (AI) has been coined as an academic discipline since the 1950s [1]. This field underwent the first surge of optimism from the 1950s to 1970s, however, followed by decades of setbacks. The biggest obstacle at that time was the lack of computing power noauthor_history_2017. In the last decade, AI has experienced explosive growth. Three sources fuel the advancement of AI: (1) substantial development of AI algorithms, especially in machine learning and neural network models (alias for 'deep learning' [2]); (2) the abundant amount of available information in the 'big data' era; and (3) the rise of computing power as predicted by Moore's Law, together with new hardware (e.g. graphics processing unit (GPU)) and infrastructures (e.g. cloud-based servers).

State-of-the-art AI algorithms, by and large, are implemented using neural networks, a computing model inspired by the brain's neuro-synaptic framework. Today, nearly all AI algorithms are running on digital computers based on von Neumann architecture, a computing architecture that has dominated computing design since it was invented but is nothing like the brain. This architecture consists of a centralized processing unit (CPU) that performs all operations specified by the program's instructions and a separate memory that stores data and instructions. It processes information sequentially in a serialized manner. However, neural network models are radically different from von Neumann architecture in some key features. First, neural networks are highly parallel and distributed, whereas von Neumann architecture is

inherently sequential (or, in the best case: sequential-parallel with multi-processors). Second, in neural networks, computing units (neurons) and storage units (synapses) are co-located. In contrast, computing units (CPUs) and storage units (dynamic random access memories (DRAMs)) are physically separate chips in digital computers. The sharp contrast between the two architectures slows down the computing speed and increases the power consumption, which, as a result, necessitates reinventing conventional computers for efficient information processing.

Neuromorphic (i.e. neuron isomorphic) computing promises to solve these problems by creating radical new hardware platforms that can emulate the underlying neural structure of the brain. The general idea is to build circuits composed of physical devices that mimic the neuron biophysics interconnected by massive physical interconnects with co-integrated non-volatile memories. In doing so, neuromorphic hardware could break performance limitations inherent in von Neumann architectures and gain advantages in speed and efficiency in solving intellectual tasks. Achieving this goal requires significant advances in a wide range of technologies, including materials, devices, device fabrication, system integration, platform co-integration, packaging etc [3–5].

Neuromorphic hardware has been built in electronics on various platforms, including traditional digital CMOS [6–8] and hybrid CMOS-memristive technologies (discussed next) [9,10]. Neural network models highlight the essential needs of high-degree physical interconnections, which, in electronic neuromorphic hardware, is achieved by incorporating a dense mesh of wires overlaying the semiconductor substrate as crossbar arrays. Unfortunately, electronic connections fundamentally suffer harsh trade-offs between bandwidth and interconnectivity [11,12]. A major limitation for neuromorphic electronics is interconnect density, thus confining the neuromorphic processing speed and associated application space within the MHz regime.

Photonics has unmatched feats for interconnects and communications in terms of bandwidth, which can negate the bandwidth and interconnectivity trade-offs [5,13–15]. The advantages of photonics for neural networks were recognized decades ago. The photonic neural network research was pioneered by Psaltis and others who adopted spatial multiplexing techniques enabling all-to-all interconnection [16]. However, low-level photonic integration and packaging technologies hindered the practical applications of photonic neural networks at that time. Nevertheless, the landscape of photonic neural networks has changed tremendously with the emergence of large-scale photonic fabrication and integration techniques [17,94,5,18]. For example, silicon photonics provides an unprecedented platform to produce large-scale and low-cost optical systems [18–20]. In parallel, a broad domain of emerging applications (such as solving nonlinear

optimization problems or real-time processing of multichannel, gigahertz analog signals) is also looking for new computing platforms to fulfill their computing demands [De Lima et al., 2019; 21–24]. All these changes have shed light on new opportunities and directions for photonic neural networks [5,25].

This paper is intended to provide an intuitive understanding of photonic neural networks and why, where, and how photonic neural networks can play a unique role in enabling new domains of applications. First, we discuss the challenges of digital versus analog approaches in implementing neural networks. Next, we provide a rationale for photonic neural networks as a compelling alternative for neuromorphic computing compared to electronic platforms. Then, we outline the primary technology required for evolving neuromorphic photonic processors, review existing approaches, and discuss challenges. In the subsequent sections, we provide a survey of new applications enabled by photonic neural networks and highlight the role of photonic neural networks in addressing the challenges in these applications.

2. Digital vs. analog neural networks

In this Section, we briefly compare the state-of-the-art electronic implementations of neural networks in digital and analog domain. We establish the advantages and limitations of analog implementations in general to then make the case for analog photonic implementations in the following Sections.

Deep neural networks (DNN) model complex nonlinear functions by composing layers of linear matrix operations with non-linear activation functions. Computationally, DNNs are mostly matrix-multiplication, with matrix-multiplications taking more than 90% of the total computations in a DNN [26–28]. Due to the underlying array-based operation in the matrix multiplication, digital electronic neural network hardware is usually composed of basic units, referred to as processing elements (PEs), in a 2D array structure [Y.-H. 29]. Such a structure enables the matrix multiplication operation to be N $\times$ faster than CPUs, where N is the input vector length. Usually, PEs are composed of digital multipliers and adders, with precision up to 32 bits, to perform a single multiply and accumulate (MAC) operation, similar to the arithmetic logic unit in the CPU core.

Since DNNs consume a huge chunk of energy in data movement, many digital neural network hardware focus on optimizing dataflow to save energy. Based on the connection of PEs and the interconnects, various dataflows can be described. An output-stationary dataflow performs all the MAC operations for a single output before moving to the next. All the inputs and weights required are fetched from the memory, multiplied, and added to the partial sum, which is stored inside PE [Y.-H. 30]. On the other

hand, a weight-stationary dataflow holds the weights inside PE to maximize weight reuse. The partial sum accumulation occurs across multiple PEs while the input vector is fed in a staggered style allowing the PEs to perform MAC operation with the internally stored weights. Further dataflow optimizations combining different types of data reuse are also possible to reduce energy further [Y.-H. 30, 31].

Implementing the MAC operations in analog domain can help in reducing the energy consumption. Analog electronic elements, such as charge, current and time can be used to represent the data values. *An inherent advantage in analog techniques is the built-in addition operation without requiring additional circuits*. To perform the MAC operation with analog electronics, switched-capacitor techniques charge a capacitance sized proportionally to the weight with a current sized proportionally to the input [32], current-steering techniques control the magnitude of current flowing through transistors [33], and time-domain techniques modulate the pulse width of a signal using controlled oscillators (Cao, Chang, & Raychowdhury, 2020). Such analog techniques have shown to decrease energy significantly for small DNN models: 4 × using switched-capacitor on BinaryNet, 67 × using current steering on Matched Filter, and 1.4 × using time-domain techniques on mobile reinforcement learning. The shortcomings of analog techniques include: limited size of DNN models, low bit precision ($<$ 4 b) and associated accuracy loss, analog-to-digital converter/digital-to-analog converter (ADC/DAC) overhead, susceptance to noise and process, voltage and temperature (PVT) variations.

Analog implementations have a direct consequence for noise and noise propagation [34]. Since the probability of corrupting a symbol usually is identical for all bits in a sequence, the impact of a noise-induced Boolean symbol modification can be dramatic. Compared to that the corruption of an analog signal is usually more subtle as signal perturbations are mostly proportional to noise amplitude. Digital encoding therefore requires that thresholding levels significantly exceed all noise amplitudes; however, the signal propagation is then practically noiseless as the noiseless symbolic representation is continuously re-established. Furthermore, increasing a digital signal's resolution is comparatively economic as the number of digitization levels grows exponentially.

Well-designed circuits readily approach the thermodynamic noise limit to better than one order of magnitude [35]. Thermodynamics, therefore, establishes the link between such information centered arguments and the energy fundamentally required for a certain Signal-to-noise ratio (SNR). Digital encoding is penalized with a large constant energy penalty due to the required high encoding fidelity. However, the logarithmic scaling of digital encoding with precision in comparison to the quadratic scaling of analog encoding makes digital more energy efficient beyond SNR 10^4 [36–41].

Recently digital implementations of neural networks significantly reduced their bit-precision, with several systems today running with 8 or less bit resolution during inference. Finally, spiking NNs occupy a middle ground and potentially are superior in energy efficiency to analog for SNR $>10^2$ and to digital for SNR $<10^7$ [36].

Finally, the accumulation of noise can be strongly managed using the connections of a neural network. Studies based on linear, symmetric, i.e. untrained networks of noisy linear neurons show that neural network analog in and output neurons are the chief noise source, while in particular noise uncorrelated across neurons is essentially fully suppressed through the network's connections [42]. New studies in fully trained networks of noisy nonlinear units show that nonlinearity also efficiently decorrelates noise from correlated noise-populations [43]. This is important as such noise can for example be induced by a common power supply. Finally, rather weak requirements allow to fully freeze the propagation of noise through a network, and an analog photonic neural network's output can, therefore, approach the SNR of a single neuron [43].

Another possible method to reduce data movement energy is moving the computing inside the memory modules itself. Such architectures are referred to as In-Memory computing (IMC), and use the memory cells as an analog circuit to perform the MAC operations, generally in a weight-stationary dataflow. The inputs are analog currents or voltages on the wordlines, while their weights are either binary, ternary or digitally stored over multiple memory cells. The accumulation happens inherently in the bitlines in the memory array, resulting in an analog output [44, Jintao Zhang, Zhuo Wang, & Verma, 2016]. IMC architectures, while significantly enhancing the throughput of the system, operate in analog which call for adding more system level design considerations to meet the output signal to noise ratio (SNR) requirements and size.

IMC can also be performed in crossbar arrays of emerging memories, such as resistive RAM (ReRAM), conductive bridging, magnetic tunnel junctions, and phase change memories [45]. Explicit multipliers, adders and PE interconnects are not needed in IMC. Rather, the equivalent PE array in digital is implemented in IMC in just the area required to implement the memory array, plus the ADC and DAC at the array periphery [45]. Increased area efficiency allows packing far more parallel units, hence processing operations can be accelerated by more than an order of magnitude. For comparison, the area required to implement the PE in digital implementations is 13.4×10^6 F^2 [Y.-H. 29,47,49], whereas the average memory cell sizes in IMC are less than 100 F^2 [46], where F is the minimum feature size of the technology. Furthermore, energy consumption can also

be reduced by an order of magnitude over the equivalent digital systems, since several MAC operations are performed almost at the cost of a single read operation of the memory array.

IMC suffers from constraints similar to the analog MAC implementations. Furthermore, the nonidealities of analog memory cells and their interconnects pose limitations for achieving high accuracy and scaling [47]. For example, the nonlinearities of memory cells and the resistance of the interconnect in ReRAM IMC was shown to degrade the accuracy of computation with scaling (Peng Gu et al., 2015). Noise limitations have shown to saturate the computing accuracy of analog crossbars to 8 bits [48]. In comparison with digital implementations, the energy, area and latency advantages from IMC have been shown to reduce with the increased precision 8,804,680. But precision reduction techniques such as variable layer precisions and nonuniform quantization [26,49], have shown equivalent accuracy to 32-bit digital implementations even after reducing precision down to 2-bit [J. 50].

3. The case for photonics for neuromorphic processors

In an artificial neural network (ANN), neurons are interconnected by synaptic weights (a memory element) (Figure 1). Signals from many neurons are weighted before being summed by the receiving neuron. This many-to-one (N:1) connection is called fan-in, and the weighted sum–a linear operation–is the dot product of the output from connected neurons attenuated by a weight vector. A neuron then performs a nonlinear operation on the weighted sum to implement a thresholding effect which is output to many neurons. This one-to-many (1:N) connection is called fan-out.

Electronic and photonic neuromorphic approaches to implement neural networks face different challenges. Physical laws that restrain electronics do not necessarily apply to optics. For high-speed data transfer, compared to electronic wires, optical waveguides have a lower attenuation, have no inductance (minimal frequency-dependent signal distortions), and photons hardly interact with other photons (unless intermediated by matter) which

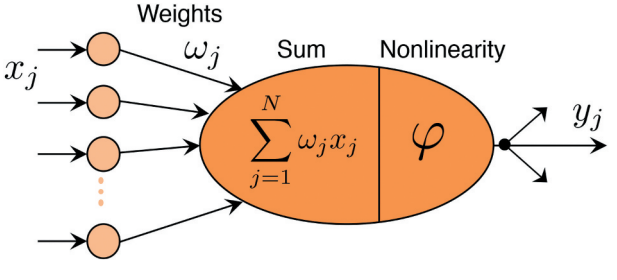

Figure 1. Artificial neuron model.

allows for wavelength multiplexing. While these fundamental properties have proven to be advantageous in optical communications, they are also important for neural networks including interconnectivity [Goodman, Leonberger, Sun- Yuan Kung, & Athale, 1984; 51] (i.e. neuron-to-neuron, neurons-to-neurons and neurons-to-memory communications), and parallel and linear processing, i.e. matrix multiplication [51]. However, these same properties make it challenging to implement nonlinear operations [52]. We refer the reader to the quantitative analyses comparing electronic and optical interconnects, and electronic and photonic computing performed by Miller [11,51] and Nahmias et al [12,53], respectively. Here, we qualitatively summarize these concepts comparing and contrasting optics and electronics.

Interconnects: Conventional electronic processors rely on point-to-point memory processor communication and can take advantage of state-of-the-art transmission line and active buffering techniques. However, a neuromorphic processor typically requires a large number of interconnects (i.e. hundreds of many-to-one fan-in per processor) [54] where it is not practical to use line and active buffering techniques for each connection at high speed. This creates a communication burden which in turn, introduces fundamental performance challenges that result from RC and radiative physics in electronic links, in addition to the typical bandwidth-distance -energy limits of point-to-point connections [51]. While some electronic neuromorphic processors incorporate a dense mesh of wires overlaying the semiconductor substrate as crossbar arrays, large-scale systems employ time-multiplexing or packet switching, notably, address-event representation [AER). These virtual interconnectivities allow electronic approaches to exceed wire density by a factor related to the sacrificed bandwidth. As argued by 5, for many applications, however, bandwidth and low latency are paramount, and these applications can be met only by direct, non-digital photonic broadcast (that is, many-to-many] interconnects.

Apart from bandwidth, power dissipation in interconnects is another major challenge in neuromorphic processors. Data movement has posed severe challenges to the power consumption of today's digital computers as parallel computing is widely deployed. A large amount of data has to move among processors and memory units, especially in those distributed programming models like neural networks. In these models, most energy is lost in data movement [55] due to capacitive charge-discharge events at high frequencies. In contrast, optics is more energy efficient in data movement at high speeds.

Linear operations and weighting: Optical physics is very well suited for linear operations. The electric field or intensity of the light can also be used as an analog metric for neuromorphic computations. The analog input in the electrical domain (voltage) can be used to modulate components such as

Mach-Zehnder modulators or ring resonator modulators. The weights (i.e. linear operations) can be implemented by scaling the field with Mach-Zehnder interferometers (MZIs) [56] or the intensity with RRs [Tait, Jayatilleka, et al., 2018; 57]. Optical signals can be added through wavelength-division multiplexing (WDM) by accumulation of carriers in semiconductors [57], electronic currents [58,59] or changes in the crystal structure of a material induced by photons [17,60–62]. Such photonic neural networks, being analog in nature, share the advantages and constraints of their analog electronic counterparts. In addition, modulations can be done at 10 s of GHz, addition can be done at the speed of light in coherent MZI networks and at 10 s of GHz in ring resonator networks, and weight multiplication can be done at the speed of light, all of which provides much lower latency than analog or digital electronic MAC operations.

Dynamic power consumption is also much lower, since multiplication does not consume any energy and energy efficiency of the E/O modulation can be reduced to as low as 10 fJ per operation [63]. However, static power consumption is dependent on laser wall plug efficiency, waveguide losses, energy consumed in maintaining the weights, etc., and must be minimized. Assuming an estimated energy consumption of $>$ 100 fJ per operation for digital implementation [64], analog IMC brings 2.1–7.8× reduction [46,65]. Photonic implementations can achieve energy consumption similar to IMC, but with orders of magnitude reduction in latency. Like their IMC counterparts, photonic neural networks can operate up to 8 bits of precision [58, 66, 67, 68].

Nonlinear operations: The same properties that allow optoelectronic components to excel at linear operations and interconnectivity are at odds with the requirements of nonlinear operations for computing [5]. The implementation of photonic neurons relies on the nonlinear response of optical devices. The approaches fall into two major categories based on the physical representation of signals within the neuron: optical-electrical-optical (O/E/O) vs. all-optical.

O/E/O neurons involve the transduction of optical power into electrical current and back within the signal pathway. Their nonlinearities occur in the electronic domain and in the E/O conversion stage using lasers or saturated modulators [69–72]. In the authors' recent work [72], neurons are implemented using silicon modulators that exploit the nonlinearity of the electro-optic transfer function. Modulation mechanisms can change the real and imaginary parts of the refractive index of the material and subsequently alters the index (termed E/O modulators) and loss (termed electro-absorptive modulators (EAMs)) of the optical propagating mode. Research on high speed and power-efficient modulators are very active, with a general focus on maximizing the interaction between the active material and the light. Such approaches include lithium niobate (LiNbO3) modulators based

on the Pockels effect [C. 73], III–V semiconductor- based quantum-confined Stark effect modulators [74], silicon modulators based on plasma dispersion effect [75, Q. 76], and hybrid modulators incorporating novel materials (such as ITO and graphene) to silicon-based modulators [69, 77, 78, M. 79].

All-optical neurons rely on the semiconductor carriers or optical susceptibility that occur in many materials. A perceived advantage of all-optical neuron implementations is that they are inherently faster than O/E/O approaches due to relatively slow carrier drift and/or current flow stages in the latter. All-optical perceptron has been demonstrated based on single-carrier optical nonlinearities, including through carrier effect in MRR [80–86], changing a material state [17,87], such as via a structural phase transition, and saturable absorbers and quantum assemblies heterogeneously integrated in photonic integrated circuits (PICs) [88].

Implementing on-chip optical neurons remains a challenge. The main challenge comes from maintaining layer-to-layer cascadability, which describes the ability of one neuron excited with a certain strength to evoke at least an equivalent response in a downstream neuron Lima et al. [2019). Maintaining the neuron cascadability must compensate for the optical device's insertion loss, the loss due to fan-out, and less than unity device efficiency. A straightforward approach is to introduce active amplifiers providing energy gain in the optical or electrical domain. For example, Tait et al. 72,demonstrated an O/E/O neuron composing a silicon modulator and photodetector pair both having a capacitance of a few tens of femtofarads. Thus this neuron requires a voltage swing of a few volts, which can be provided by a transimpedance amplifier (TIA] [28]. The monolithic silicon photonic integration platform (i.e. zero-change silicon photonics) [89] can provide seamless integration of photonic devices and microelectronic modules. In this approach, cascadability has to trade-off with power consumption at this system level.

Cascadability highlights the importance of studying power-efficient optical devices, i.e. optical modulators with low switch voltage, sensitive photodetectors, or nonlinear optical devices with low nonlinear threshold power. For this purpose, a common idea is to optimize the device geometry to maximize the optical mode overlap with the active material region. The emerging nanoscale devices [63,90], based on nano waveguides, photonic crystal nanocavities, or plasmonic nanocavities, can concentrate the propagating mode's field into a nanometer-thin region that overlaps with the actively index-modulated material. Another approach is to explore novel materials, such as graphene [M. 79], Indium Tin Oxide [91] and Lithium niobate on insulator [C. 73], and III–V hybrid integration [78,92–96]. Similarly, all-optical neurons based on third-order nonlinearity are

improved using InP-based two-dimensional photonic crystal nanocavity with quantum wells (QWs), paving the way to cascadable all-optical spiking neurons [97].

The highly-efficient modulation techniques can be potentially applied to the linear operation block for high energy-efficient and high-speed weight tuning. However, linear and nonlinear operation blocks consider different trade-offs in choosing modulation devices. For example, nonlinear operations require fast modulation (10s of GHz) to improve the line rate of a photonic neural network. Silicon modulators based on plasma dispersion effect are widely used for high-speed [> 40 G) modulation for optical interconnects in data centers. Exploring this modulator technology, fast O/E/O neurons are demonstrated by 72,using a silicon microring modulator with a reverse-biased PN junction embedded in the microring, and 98,with a Mach-Zehnder type modulator. On the other hand, for linear operations, the tuning speed sometimes is not the paramount metric to be considered. For example, in deep learning inference, once the neural network is trained, the weights don't need to be changed often. In these applications, power consumption is a more important parameter. Power consumption comes from the signal power dissipated during its propagation (i.e. device insertion loss] and the power needed to tune and hold the weight values. Therefore, modulation devices with low insertion loss and low tuning power can be better candidates for linear operation blocks. Silicon modulators used for nonlinear operations have a typical insertion loss of $\sim$ 4–5 dB. Apparently, the insertion loss of a silicon modulator needs to be further optimized for linear operations, because the loss would accumulate along with the large-scale Mach-Zehnder meshes. Micro-heaters based thermal tuning is easy to use and is most widely used for reconfigurable photonic circuits, including photonic neural networks. The Mach-Zehnder modulator with heaters has a much smaller insertion loss of $\sim$ 0.3 dB, however it dissipates several milliwatts of electrical power when tuning and holding the states. Non-volatile actuators using phase change materials, can set their state and then maintain their state without needing a power to hold the state, thus consuming almost zero power for matrix computations. Table 1 shows a variety of modulation mechanisms and state-of-the-art devices from the literature.

Memory: On-chip nonvolatile memories that can be written, erased, and accessed optically are rapidly bridging a gap toward on-chip photonic computing [106]; however, they cannot usually be written to and read from at high frequencies. As described by 5, future scalable neuromorphic photonic processors will need to have a tight co-integration of electronics with potentially hybrid electronic and optical memory architectures, and take advantage of the memory type (volatile versus non-volatile) in either digital or analog domains depending on the application and the computation been performed.

Table 1. Efficiency and speed of various index modulation techniques on silicon photonics. Options for phase modulation of silicon waveguides. a) thermal tuning with TiN filament; b) thermal tuning with embedded photoconductive heater; c) PN/PIN junction across the waveguide for injection and/or depletion modulation; d) III–V/Si hybrid waveguide; e) metal-oxide-semiconductor (MOS), where the 'metal' is actually an active semiconductor; f) lithium niobate cladding adds a strong electrooptic effect; g) 2 single-layer-graphene (SLG) 'capacitor'; h) non-volatile phase change material. [1]This bandwidth was not yet shown experimentally. A big challenge is to reduce the contact resistance with Graphene, reducing RC-loading effect. [2]Not experimentally shown at high-speed. [3]Demonstrated up to 20 MHz.

Modulation Effect	Speed	Efficiency	Ref.
Thermo-optic TiN [a)	5.6μs	$P_\pi L = 6.8$mW mm	99,
Thermo-optic N+/N/N+ Si [b]	μs	$P_\pi L = 0.8$mW mm	Jayatilleka et al. (2015)
Reverse-biased PN [c)	41 GHz	$V_\pi L = 46$V mm	100,
Graphene SLG [g]	30 GHz[1]	$V_\pi L = 28$V mm	101,
LiNbO3/Si Hybrid [f]	70 GHz	$V_\pi L = 22$V mm	102,
III–V MQW/Si Hybrid (d]	27 GHz	$V_\pi L = 2.4$V mm	H.-W. Chen et al. (2011)
III–V/Si MOS [e)	2.2 GHz	$V_\pi L = 0.9$V mm	103,
ITO MOS [e]	GHz[2]	$V_\pi L = 0.52$V mm	104,
Forward-biased PIN [c]	0.5 GHz	$V_\pi L = 0.36$V mm	105,
PCM (h]	0.8 GHz[3]	$E_\pi = 400$pJ	Ríos et al. (2015)

Current approaches with photonic neural networks are driven by electronic circuits or micro-controllers to load matrices. The integration and packaging of large-scale optical and electronic circuits can be challenging in terms of cost and power. In some machine learning applications [such as deep learning inference) the weights, once trained, do not have to be updated often or at all. In these cases, the integration of novel photonic memory technologies can limit the need to read from and write to electronic memories with DACs and ADCs. All-optical memories have been demonstrated using various optical components such as nonlinear switches 107, MZIs (108], laser diodes [109], semiconductor optical amplifiers (SOAs), bandpass filters (BPFs) and isolators [J. 110]. The access time of an optical memory cell is small and attractive for photonic neural networks [111]. Non-volatile photonic memories with phase-change materials (PCMs) set and retain the weights without further holding power after being set [17], resulting in almost zero power consumption in performing matrix multiplication operations. Here a crystalline phase transition (from amorphous to crystalline and back) constitutes reversible (WRITE & RESET) memory programmability, therefore enabling dynamic synaptic plasticity and online learning in photonic neural networks. For online training of the photonic neural network, a high-speed WRITE would be ideal, and experimentally MHz speeds are possible for known PCMs such as Ge2Sb2Te5 (GST) or GST alloys, just limited by the heat-capacitance of these photonic RAM. The READ speed, interestingly, is on the order of ps, and simply given by the time-of-flight of the signal photon through such photonic RAM. However, significant improvements must be made in reducing the energy consumption (taking into account optical losses) and size, reliability and ease of

interfacing for their practical deployment in photonic neural networks. The current prominent material GST has prohibitively high optical losses even for the lower-loss amorphous state with an extinction coefficient (kappa = 0.2), thus leading to high insertion loss. Future research should focus on low-optical loss solutions. Emerging P-RAMs would also eliminate the long-standing memory access bottleneck known from electronics, and a successful P-RAM integration constitutes a similar shift in optics from centralized to decentralized computing, known as in-memory computing, one of the very active research fields left in circuit design, after transistor scaling stopped.

4. Architectures of neuromorphic photonic processors

Neuromorphic processors and photonic neural networks alike require a number of fundamental building blocks; i) synaptic MAC operation, ii) nonlinear activation function (NLAF); iii) state-retention, i.e. memory, iv) data input/output (I/O) ports and, depending on the applications, data domain crossings such as v) photonic to electronic and/or vi) analog-to-digital, with the latter including DAC and the ADC counterparts. However, a generic architecture of a current state-of-art photonic neural network system is given in Figure 2.

Photonic synaptic MAC operations and VMMs: The most common form of photonic neural networks focuses on accelerating the mathematical computation of multiplications using optics; this is not surprising, since photonic programmable circuits allow for a non-iterative mathematical multiplication by simple preparing the state of the programmable element followed by sending the optical signal through this element. Then, the multiplication is performed 'on-the-fly' at pico-second delays in PICs enabling a notion of 'real-time' (i.e. zero-delay) computing, which is incidentally not to be confused with computing schemes that 'expect' a system to complete a task at a pre-determined and deterministic time. Options for VMM are plenty, but the most common ones are a mesh network of cascaded MZIs, or MRR filters, or a photonic tensor core (PTC) processor [56,117,118]. Accelerating VMMs is a worthy aim, since the mathematical computational complexity of VMMs scales with N3, where N is the matrix size (assuming a square matrix). However, the mathematical complexity is technically speaking irrelevant (at first order), since for hardware implementations considered here, the runtime complexity is actually of interest. And it is this runtime complexity that is non-iterative in analog photonic neural networks (assuming the neuron's weights are set, i.e. programmed).

ully programmable VMM do still present a formidable challenge to physical implementations considering the size of competitive neural networks. In particular the optimization of each matrix element is

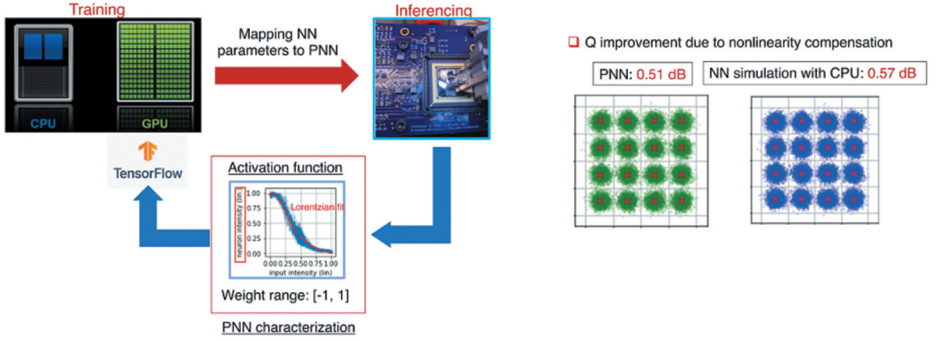

Figure 2. Generic system schematic of an analog photonic neural network or photonic tensor core accelerator. The digital-to-analog domain crossings are power costly and ideally a photonic DAC without leaving the optical domain, e.g. Ref [112] would be used instead, thus reducing system complexity and hence allowing for extended scaling laws. The optical processor itself can perform different operations depending on the desired application; is the processor a tensor core accelerator, then it will perform MAC operations such as used directly in VMMs or convolutions, e.g. Ref. 138. Is the processor a neural network, then nonlinearity and summations are needed in addition to linear operations. For higher biological plausibility such as for temporal processing of signals at each node, then mapping of partial differential equations onto photonic hardware is needed. This includes spiking neuron approaches where event-driven scheme play a role) [72,85,113–116].

technologically challenging. However, one can leverage reservoir computing (Jaeger and Haas (2004)) to mitigate this challenge to a certain degree. In reservoir computing, the connection matrix creating the neural network's hidden state can be ad-hoc defined, and training is restricted to the readout layer. This allows to encode the VMM of hidden layers using photonic components creating large scale random projections fully in parallel. First hardware implementations used photonic delay systems (Brunner, Soriano, Mirasso, and Fischer (2013); Larger et al. (2012)) leveraging temporal multiplexing of a single nonlinear photonic element and demonstrating photonic neural networks comprising 100 s of photonic neurons. However, full parallelism can only be obtained when avoiding temporal multiplexing. Parallel reservoirs were demonstrated using spatially multiplexed photonic neural networks integrated on a silicon chip (Vandoorne et al. (2014)) and based on random scattering in an electro-optical system [119,120] which surpassed a high-performance GPU in terms of scalability while achieving comparable computing accuracy. Utilizing the coupled modes of a large-area vertical-cavity surface emitting laser as a reservoir, and a digital micro-mirror to realize programmable readout weights [121], a fully hardware implemented photonic reservoir computing providing results without pre- or post-processing was demonstrated (Porte et al. (2021)). Generally, reservoir computing architectures offer competitive performance in various benchmark tasks, and their conceptual simplicity provides photonic systems with a viable strategy to fully leverage their parallelism.

NLAF, or threshold: Depending on the underlying algorithm model of the neural network, the NLAF can be as simple as a step-function such as used in binary classification problems, or a more elaborate function such as a sigmoidal, tangent hyperbolic, or population growth functions. However, when performing neural network training using gradient descent (GD) backpropagation (BP) methods, the problem of vanishing (or exploding) gradients can occur, where during each training cycle (called epoch) the available gradient on which a differentiation is performed is becoming consecutively smaller to the point of noise-level dominated and training would seize. To prevent this, rectifying linear units (ReLU) or Gaussian error linear units (GELU) could be used instead. In a ReLU, for example, the output is 'zero' up until a certain input level, upon which, the output is simply a linear function. This linearity ensures a constant and non-zero gradient during GD BG training cycles. Since differentiation at a step is mathematically not defined, a Soft-ReLU is often used instead to ensure continued loss-function minimization and neural network performance improvement with training. Note, the known problem of overfitting neural networks does also apply to photonic neural networks. However, pruning techniques and hyperparameter adjustments during training are known and

interesting method (yet time consuming) to improve ANN performance. Interestingly, for the context of photonic neural networks, the network interconnectivity sparseness created in pruning steps, saves fan-out (fan-in) connections as well as waveguide connections. This is a blessing, since the bulkiness of photonic components (in contrast to electronic counterparts on a per-unit basis) increases functional density and thus performance per unit chip footprint. Indeed, future work should explore pruning techniques further for photonic neural networks. An initial study was, for instance, performed on the MZI N × N mesh network that was originally developed by the MIT groups [56], where it was shown that the number of required MZIs can be reduced by 3.7 × -7.6 × compared to the brute-force method [56,122]. Incidentally, in performance-optimized DNNs, the specific shape of the NLAF should be adjusted with layer depth; for instance, a ReLU that is not saturating is more useful in the upstream layers, while a saturating sigmoidal function supports a decision-making process at the fully-connected (FC) layer at the output in classification problems. In most neural networks demonstrations, the primary performance driver is power efficiency, while speed is a secondary consideration. Therefore, early photonic neural network demonstrations implement NLAF in the digital electronic domain, limiting the speed of neural network to the clock speed (hundreds of MHz to a few GHz). Nevertheless, optical NLAF, as discussed in Section 3 offer a unmatched speed over ten GHz in performing NLAF, thus becoming essential elements in order to solve many compelling applications that requires online (i.e. real-time) learning and inference or for neural networks with gigahertz bandwidths.

Memories: The memory is usually implemented electronically and various choices exist. Static RAM (SRAM) and dynamic RAM (DRAM) are used in neural network architectures to store inputs, weights, training parameters and look up table (LUT) values. At the cell level, SRAM typically uses 6 transistors (6 T) to store a single bit, whereas DRAMs use a single transistor and a single capacitor (1T1C). SRAMs are indispensable for neural network implementations given their faster access time; in addition, they can be leveraged for data reuse to reduce data fetch energy [123,124]. DRAM with its larger storage capacity is used to store all activations and training weights. However, owing to off-chip implementation, DRAM is slower with higher energy consumption than SRAM. Depending on the size of an SRAM and its location, the energy for read and write can scale from sub-pJ to 10 pJ per byte for SRAMs [31,125]. With an estimated SRAM size of 64KB used for DNN, the corresponding read and write latencies are ∼1 ns [126], which can be projected to reduce to 0.25 ns in 7-nm CMOS. Given the size difference and off-chip implementation in comparison to SRAMs, off-chip DRAMs consume more than two orders of magnitude of energy for data fetch [124]. High bandwidth memories (HBM) are used to reduce the

energy consumption of DRAM's data fetch [127–130], by moving DRAM modules closer to the chip. Alternatives to DRAM include resistive memories and PCMs. Utilizing 1T1R in ReRAM crossbar topology was demonstrated with low energy consumption and write latency [131,132]. On the other hand, using PCMs as unit cells requires high energy to write, and has poor resolution and linearity [133].

The PCMs operation is rooted in the switch between a crystalline (c) and an amorphous phase (a), typically via applied thermal stimulus (electrically or optically induced). Once switched into a particular state it remains there until a sufficiently high energy pulse (usually in the 600–1000 degree Celsius range) is delivered to the material. The switching energy is typically on the order of Nj, but, naturally, is a function of the thermal capacitor, hence scales with the device size. The memories' WRITE and RESET speed (e.g. from 'a' to 'c' and back, or vise-versa) is determined by the heat source's ability to deliver the energy in a short time, and the thermal capacitor's dissipation rate. Together they set up a memory switching time, tau_s. There seem to be some confusion in the community about this; for instance, work from the Oxford group uses a ns-pulse laser to program GST memories, but the reported response times of the device is in units of seconds [J. W. 134, 135, Y. Zhang et al. [139], 136]. Similar orders of magnitude were reported for VO_2 [137]. Numerical 3D thermo-optical analysis of 1–10 micrometer small PCM pads based on GST or its alloys forecast a temporal response on the order of microseconds [135]. As it stands, WRITE and RESET speeds that allow for a few dB of signal modulation (and not just 1–2%) on the order of microseconds or faster, have yet to be experimentally demonstrated. Once PCM are introduced to photonic-electronic neural networks, however, their usefulness can be rather high, since near zero-static power consumption's can be realized by SET kernels using the PCM approach, thus enabling a compute-in-memory (CIM) paradigm enabling high system efficiency (this is for applications where the kernel is updated infrequently).

An analog memory cell made up of a capacitor has drawbacks such as need for ADC/DACs to interact with digital cells, low density compared to SRAM cell, need for refresh, and sensitivity to noise and crosstalk. However, for photonic computing, placing a capacitor memory cell next to photonic elements is attractive. This is because the computing is analog in nature, and photonic compute elements are much bigger than digital compute elements, so the large size of the capacitor memory cell in comparison to SRAM is not a major concern. Such an architecture can eliminate the data movement bottleneck, significantly reducing the access time and energy consumption. Monolithic photonic processes supporting metal capacitors are ideally suited for such implementations.

For most applications, training the neural network is time consuming, spanning hours to days or even weeks. Thus, it is realistic to assume that for most applications, the neurons' weights or the kernel of a PTC is fixed and does not change often in time. For this reason, a non-volatile solution that retains information-of-state (i.e. memory), is of high interest. If achieved, a (near) zero static power consumption can be achieved in photonic neural networks, allowing them to be rather efficient. Note, 'static' refers here to the MAC operation and not to possible signal modulation, which would be considered part of the I/O of the system. Fortunately, such state-retention is recently achieved in electro-thermal programmable PCM.

Data I/Os: Photonic accelerators such as photonic neural networks and PTCs allow for high-throughputs approaching P-OPS (peta operations per second). Such a photonic 'highway', while promising, may not demonstrate its full potential, if the to-be-processed input data is not provided at a sufficiently high data rate to the optical accelerator. This can be assured in two ways; either the I/O data bandwidth is sufficiently high such as provided by a FPGA or the data is already prevailing in the optical domain (such as of an optical aperture from a camera system, for example). The latter is elegant, since it not only eliminates the needs to drive power-costly EO modulators for signal encoding, but more importantly eliminates the requirements for DACs/ADCs (see next paragraph). Indeed, high-speed DACs would consume about a third of the total photonic neural network system's power. For the case of optical data as the input, some PTCs show a dramatic power drop from about 80 W down to 2 W when DACs are not needed [138].

Domain crossings: digital/analog domain crossings: DACs and ADCs are required to interface a photonic neural network with digital signal processing (DSP) units (typically back-end) or when receiving data input data digitally such as from a server/computer etc. High sampling rate DACs used to drive input modulators mostly utilize current steering schemes and dissipate >5.5 pJ of energy for 6b-8b of conversion resolution [140,141]. The contribution of such converters to the overall energy efficiency of the neural network is reduced by 1/N as the network is scaled with N. Given the reusability of the weights over a given batch size, low speed capacitive DACs can be used. These DACs are usually adopted in textcolorredsynthetic aperture radar (SAR) architectures and typically contribute to most of the ADCs' energy. Charge average switching, merge and split, charge recycling, and common mode voltage (Vcm) based charge recovery are some of the techniques used to reduce the DAC capacitance and consequently its switching energy [142–144]. In the Vcm-based scheme, the differential DAC arrays are connected to a common mode voltage Vcm which reduces the DAC's switching energy. The power consumption of high speed ADCs scales exponentially with the resolution and linearly with the conversion rate (Murmann, n.d.). For

photonic neural network applications where the sampling frequency of the analog frontend is between 1–10 GS/s, and the resolution requirement is $<$ 8b, the energy consumption can be estimated as $\sim$1 pJ for state of the art ADCs (Murmann, n.d.). The architecture for ADC depends upon the photonic neural network. In recurrent neural network and long short-term memory (LSTM) networks, or in training back-propagation, the ADC is in a feedback network. Hence, a low-latency ADC architecture must be chosen. The lowest achievable latency is achieved in Flash ADCs, approximately $\sim$100 ps (with an estimated delay of 80 ps for dynamic comparators and 30 ps for the encoding gates). Therefore, latency-dependent photonic neural networks are limited to $\sim$10 GS/s of operation. However, high-speed Flash ADCs also consume large power. On the other hand, neural network architectures such as convolutional neural networks which do not rely on feedback relax the low latency constraints for the ADC. In such implementations, pipeline and time interleaved SAR ADCs are better suited given their low energy consumption and high conversion rate.

5. Training of photonic neural networks

Training is one of the key steps of most ANN algorithms. In common ANN, data are fed into the network and the weights are updated iteratively using error backpropagation algorithms, which requires significant computational resources. Photonic neural networks based on on-chip reconfigurable integrated photonic devices such as the MZI have been demonstrated as a promising energy efficient way to perform vector-matrix multiplications. For the deployment of photonic neural network to solve practical problems, these networks must be trained.

Current neural network optimization is largely based on back-propagating error gradients, and today's boom in neural network applications is closely linked to the successful implementation of this concept in digital hardware. However, in hardware with unidirectional, i.e. forward flow of information, its implementation requires calculating the error-gradient for each network connection according to the chain rule of differentiation. In digital hardware this creates an enormous overhead, while in analog networks each weight and neuron parameter needs to be probed and stored, which ultimately is prohibitively complex in most settings. One hardware friendly alternative can leverage local learning rules such as STDP or Hebbian learning, yet their standard implementations usually result in sub-optimal performance. Recently a new class of optimization rules identify strategies which are more hardware friendly. The different versions of feedback alignment relax the requirement on the back-propagation of an error signal [145]. Equilibrium propagation [146], an

energy-based model, entirely prevents error back-propagation yet achieves equivalent performance adjusts weights using local contrastive Hebbian learning. For that, the system's input remains clamped, and an error signal is applied to the ouput. In voltage-based systems, this provides a local correction signal and the concept was successfully transferred to hardware, yet a mapping onto the governing physical laws of optics has not yet been established.

5.1. Training analogue photonic neural networks

Photonic optimization and wave propagation In systems where information can symmetrically propagate forward as well as backward, such as often the case in photonics, a regulatory error signal could be sent backwards. However, in order to physically implement weight optimization according to error back propagation, a neuron's nonlinearity in backward direction needs to be the gradient of its nonlinearity in forward direction. Such asymmetric neurons are a challenge that remains largely out of reach until the current day, and the only viable photonic concepts rely on phase conjugation [147–149]. An attractive alternative is to rely on networks comprising photonic neurons whose activation function and its derivative are proportional. This allows to use pump-probe methods, where inference is created by the forward propagation pump, while local error signals are provided by the backward propagating low-intensity probe (Guo, Barrett, Wang, & Lvovsky, 2021).

Other approaches optimize a set of weights using in-silicio simulations on a classical computer, which are then transferred to the optical system. However, such offline training significantly suffers from discrepancies between the numerical model and the physical substrates, making substantial subsequent training on the physical substrate essential. However, combining such training can result in state of the art performance, as was demonstrated using a cascaded electro-optical free-space setup that emulates deep neural networks (Zhou et al., 2021) and achieved performance superior to the seminal LeNet-5 architecture while outperforming a state of the art GPU in terms of speed and energy efficiency. This is in sofar remarkable as LeNet-5 has been a reference around 10 years ago, hence trained photonic neural networks are closing the gap. An alternative to error back-propagation is to optimize weights using competitive statistical optimization tools such as genetic algorithms [139] or Bayesian optimization [150].

The optimization of integrated photonic circuits presents a challenge as it required probing local optical intensities at numerous sections across the entire circuit. One approach to mitigate the resulting complexity is leveraging the adjoint method, which substantially reduces the complexity due to

a simplification of the associated mathematical model (Hughes, Minkov, Shi, & Fan, 2018), as shown in Figure 3. In this work, a photonic analogue of the backpropagation algorithm is implemented based on the adjoint variable method (Hughes, Minkov, Shi, & Fan, 2018). By physically propagating the original field, the adjoint field and the interference field, the gradient of the loss function with respect to each phase shifter can be obtained simultaneously. The backpropagation procedure proceeds in a layer-by-layer fashion. In the l-th layer of the neural network, the *in situ* backpropagation training steps are: (i) Forward propagation (Figure 3a): Encode data as the original field amplitudes $\mathbf{X}_{l-1}$, send into the network and measure the field intensities $\left|\mathbf{e}_{og}\right|^2$ at each phase shifter. (ii) Error backpropagation (Figure 3b): Back propagate the error vector δ_l, measure the field intensities at each shifter $\left|\mathbf{e}_{aj}\right|^2$ and record the complex adjoint field as $\mathbf{X}_{TR}^*$. (iii) Interference measurement (Figure 3c): Send in the combination of the original and time-reversed adjoint fields $\mathbf{X}_{l-1} + \mathbf{X}_{TR}$ and measure the field intensities $\left|\mathbf{e}_{og} + \mathbf{e}_{aj}^*\right|^2$ at each phase shifter. (iv) Gradient calculation: Subtract the measured field intensities of (i) and (ii) from (iii) to get the gradient of loss function with respect to the phase change at each phase shifter. This procedure for gradient measurements is exact in lossless system and has been demonstrated numerically to be effective for systems with non-negligible, mode-dependent losses. The protocol provides an efficient way to compute the gradient of loss function with respect to arbitrary number of tunable parameters in constant time, which opens the possibility for in situ training of large-scale photonic circuits. This method for *in situ* measurements of device sensitivities can also be broadly applied to other reconfigurable systems and machine learning hardware platforms such as quantum optical circuits [151] and optical phased arrays [20].

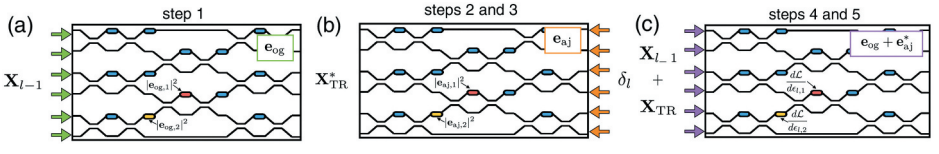

Figure 3. Schematic illustration of photonic neural networks training through in situ backpropagation. Colored squares represent phase shifters. (a) Forward propagation: send in $\mathbf{X}_{l-1}$ and record the intensity of the electric field $\left|\mathbf{e}_{og}\right|^2$ at each phase shifter. (b) Backpropagation: send in error signal δ_l and record the electric field intensity $\left|\mathbf{e}_{aj}\right|^2$ at each phase shifter and the complex conjugate output $\mathbf{X}_{TR}^*$. (c) Interference measurement: send in $\mathbf{X}_{l-1} + \mathbf{X}_{TR}$ to get gradient of the loss function $\mathcal{L}$ for all phase shifters simultaneously by extracting $k_0^2\mathcal{R}\{\mathbf{e}_{og}\mathbf{e}_{aj}\}$. (Tyler et al., 2018) Hughes et al. (2018).

Boolean learning via coordinate descent Coordinate descent is a practical as well as efficient alternative to error back propagation and currently is heavily explored in the field of machine learning. Individual or sets of weights i.e. coordinates, usually drawn at random, are modified in order to probe the error landscape's local gradient. Probing is followed by updating weights opposite to the gradient's direction and with a weighting factor dubbed the learning rate. In photonic hardware Boolean neural network connection weights have been realized via a digital micro-mirror device (DMDs) [121], and in a Boolean context the error gradient is probed by inverting a set of weights. Should the associated error gradient be negative, then the last modification is kept, otherwise it is discarded and the connections revert back to the previous configuration.

Boolean weights are currently explored in the general context of neural networks [152] as well as in special-purpose electronic hardware [153]. In photonic reservoir computing [121] it was shown that Boolean coordinate descent converges exponentially and achieves chaotic signal prediction accuracy only slightly below a comparable photonic reservoir [154] where double-precision weights were optimized offline. Furthermore, instead of a fully random, i.e. Makovian selection of descent coordinates, leveraging a greedy selection strategy allowed the system to converge twice as fast.

Boolean photonic weights implemented via a DMD enable programmable photonic neural networks comprising thousands of connections. In digitally emulated (Courbariaux et al., 2016), electronically [153] and photonically implemented [121] neural networks, such binarized weights resulted only in slight performance penalties. DMD-based Boolean coordinate descent in photonic neural networks therefore harbors great prospects for future, practical yet high performance neural networks, which noteworthy can be readily programmed based on classical software tools.

5.2. Ultrafast learning and spike timing dependent plasticity (STDP)

What makes neuron fascinating is its ability to learn and adapt, which is a powerful capability that governs our actions, thoughts, and memories. These important functions of humans are relying on the synaptic plasticity between neurons, which is self-adjusted based on the information being processed and the response of the neuron itself. Among different synaptic weight plasticity models, STDP is the most popular one, which is a biological process that adjusts the interconnection strength between neurons based on the temporal relationship (i.e. timing and sequence) between pre-synaptic and post-synaptic activities [155]. The more you are running a certain neural circuit, the stronger the circuit becomes.

In STDP, the interconnection strength between two neurons (N1 and N2) is determined by the relative timing and sequence between the presynaptic (red) and post-synaptic spikes (blue), as illustrated in Figure 4(a). If N2 spikes shortly after the stimulation from the presynaptic spike, the interconnection strength will be significantly increased and results in long-term potentiation (LTP) of the connection strength, as illustrated by the shaded purple region in Figure 4(b). However, if N2 spikes before the stimulation from the pre-synaptic spike, the interconnection strength will be significantly decreased, resulting in long-term depression (LTD) of the synaptic connection, as illustrated by the shaded brown region in Figure 4(b). The exact amount of synaptic connection strength increment/decrement depending on the precise timing difference between the pre-synaptic and post-synaptic spikes of N2.

To enable ultrafast learning in photonic neuron network, STDP has to be implemented using photonics and integrated into the photonic neuron network [156,157]. One promising solution to implement STDP in photonic neuron network is using semiconductor optical amplifier (SOA) [156–158]. SOA has a unique gain dynamic that is sensitive to the timing and sequence of the input stimulations, which is similar to the STDP in neurons. One example is to use both cross-gain modulation and cross-polarization modulation in SOA to mimic the LTP and LTD responses in biological neurons [158], the resultant photonic STDP is shown in Figure 4(c) that reassemble the biological STDP. It has been shown that supervised learning can be achieved using a SOA based STDP and two SOAs based neurons [156,157].

6. Applications of photonic neural networks

Integrated optical neural networks will be smaller (hundreds of neurons) than electronic implementations (tens of millions of neurons). But the bandwidth and interconnect density in optics is significantly superior to

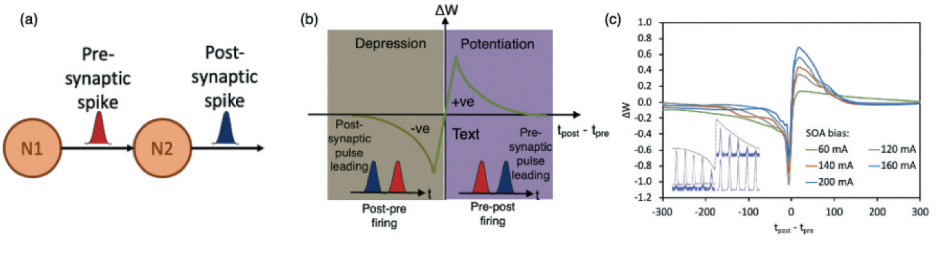

Figure 4. (a) Illustration of pre-synaptic and post-synaptic spikes from two neurons N1 and N2. (b) Illustration of a STDP response. Right purple region: long-term potentiation; Left brown region: long-term depression. tpost-tpre: time difference between the firing of the post- and pre-synaptic spikes. (c) Experimentally measured STDP curves from the photonic based STDP circuit.

that in electronics. This raises a question: what are the applications where sub-nanosecond latencies and energy efficiency trump the sheer size of processor? These may include applications 1) where the same task needs to be done over and over and needs to be done quickly; 2) where the signals to be processed are already in the analog domain (optical, wireless); and 3) where the same hardware can be used in a reconfigurable way. This section will discuss potential applications of photonic neural networks in computing, communication, and signal processing.

6.1. High-speed and low-latency signal processing for optical fiber communications and wireless communications

6.1.1. Optical fiber communications

The world is witnessing an explosion of internet traffic. The global internet traffic has reached 5.3 exabytes per day in 2020 and will continue doubling approximately every 18 months. Innovations in fiber communication technologies are required to sustain the long-term exponential growth of data traffic [159]. Increasing data rate and system crosstalk has imposed significant challenges on the DSP chips in terms of ADCs performances, circuit complexity, and power consumption. A key to advancing the deployment of DSP relies on the consistent improvement in CMOS technology [160]. However, the exponential hardware scaling of ASIC-based DSP chips, which is embodied in Moore's law as other digital electronic hardware, is fundamentally unsustainable. In parallel, many efforts are focused on developing new DSP algorithms to minimize computational complexity, but usually at the expense of reducing transmission link performances [161].

Instead of embracing such a complexity-performance trade-off, an alternative approach is to explore new hardware platforms that intrinsically offer high bandwidth, speed, and low power consumption [162, 27, 22]. Machine learning algorithms, especially neural networks, have been found effective in performing many functions in optical networks, including dispersion and nonlinearity impairments compensation, channel equalization, optical performance monitoring, traffic prediction, etc. [23].

PNNs are well suited for optical communications because the optical signals are processed directly in the optical domain. This innovation avoids prohibitive energy consumption overhead and speed reduction in ADCs, especially in data center applications. In parallel, many PNN approaches are inspired by optical communication systems, making PNNs naturally suitable for processing optical communication signals. For example, we proposed synaptic weights and neuron networking architecture based on the concept of WDM to enable fan-in and weighted addition [57]. This architecture can provide a seamless interface between PNNs and WDM systems, which can be applied as a front-end processor to address inter-wavelength or inter-

mode crosstalks problems that DSP usually lacks the bandwidth or computing power to process (e.g. fiber nonlinearity compensation in WDM systems). Moreover, PNNs combine high-quality waveguides and photonic devices that have been initially developed for telecommunications. Therefore, PNNs, by default, can support fiber optic communication rates and enable real-time processing. For example, The a scalable silicon PNN proposed by the authors is composed of microring resonator (MRR) banks for synaptic weighting and O/E/O neurons to produce standard machine learning activation functions. The MRR weight bank is inspired by WDM filters, and the O/E/O neurons use typical silicon photodetector and modulator. Therefore, the optimization of associated devices in PNNs can utilize the fruits of the entire silicon photonic ecosystem that is paramountly driven by telecommunications and data center applications.

In order to truly demonstrate photonics can excel over DSP, careful considerations are required to identify different application scenarios (i.e. long-haul, short-reach) and system requirements (i.e. performances, energy). Continuous research is needed to improve photonic hardware and to develop hardware-compatible algorithms. In the following session, we discuss two approaches to apply PNNs for optical communications. One approach leverages standard deep learning algorithms (e.g. backpropagation) to train every parameters in PNN. The PNN can be feed-forward or recurrent. We provide an example demonstration of using this approach for fiber nonlinearity compensation in long-haul fiber communication systems. Another approach is reservoir computing which is constructed with a network of nonlinear nodes with random, but fixed, recurrent connections, followed by a read-out layer. Only the read-out layer is trainable. Photonic reservoir computing system constructed with different photonic circuits and devices have been exploited as a channel equalizer to solve linear and nonlinear impairments in short-reach optical communication systems.

Deep learning for fiber nonlinearity compensation Long-haul communication systems prioritize high performances in terms of distance reach and spectral efficiency. This requirement allows the use of coherent technology, along with dense wavelength multiplexing and polarization multiplexing schemes, to maximize the fiber capacity. In long-haul fiber optic transmission systems, fiber nonlinearity remains a challenge to the achievable capacity and transmission distance. One reason is that the nonlinear interplay between signal, noises, and optical fibers negates the accuracy of conventional nonlinear compensation algorithms based on digital backpropagation. Another reason is, the implementation of most nonlinear compensation algorithms in DSP chips demands excessive resources. In contrast, the neural network approach can learn and approximate the nonlinear perturbation from the abundant training data, rather than solely

relying on the physical fiber model (known as stochastic nonlinear Schrodinger equation). Based on the perturbation methods, the derived neural network algorithm has enabled compensating the nonlinear distortion in a 10,800 km fiber transmission link with 32 Gbaud signals [S. 163].

118,developed a photonic neural network platform based on the so-called **'neuromorphic'** approach, aiming to map physical models of optoelectronic systems to abstract models of neural networks (which differs from the reservoir approaches discussed next). By doing so, the photonic neural network system can leverage existing machine learning algorithms (i.e. backpropagation) and map training results from simulations to heterogeneous photonic hardware. The concept is shown in Figure 5. A proof-of-concept experiment demonstrates the real-time implementation of a trained feed-forward neural network model using an integrated silicon photonic neural network chip for fiber nonlinear compensation [22]. In this work, the authors experimentally demonstrated that the silicon photonic neural network can produce a similar Q factor improvement compared to the simulated neural network for nonlinear compensation as shown in Figure 5, but it promises to process the communication data in real-time and with high bandwidth and low latency.

We also proposed a photonic architecture enabling all-to-all continuous-time recurrent neural networks (RNN) [118]. Recurrent neural networks can resemble optical fiber transmission systems: the linear neuron-to-neuron connections with internal feedback is analog to linear multiple-input multiple-output (MIMO) fiber channel with dispersive memory. With neuron nonlinearity, RNNs can be ideally used to approximate both of linear and nonlinear effects in a fiber transmission system and compensate for different transmission impairments. RNNs, consisting of

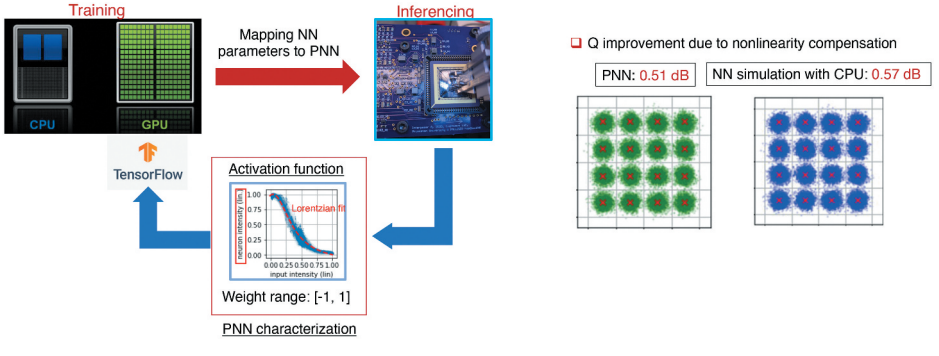

Figure 5. (left) Concept of training and implementing photonic neural networks. Inset shows the NLAF of the photonic neural network measured with real-time signals. The NLAF is realized by a fast O/E/O neuron using a SiGe photodetector connected to a microring modulator with a reverse-biased PN junction embedded in the microring. (right) Constellations of X-polarization of a 32 Gbaud PM-16QAM, with the ANN-NLC gain of 0.57 dB in Q-factor and with the PNN-NLC gain of 0.51 dB in Q-factor [22].

many feedback connections, are considered to be computationally expensive for digital hardware and require at least milliseconds to conduct a single inference. Contrarily, in photonic RNN, the feedback operations are simply done by busing the signals on photonic waveguides, allowing photonic hardware capable of converging to the solution within microseconds. This architecture thus allows to train PNNs externally using standard machine learning algorithms, e.g. backpropagation through time [164].

Reservoir computing for channel and/or predistortion equalization Short-reach fiber-optic communication systems (FOCS) have recently seen increasing demand driven largely by the proliferation of cloud-based computing architectures and fronthauling in cloud radio access networks (C-RAN). DSP-based coherent transceivers are optimized for reach and capacity and generally considered commercially unviable for short-reach optical fiber links due to their high cost, footprint, and latency. Legacy systems employing intensity modulation and direct detection (IM/DD) with limited signal processing capabilities can provide low-cost solutions for inter-datacenter applications, however they cannot scale with the ever-increasing capacity requirements in modern communication networks. This has motivated renewed research interest in low-cost and low-complexity transceivers with bitrates optimized over short transmission distances (10–100 km) where channel impairments are dominated by dispersion with some nonlinear distortion [165].

Photonic neural networks based on reservoir computing techniques have shown promising results for channel equalization in fiber-optic links. Reservoir computers (RC) are a class of recurrent neural networks that consist of a reservoir of sparsely connected neurons with randomized fixed weights. Contrary to feed-forward recurrent neural networks which are trained using backpropagation or Hessian-free optimization, reservoir computers only require the output weights to be trained, which can be achieved by linear regression. Optical RCs have attracted significant research interest as the reservoir can be realized by a single nonlinear element with a delayed feedback loop [166] which has size- and cost-efficient implementations in photonic circuits. The first demonstrations of this technology for signal equalization tasks in FOCS used a semiconductor laser as a nonlinear element with a fiber delayed-feedback line and showed results competitive with DSP-based techniques [162]. This approach is illustrated in Figure 6. Real-time operation of this photonic RC faces challenges, however, as the input layer is time-multiplexed by electronically masking each bit before injection into the reservoir, which incurs a speed penalty as the bit time must be stretched to match the delay of the feedback loop [167]. An all-optical implementation of a dual quadrature RC has also been proposed that could enable high bandwidth signal processing for coherent optical receivers [167].

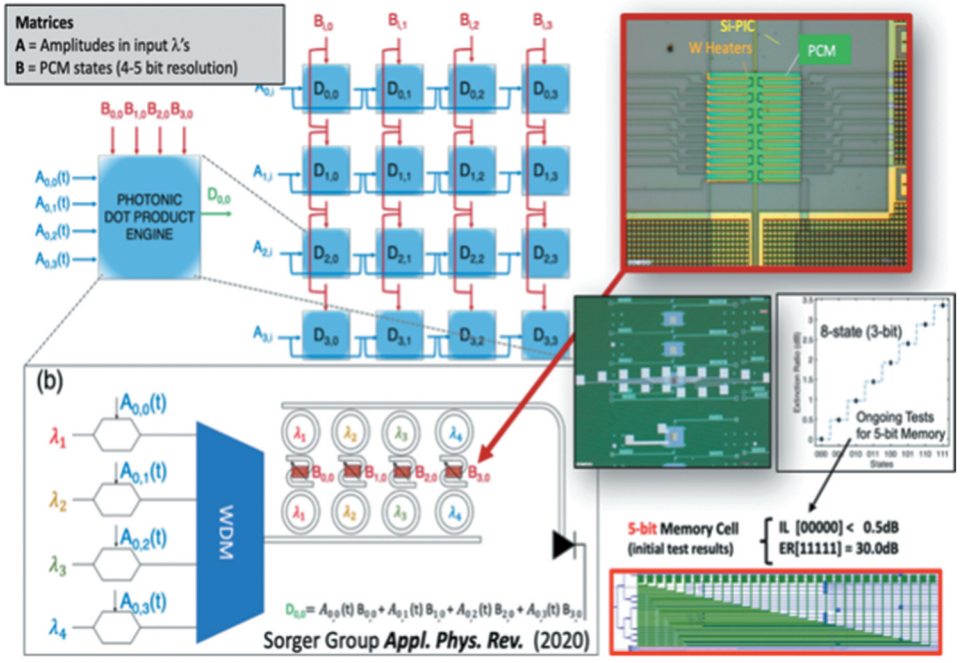

Figure 6. (a) Generic model of a reservoir computer [b) Illustration of the RC equalization scheme in 162,based on a single nonlinear node with time delayed feedback. The input $\mathbf{a}^i$ is a vector of N samples representing a single bit or symbol. The mask vector $\mathbf{m}$, which defines the input weights, is multiplied by the elements in $\mathbf{a}^i$ and injected sequentially into the reservoir. The output is a linear combination of the virtual nodes in the reservoir with weights optimized to produce an estimate $\mathbf{b}^i$ of the bit or symbol value.

In this design, nonlinear transformation of the input signal is achieved through the Kerr effect in a highly nonlinear fiber with a modulated pump to select the desired signal quadrature. Optoelectronic approaches have also been considered, with optical pre-processing via spectral slicing to improve the dynamics of a digital reservoir computer. This architecture addresses the system losses incurred in all-optical RCs and achieves significant reach extension compared to legacy IM/DD systems at the cost of higher complexity as the number of photodetectors scales linearly with the number of spectral slices [168]. Despite the SNR penalties, photonic RCs have merit over conventional DSP-based linear techniques when there is significant nonlinear distortion in the channel. Increasing the launch power at the transmitter may offset the system losses to achieve a higher optical SNR (OSNR) with improved performance in the presence of nonlinear impairments compared to linear equalizers. This hypothesis is supported by [169] which compared the bit error rate (BER) vs OSNR trade-off for a state-of-the-art DSP-based equalizer against a photonic RC in a 100 km 56 Gbd dense WDM transmission system. As expected, the DSP equalizer is the better choice at low OSNR values, however the results show the RC-based equalization outperforms the digital receiver at high OSNR where the nonlinear perturbations are strong.

6.1.2. Jamming avoidance response

The dramatic increased demand in mobile RF systems has significantly worsened the spectral scarcity issue, the overcrowded RF spectrum increases the likelihood of inadvertent jamming (170, 171]. Inadvertent jamming is one type of jamming that comes from a friendly source, that is usually aimless and unforeseeable. However, inadvertent jamming could easily corrupt the transmission channel if not being mitigated properly. In fact, inadvertent jamming does not only happen in our communication systems. Eigenmannia, a genus of glass knifefishes uses electric discharge to communicate with their own species and to recognize different species. Since Eigenmannia does not have FCC to regulate their frequency allocation, their frequency usage is dynamic. To avoid jamming, the Eigenmannia has an effective Jamming Avoidance Response (JAR) that helps the fish to identify potential jamming and automatically move their electric discharge frequency away from the potential jamming frequency [172,173]. The JAR in Eigenmannia mainly consists of four functional blocks (Figure 7(a), they are (i) Zero-crossing point detection unit, (ii) Phase unit, (iii) Amplitude unit, and (iv) Logic unit.

Text for Footnote 2

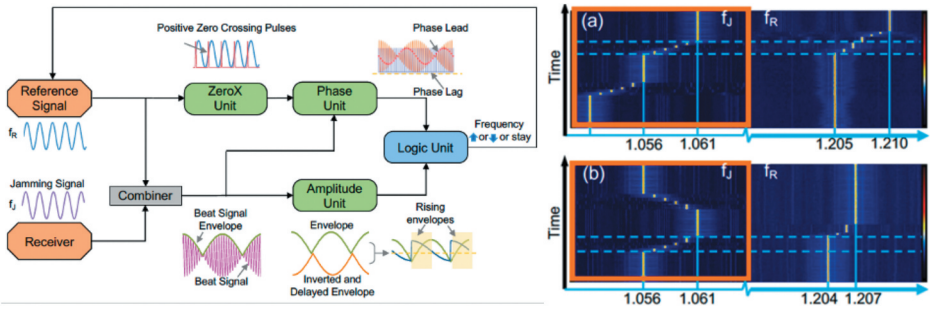

Figure 7. (a) Illustration of the JAR design and the four functional units. (b) Spectral waterfall measurement of the photonic JAR in action with sinusoidal reference signal fR and jamming signals fJ = 150 MHz. (i) fJ is approaching fR from the low frequency side and triggers the JAR, (ii) fJ is approaching fR from the low frequency side and triggers the JAR, and then is moved away.

First, the ZeroX unit identifies the positive zero crossing points in the reference signal. Then, phase comparison between the reference signal and the beat signal takes place at the Phase unit. Amplitude unit takes the envelope of the beat signal and marks the rising and falling amplitudes differently. Lastly, the Logic unit takes the phase and amplitude information obtained and determines if the electric discharge frequency should be increased or decreased to avoid jamming, and if there is no potential jamming threat.

This powerful JAR can be implemented using photonics to allow the JAR to be used in our communication frequency range [174,93,175]. The major device to achieve JAR is a semiconductor optical amplifier (SOA), where self-phase modulation is used in the Zero-crossing point detection unit, cross-gain modulation is used in both the Phase unit and amplitude units [174,93]. As the jamming frequency is moving closer to the jamming range, the JAR will be activated and move the emitting frequency away gradually until it is out of the jamming range, as illustrated in the spectral waterfall measurement in Figure 7(b). The photonic JAR works well for frequencies from hundreds of MHz to tens of GHz, that provides an adaptive and intelligent solution to inadvertent jamming in emerging communication systems.

6.1.3. Multivariate photonics – Principal component analysis (PCA), Independent component analysis (ICA), Blind source separation (BSS)

Multi-antenna systems provide an orthogonal dimension with which to share the electromagnetic spectrum. Signals received by phased-array antennas have a high degree of redundancy, so they are typically combined in such a way to distill the most salient information. Central to this is the concept of dimensionality reduction, in which information across many channels is fused in an

intelligent way to extract a smaller number of information-rich signals. Dimensionality reduction relies on learning the optimal combination of inputs based on statistical analysis. Blind source separation (BSS) is powerful tools that can reduce many received signals into a salient estimate of independent transmitters (Figure 8). BSS can essentially be achieved by sequentially using two well-known techniques: principal component analysis (PCA) and independent component analysis (ICA) on input signals. PCA distinguish variables by their variance (or second-order moment) of the output, whereas ICA also use its kurtosis (or fourth-order moment).

A typical radio receiver includes antenna(s) with a low-noise amplifier (LNA), an intermediate frequency (IF) mixer, and in some cases, filtering and demodulation. The received signal is then digitized (i.e. sampled and quantized) by an ADC and sent to a DSP backend. Key metrics that will be used to quantify performance in future radio spectrum access include frequency agility, power consumption, spatial utilization, reconfigurability and cost. Taking these metrics into consideration, it becomes clear that incremental improvements to the current state-of-the art will be unable to support the needs of future spectrum access. With conventional DSP, all of the signals must go through ADC quantization before processing can be applied (Figure 9). ADCs, however, are a bottleneck in RF systems [176], facing fundamental tradeoffs in sample rate, effective number of bits (ENOB), and power consumption (Sundstrom, Murmann, & Svensson, 2009; Walden, 1999). This creates a practical 'tunnel vision' effect, in which a large quantity of potentially useful information beyond the practical limit of ADC conversion (i.e. wider spectral window, distinguished directions of radiation, etc.) becomes unobservable.

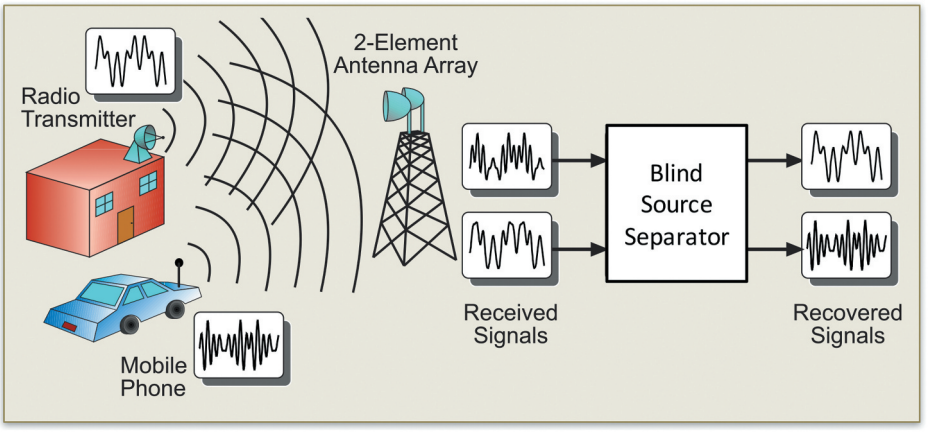

Figure 8. Concept of blind source separation.

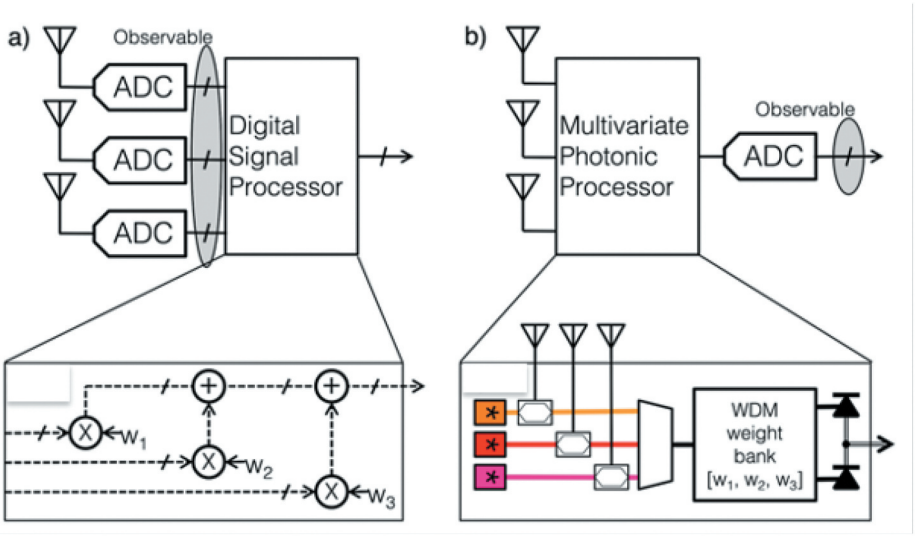

Figure 9. (Comparison of multi-antenna radio front-ends followed by dimensionality reduction. a) Dimensionality reduction with electronic DSP in which each antenna requires an ADC. b) Dimensionality reduction in the analog domain [a.k.a physical layer) in which only one ADC is required. A photonic implementation of weighted addition is pictured, consisting of electro-optic modulation, WDM filtering, and photodetection. From .177

Analog RF photonics provide a completely different set of physical processing constraints that lead to disruptive enhancements in the stated metrics [177]. In an RF photonic receiver, the input RF signal modulates the intensity of an optical carrier wave i.e. E/O conversion which is then detected to yield the output RF signal i.e. O/E conversion [178]. Between E/O modulation and O/E detection, the optical signal is processed with tunable optical components (Cap- many et al., 2013). By effectively upconverting to a 193THz IF, photonic processors are nearly frequency independent – even GHz signals are considered narrowband because optical waveguides have a flat frequency response over a 5THz window.

By using multiple RF signals from N antennas to modulate N distinct carrier wavelengths, these signals can be wavelength-division multiplexed (WDM) in an optical waveguide (Figure 9). Essentially, this one waveguide carries the complete picture of the wideband and spatially resolved spectral environment, and this proximity of information enables entirely novel approaches to spatial RF problems (Chang & Cheng, 2016). This analog combining has the profound impact on power use in that power is only related to N (one laser needed for each channel) and not bandwidth, resulting in the potential to revolutionize wideband, multi-antenna systems.

Figure 10 presents the experimental results of two-channel multi-band photonic BSS where we first run the photonic BSS pipeline to find the PC/IC vectors in free-running mode, and then apply them at the RF photonic front-

end to obtain the waveforms of estimated sources in synchronized mode (for evaluation] [24]. The band whose central frequency is at 880 MHz (Figure 10 (b)) exhibits relatively clean mixtures (blue curves in left column) and source separations (red curves in right column), which can be attributed to the fact that antennas hold the best transmission window around 900 MHz. The top mixture (RX1) is received by the antenna that is closer to the BPSK transmitter, while the bottom mixture (RX2) is received by the antenna that is closer to the ASK transmitter. The photonic BSS pipeline demonstrates its effectiveness by successfully separating these two sources, resulting in the top estimated source (IC1) being the BPSK (cannot tell zeros and ones from its intensity) and the bottom estimated source (IC2) being the ASK (zeros are at the ground level, while ones are represented by those envelopes). The other two bands whose central frequencies are at 781 MHz (Figure 10(a)) and 1000 MHz (Figure 10(c)) experience more difficulty of accurate source estimation caused by higher signal-to-noise in received mixtures, though they still manage to achieve perceivably effective source separation.

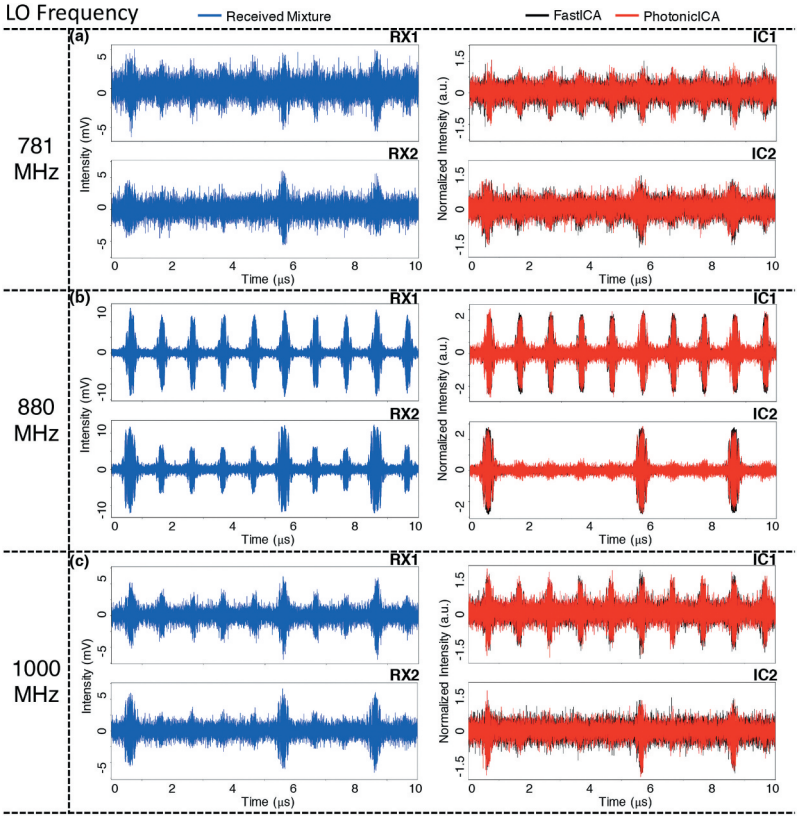

Figure 10. (Experimental waveforms of the received mixtures (left column) and corresponding estimated sources (right column) when performing FastICA (black) and photonic BSS (red) at three frequency bands whose central frequencies are at: (a) 781 MHz, (b) 880 MHz, and [c) 1000 MHz. From .24

6.2. AI/Machine learning

6.2.1. Vector-matrix multipliers

With an ongoing trend in computing hardware towards increased heterogeneity, domain-specific coprocessors are emerging as alternatives to centralized paradigms. The tensor core unit has shown to outperform graphic process units by almost 3-orders of magnitude enabled by higher signal throughout and energy efficiency. In this context, photons bear several synergistic physical properties while PCMs allow for local nonvolatile mnemonic functionality in these emerging distributed non von-Neumann architectures. After earlier theoretical tensor core proposals 138, recently photonic-based tensor operation ASICs were demonstrated Feldmann et al. (2021], X. Xu et al. (2021) An integrated photonics-based tensor core unit can be designed by strategically utilizing i) a photonic parallelism via wavelength division multiplexing, ii) high 2 Peta-operations-per second throughputs enabled by 10's of picosecond-short delays from optoelectronics and compact photonic integrated circuitry, and iii) near-zero power-consuming novel photonic multi-state memories based on PCMs featuring vanishing losses in the amorphous state. Combining these physical synergies of material, function, and system, we show, supported by numerical simulations, that the performance of this 4-bit photonic tensor core unit can be one order of magnitude higher for electrical data, whilst the full potential of this photonic tensor processor is delivered for optical data being processed, where we find a 2–3 orders higher performance (operations per joule) as compared to an electrical tensor core unit whilst featuring similar chip areas. This work shows that photonic specialized processors have the potential to augment electronic systems and may perform exceptionally well in network-edge devices in the looming 5 G networks and beyond. (Figure 11)

6.2.2. Convolutions (inference accelerator)

Convolutions of an image with a filter, i.e. the convolution kernel, are among the most heavily employed operations in computational imaging. During convolution, a kernel slides across an image and at each position the overlap integral between image and kernel provides the convolution's value. This technique is extremely efficient in identifying entire objects [179] or only local features. It therefore comes at not too much of a surprise that functional topologies implementing convolutions similar to Gabor filters have been identified in the visual cortex of mammals [180]. Convolutional neural networks (CNNs) leverage these concepts in a modern information processing context, and an image classification performance superior to humans makes CNNs relevant for a wide range of applications [2].

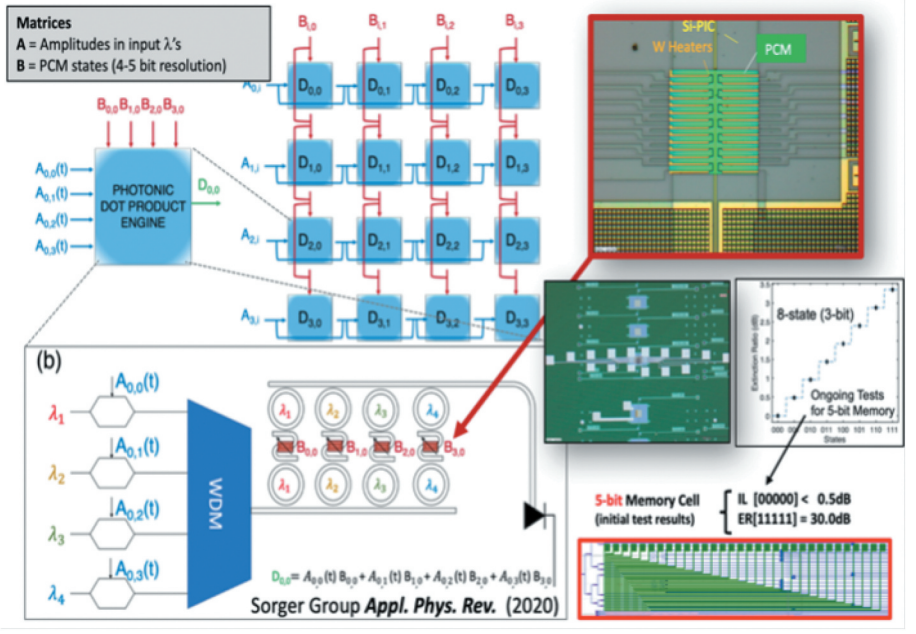

Figure 11. Photonic Tensor Core [PTC] featuring Peta-OPS throughput and 10s ps-short latency to process VMM in PICs 138. (left] schematic of a modular exemplary PTC composed of 4 × 4 photonic dot-product engines (i.e. 1 × 4 vector-multiplications) using ring resonators enabling WDM-based parallelism. Data entered via (exemplary) EO modulators at 50 Gbps are MUXed then dropped at passive ring resonators. Being 'passive' for the ring resonator is importantly beneficial since it allows for reduced complexity, less real-estate used on the chip, and does not consume power unlike thermally-tuned approaches to perform the optical multiplication. (right) The latter can be elegantly executed using nonvolatile photonic random access memories (P-RAM) based on electrically-programmable phase change materials (PCM). The image shows a prototype of a 3-bit (8-states) binary-written P-RAM element on a Silicon PIC. Using non-GST PCMs enables a rather low (<0.02 dB/state) insertion loss [135].

Free-space optical convolution: A fundamental aspect of computational imaging is that the primary information is distributed in two dimensions (2D). Free-space optical convolution setups inherently respect this encoding principle. Task-specific accelerators based on free-space optics bear fundamental homomorphism for massively parallel and real-time information processing given the wave-nature of light. However, initial results are frustrated by data handling challenges and slow optical programmability. Sorger's group recently introduced an amplitude-only Fourier-optical processor paradigm capable of processing large-scale ~(1000 × 1000) matrices in a single time-step and 100 microsecond short latency [117] (Figure 12). Conceptually, the information-flow direction is orthogonal to the two dimensional programmable-network, which leverages 106 -parallel channels of display technology, and enables a prototype demonstration performing convolutions as pixel-wise multiplications in the Fourier domain reaching peta operations per second throughputs. The required

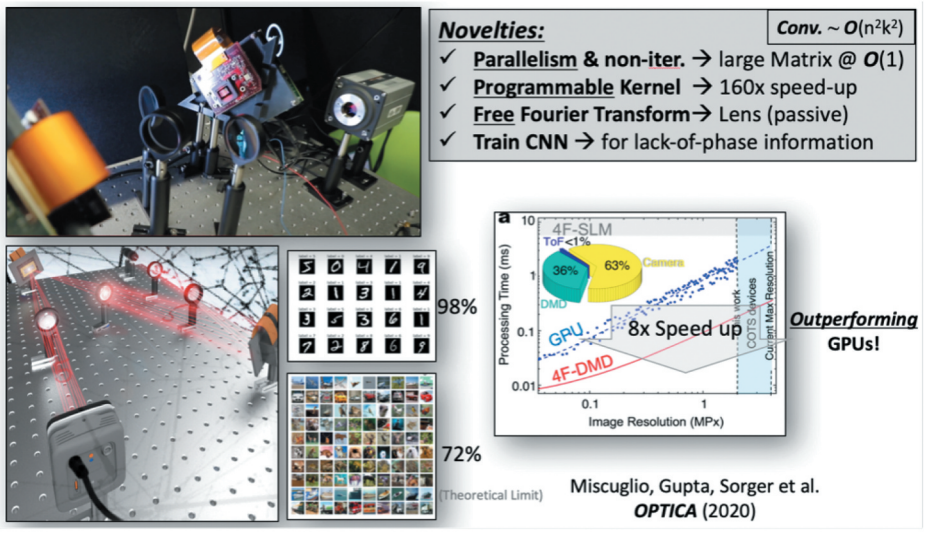

Figure 12. Example of an Optical Convolutional Neural Network (CNN) accelerator exploiting the massive (10^6 parallel channel) parallelism of free-space optics. The convolutional filtering is executed as point-wise dot-product multiplication in the Fourier domain. The conversion into and out-of the Fourier domain is performed elegantly and completely passively at zero power While SLM-based systems can perform such Fourier filtering in the frequency domain, the slow update rates does not allow them to outperform GPUs. However, replacing SLMs with fast 10s kHz programmable digital micromirror display [DMD) units, gives such optician CNNs an edge over the top-performing GPUs. Interestingly, the lack-of-phase information of these amplitude-only DMD-based optical CNNs can be accounted for during the NN training process. For details refer to .117

real-to-Fourier domain transformations are performed passively by optical lenses at zero-static power. Previously, realized a convolutional neural network (CNN) performing classification tasks on 2-Megapixel large matrices at 10 kHz rates, which latency outperforms current GPU and phase-based display technology by one and two orders of magnitude, respectively. Training such an optical convolutional layer on image classification tasks and utilizing it in a hybrid optical-electronic CNN, showed classification accuracy of 98% (MNIST) and 54% (CIFAR-10), respectively. Interestingly, we found that the amplitude-only CNN is inherently robust against coherence noise in contrast to phase-based paradigms and features an over 2 orders of magnitude lower delay than liquid crystal-based systems. Beyond contributing to novel accelerator technology, scientifically such an amplitude-only massively-parallel optical compute-paradigm shows that lack of phase information can be accounted for in optical image processing by training the system.

2D integrated optical convolution: Deep neural networks are based on CNNs which are powerful and highly ubiquitous tools for extracting features from large datasets for applications such as computer vision and

natural language processing. The success of CNNs for large-scale image recognition has stimulated research in developing faster and more accurate algorithms for their use. However, CNNs are computationally intensive and therefore results in long processing latency. One of the primary bottlenecks is computing the matrix multiplication required for forward propagation. In fact, over 80% of the total processing time is spent on the convolution (181]. Therefore, techniques that improve the efficiency of even forward-only propagation are in high demand and researched extensively [Good-fellow, Bengio, & Courville, 2016; 182]. Recently, there has been much investigation of implementing convolution operations with integrated optics [58, Feldmann et al., 2021; 138, X. Xu et al., 2021]. These approaches can speed up convolution operations by orders of magnitude (over current electronic processors such as graphic processing units (GPU) and tensor processing units (TPU) by implementing fast and parallel vector-matrix multiplications with wavelength multiplexing techniques.

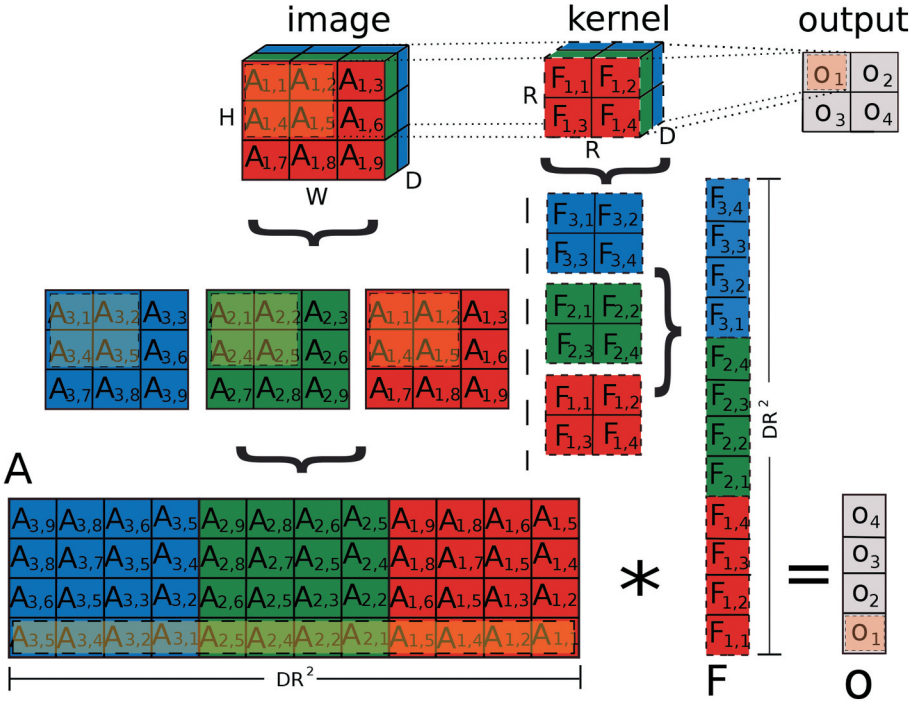

Figure 13. Schematic illustration of a convolution. An input image $\mathbb{A}$ with dimensionality $H \times W \times D$ (where H, W and D are the height, width and depth of the image, respectively) is convolved with a kernel $\mathbb{F}$ with dimensionality $R \times R \times D$, to produce an image $\mathbb{O}$. Assuming $H = W$, the overall output dimensionality is $(H - R + 1)^2$. The bottom of the figure shows how a convolution operation generalized into a single matrix-matrix multiplication. where the kernel $\mathbb{F}$ is transformed into a vector with DR^2 elements, and the image is transformed into a matrix of dimensionality $DR^2 \times (H - R + 1)^2$. Therefore, the output is represented by a vector with $(H - R + 1)$ elements.

A convolution is a weighted summation of two discrete domain functions f and g: $(f * g) = \sum_{t=-\infty}^{\infty} f[\tau]g[t-\tau]$ where $(f * g)$ represents a weighted average of the function $f[\tau]$ when it is weighted by $g[-\tau]$ shifted by t. The weighting function $g[-\tau]$ emphasizes different parts of the input function $f[\tau]$ as t changes. Convolutions are well known to perform a highly efficient and parallel matrix multiplication using kernels [183]. As depicted in Figure 13 convolution of an image $\mathbb{A}$ with a kernel $\mathbb{F}$ that produces a convolved image $\mathbb{O}$. An image is represented as a matrix of numbers with dimensionality $H \times W \times D$, where H and W are the height and width of the image, respectively; and D refers to the number of channels within the input image. Each element of a matrix $\mathbb{A}$ represents the intensity of a pixel at that particular spatial location. A kernel is a matrix $\mathbb{F}$ of real numbers with dimensionality $R \times R \times D$. The value of a particular convolved pixel is defined by: $O_{i,j} = \sum_{h=1}^{D} \sum_{q=0}^{iS+R} \sum_{p=0}^{jS+R} B_{q,p,h} A_{iS+q,jS+p,h}$, where S is the 'stride' of the convolution. The dimensionality of the output feature is: $\left\lceil \frac{H-R}{S} + 1 \right\rceil \times \left\lceil \frac{W-R}{S} + 1 \right\rceil \times K$, where K is the number of different kernels of dimensionality $R \times R \times D$ applied to an image, and $\lceil . \rceil$ stride the ceiling function. The efficiency of convolutions for image processing is based on the fact that they lower the dimensions of the outputted convolved features. Since kernels are typically smaller than the input images, the feature extraction operation allows efficient edge detection, therefore reducing the amount of memory required to store those features.

In 2014, Tait et al. [57] proposed a scalable silicon photonic neural network called 'broadcast-and-weight' (B&W) which was demonstrated in 2017 [118] concurrently with other silicon photonic neuromorphic architectures [Shain- line et al., 2017; 56]. Since the B&W architecture is based on wavelength mutiplexing, parallel matrix multiplication can be leveraged to perform operations such as convolutions. In B&W architecture, WDM signals are weighted in parallel by a bank of MRRs used as tunable filters 66, Tait, Jayatilleka, et al. (2018), summed with balanced photodiodes (BPD), and nonlinear activation functions implemented with MRR modulators. In 2020, based on this architecture, Bangari et al. [58] introduced a digital electronic and analog photonic (DEAP) architecture capable of performing highly efficient CNNs.

Figure 14 shows the silicon photonic implementation of DEAP for performing convolution operations. To handle convolutions for kernels with dimensionality up to $R \times R \times D$, R^2 lasers are required with unique wavelengths since a particular convolved pixel can be represented as the dot product of two $1 \times R^2$ vectors. To represent the values of each pixel, DR^2 add-drop modulators (one per kernel value) are required where each

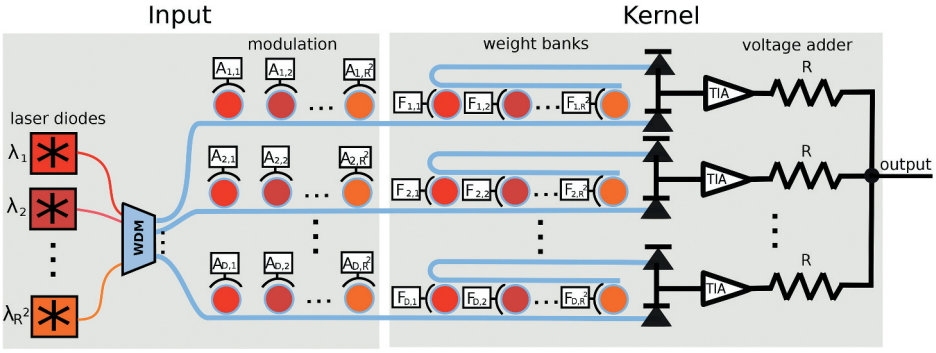

Figure 14. Photonic architecture for producing a single convolved pixel. Input images are encoded in intensities $A_{l,h}$, where the pixel inputs $A_{m,n,k}$ with $m \in [i, i+R_m], n \in [j, j+R_m], k \in [1, D_m]$ are represented as $A_{l,h}$, $l = 1, \ldots, D$ and $h = 1, \ldots, R^2$. Considering the boundary parameters, we set $D = D_m$ and $R = R_m$. Likewise, the filter values $F_{m,n,k}$ are represented as are represented as $F_{l,h}$ under the same conditions. We use an array of R^2 lasers with different wavelengths λ_h to feed the MRRs. The input and kernel values, WCIM_A_1985028 and $F_{l,h}$ modulate the MRRs via electrical currents proportional to those values. Then, the photonic weight banks will perform the dot products on these modulated signals in parallel. Finally, the voltage adder with resistance **R** adds all signals from the weight banks, resulting in the convolved feature.

modulator keeps the intensity of the corresponding carrier wave proportional to the normalized input pixel value. The R^2 lasers are multiplexed together using WDM, which is then split into D separate lines. On every line, there are R^2 add-drop MRRs (where only input and through ports are being used), resulting in DR^2 MRRs in total. Each WDM line will modulate the signals corresponding to a subset of R^2 pixels on channel k, meaning that the modulated wavelengths on a particular line corresponds to the pixel inputs $(A_{m,n,k})$; $m \in [i, i+R], n \in [j, j+R]$ where $k \in [1, D]$, and $S = 1$. The D WDM lines are then be fed into an array of D photonic weight banks (WB). Each WB contains R^2 MRRs with the weights corresponding to the kernel values at a particular channel. Each MRR within a WB is tuned to a unique wavelength within the multiplexed signal. The outputs of the weight bank array are electrical signals, each proportional to the dot product $(F_{m,n,k}) \cdot (A_{p,q,k})$; $m \in [1, R^2]$, $n \in [1, R^2]$, $p \in [i, i+R^2]$, $q \in [j, j+R^2]$, where $k \in [1, D]$, and $S = 1$. Finally, the signals from the weight banks are added together. This can be achieved using a passive voltage adder. The output from this adder will therefore be the value of a single convolved pixel.

Recently, Feldman et al. (Feldmann et al., 2021) experimentally demonstrated this approach with a photonic tensor core for parallel convolutional processing achieving bandwidths of 2 TMACs/s and compute densities of 555 GMACs/s/mm^{2}with 5-bits of precision. Their system consists of a 3×3 filter, 4 channels and filters and 14 GHz modulation and detection bandwidth, with

a MAC cell area of 285 μm $\times$ 354 μm. With improved devices, efficiency, bandwidth, and integration densities, such a tensor core could feature a computational bandwidth of 1 PMACs/s and compute density of 15.6 TMACs/s/mm^2 with 50 GHz modulation and detection bandwidths, tensor core size of 50 $\times$ 50, and 187 wavelength channels, and MAC cell area of 30 μm $\times$ 30 μm. For comparison, the Google TPU [184] has a compute density of 150 GMACs/s/mm^2 with 8-bits of precision.

3D integrated optical convolution: Information is kept within 2D spatial encoding, while light propagation through passive components along a third dimension implements the required operation, i.e. Fourier and inverse Fourier transforms as well as multiplication with a filter kernel. Novel 3D nano-fabrication [185] can now create intricate photonic waveguide circuits [186] and holograms [187] that equally leverage the primary encoding space of images. Figure 15(a) shows how Boolean convolutional topologies can be 'hard-wired' in 3D. This intricate 3D routing can then be realized using 3D photonic waveguides, fabricated using two-photon polymerization of femto-second laser pulses [185], and Figure 15(b) shows a SEM micrograph, and the resulting convolution filter's transfer function, Figure 15(c), agrees well with the target. Importantly, 3D integration makes such interconnects scalable in size [188], and motivated by identical scalability arguments similar 3D integration is already explored in electronics [92]. However, such electronics chips will face severe thermal management challenges due to capacitive energy deposition into a volumetric circuit, and photonics therefore has an inherent advantage towards scalable integration of parallel convolutional filters.

6.2.3. *Wave physics as an analog recurrent neural network*

RNNs are a powerful class of artificial neural networks that have demonstrated outstanding performance for applications such as machine translation, speech recognition and time series prediction. The latency and energy consumption are important factors for these tasks, which motives the design

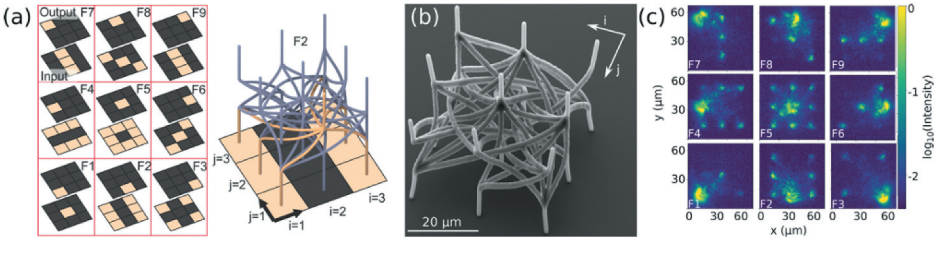

Figure 15. Fully parallel and passive convolutions integrated in 3D. (a) 9 convolutional Haar filters and their 3D integration topology using photonic waveguides. (b) SEM micrograph of the 3D printed convolutional filter unit, fabricated with direct laser writing. (c) The filter's optical transfer function agrees well with the target defined in [a). Figure reproduced from .186

of specialized hardware platforms such as neuromorphic photonic networks 118. Recently, 189,identifies the fundamental mapping between the computation of an RNN and the dynamics of wave-based physical systems. Passive photonic devices can thus be trained to operate as recurrent neural network without using analog-to-digital conversion or recurrently feeding signals back to the input.

For a standard RNN implementation, input signals are fed into the network sequentially. At each time step t, the network performs operation on the hidden state h_{t-1} and the input signal x_{t-1} with dense matrix multiplications $W^{(h)}$, $W^{(x)}$, $W^{(y)}$ and nonlinear activations $\sigma^{(h)}$, $\sigma^{(y)}$ to get the output signal y_t and to update the hidden state to h_t, as shown in Figure 16 (a). On the other hand, wave-based physical system is usually described by a second-order partial differential equation, which with a finite difference discretization can be represented as a discretized recurrent equation 189. The discretized equation has a form that is similar to the update equation of the RNN if we rewrite the hidden state h_{t-1} as the field distributions of the current and next time steps $[u_{t-1}, u_t]^T$, as shown in Figure 16(b). The Laplacian operator and the material nonlinear response in the wave equation encode the matrix multiplications and nonlinear activations of the RNN. The spatial distribution of the material's property such as refractive index corresponds to the trainable weights $W^{(h)}$, $W^{(x)}$, $W^{(y)}$ in RNN and can be optimized for different tasks. The time-dependent input source of

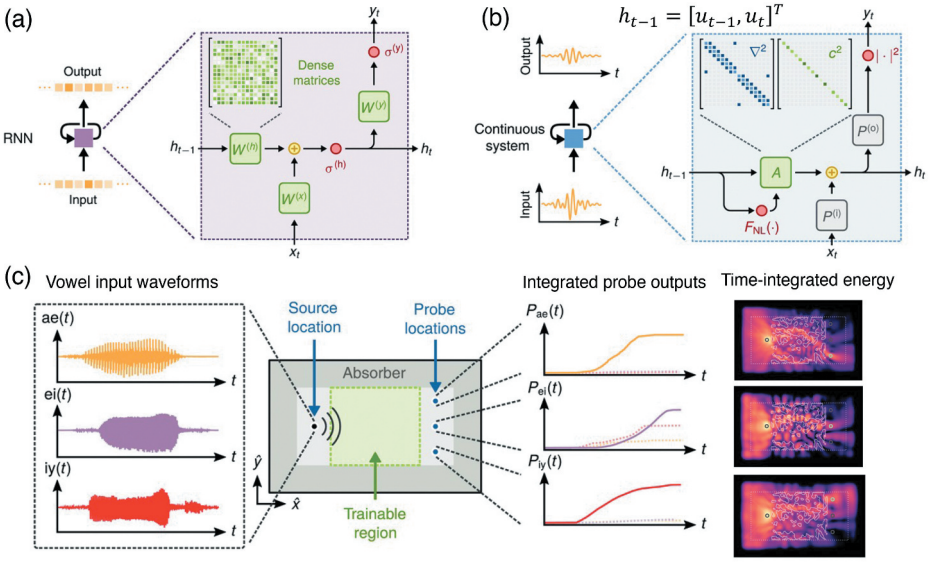

Figure 16. (a] Illustration of a recurrent neural network at time step t. (b) Illustration of the recurrent representation of the physical wave system. [c) A vowel classification task on an optical analog recurrent neural network. $x(t)$ is the waveform input signal at the source location. $P_x(t)$ is the output signal at different probe locations. Figure reproduced from .189

physical wave can be understood as the input data x_t. And the intensity measurement can be mapped as the output y_t. The comparisons indicate the equivalence between the dynamics of physical waves and the dynamics of RNN. Figure 16(c) shows how a wave RNN can be trained for vowel classification tasks. The audio signals are injected from the left at the source location, and output signals are detected at different probe locations on the right side. During the training process, the spatial distribution of material refractive index is updated by feeding audio signals iteratively using gradient-based inverse design method 190. After training, the material refractive index distribution is fixed. For different vowels, the final device routes the energy to the correct locations depending on the input vowel. The system achieves a classification accuracy of $95.5\% \pm 1.4\%$ and $90.3\% \pm 6.4\%$ on the training and testing datasets, respectively. This work connects machine learning algorithms and analog computing hardware platforms, and opens the opportunity to implement time-domain computations on wave-based physical systems such as optics, and acoustics. In addition to the advantages in speed and energy consumption, passive wave-based physical computing platforms show a promising pathway for encoding time-domain information because the recurrence relationship is encoding naturally in their wave propagation governing equations.

6.2.4. Photonic accelerators for edge or fog computing: image processing and super resolution filtering

Photonic neural networks [e.g. 138] and free-space Fourier-optics 4 f processors [117] can be particularly beneficial to tasks that involve, for instance, super-resolution on object detection performance in satellite imagery. Some modern camera sensors, present in everyday electronic devices like digital cameras, phones, and tablets, are able to produce reasonably high-resolution (HR) images and videos. The resolution in the images and videos produced by these devices is in many cases acceptable for general use. However, there are situations where the image or video is considered low resolution (LR). Examples include the following situations: (i) Device resolution limitation (as in some surveillance systems and satellite images). (ii) Object relatively small in a larger context; e.g. faces or vehicle located far away from the camera. (iii) Blurred or noisy images; being the device mounted on moving automated device (e.g. drones). (iv) Improving the resolution as a preprocessing step improves the performance of other algorithms that use the images; pattern recognition and target tracking. Super-resolution is a technique to obtain a high resolution (HR) image from one or several LR images. In this case one can use a training set to train (offline) a convolutional neural network (CNN) to learn to map between a low-resolution image and the high-resolution image within the training set; for instance, using a 4 f-based system one can obtain pixel-wise dot product in

the Fourier domain between the large input matrix (2MP) with a reprogrammable kernel that has been pre-patterned/written according to the training. The re-programmability, in a first instance, can be achieved using micromirror devices, which by shining light on top of the film, can locally change the phase of the PCM generating a 2D pattern. Although, this solution is limited by the speed at which the film can be written, therefore, in future implementations one can independently and simultaneously tune each pixel by changing their phases electrothermally in μs-time scale [requires additional circuitry), thus acting as 2D space filter in the Fourier domain or as layer of a convolutional network. Providing a forward looking view, if one would substitute the liquid crystal from the SLMs and the micromirrors from the DMDs with GHz-fast updating elements, but keep the same 1000 × 1000 pixel real-estate, the system would yet improve by a factor of 10^{10} to 10^{8} from SLM's and DMD's, respectively. Such electro-optic modulation, however, must be ultra-compact and regular photonics modulators based on carrier injection or depletion using Silicon or using the Pockel's effect in Lithium Niobate C. 73,are non-usable due to millimeter to centimeter-large modulator device footprints. An alternative approach is to utilize higher index-changing emerging EO materials and heterogeneously integrate them with photonic waveguides. ITO and its ability to produce epsilon-near-zero 91, a nonlinearity enhancements [including EO nonlinearity] can be used to demonstrate micrometer-compact modulators, for example. 77, 191–193.

6.3. Nonlinear programming

6.3.1. Solving optimization problems (model predictive control]

Solving mathematical optimization problems lies at the heart of various applications present in modern technology such as machine learning, resource optimization in wireless networks, and drug discovery. Many optimization problems can be written as a quadratic program. For example, the least squares regression method can be mathematically mapped to a relatively easy quadratic program (with a positive definite quadratic matrix). Quadratic Programming refers to algorithms related to solving the optimization problem of finding the extremes of a quadratic objective function subject to linear constraints. The general formulation of a quadratic program is usually solved iteratively, often requiring many time steps to reach the desired solution. The difficulty of quadratic programming grows exponentially with the dimension of the problem. Algorithms that can deal with large dimensions involve more computationally intensive techniques such as genetic algorithms or particle swarm optimization. As a result, conventional digital computers must either be

limited to solving quadratic programs of very few variables, or to applications where the computation time is non-critical. This is why traditional computers are not appropriate to implement algorithms depending on QP for high-speed applications such as signal processing and control systems. In machine learning, many algorithms, such as support vector machines, require offline training because of the computational complexity of QP, but would be much more effective were they trained online.

A number of high-speed control problems, e.g. controlling plasma in aircraft actuations, fusion power plants, guiding of drones etc., are currently bottlenecked by the speed and latency of the control algorithms. Model predictive control (MPC) is an advanced technique to control complex systems, outperforming traditional PID control methods because it is able to predict control violations, rather than react to it. However, its control loop involves solving a quadratic problem at every control step, and therefore it is not computationally tractable for systems requiring speeds higher than kHz. Photonic neural networks help overcome this tradeoff by using techniques such as wavelength division multiplexing, which enables hundreds of high bandwidth signals (20 GHz) to be guided through a single optical waveguide.

Neural networks were demonstrated to solve general-purpose quadratic optimization problems by Hopfield and Tank [194,195]. Until now, Hopfield networks have not been commonly implemented in hardware due to its all-to-all connectivity, which creates an undesirable tradeoff between neuron speed and neural network size – in an electronic circuit, as the number of connections increases, the bandwidth at which the system can operate decreases 57. This means a photonic Hopfield network implementation can simultaneously tackle quadratic programs with large dimensions and converge in nanoseconds [27]. Figure 17 illustrates the implementation of the MPC algorithm on a neuromorphic photonic processor. Implemented in photonic hardware, model predictive control can be employed in systems operating in the MHz regime.

6.3.2. Ordinary differential equation solving with a neural compiler

While neural networks are often used for their learning properties, they can also be programmed directly. Direct programming of analog systems has a major pitfall in that the components are unreliable and subject to parameter variation. One approach is to represent variables as population states that are robust to parameter variations. The neural engineering framework (NEF) [196] provides an algorithm to program ensembles of imperfect analog devices to perform operations on population coded variables.

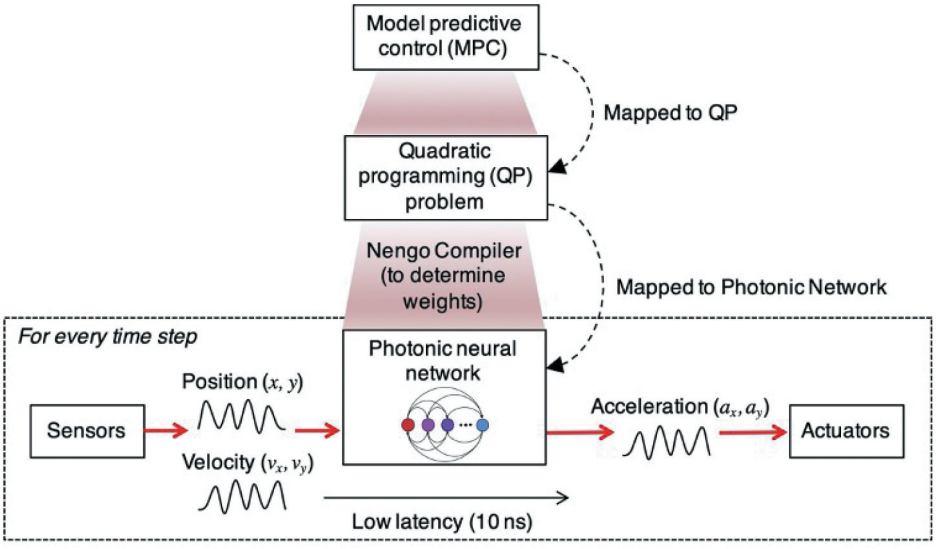

Figure 17. Schematic figure of the procedure to implement the MPC algorithm on a neuromorphic photonic processor. Firstly, map the MPC problem to QP. Then, construct a QP solver with continuous-time recurrent neural networks (CT-RNN) [194]. Finally, build a neuromorphic photonic processor to implement the CT-RNN. The details of how to map MPC to QP, and how to construct a QP solver with CT-RNN are given in De Lima et al. (2019). Adapted from De Lima et al. [27].

118,demonstrated a programmable network of two photonic neurons. The NEF algorithm was fed the responses of photonic devices, resulting in a weight matrix that allows the network to approximate variables, operations, and differential equations. It was shown in simulation how 24 neurons could emulate the Lorenz attractor. The approximation improves with more neurons.

The Lorenz attractor is an example of a task-based benchmark, which allows a performance comparison between disparate computing technologies. Figure 18 shows a comparison between a conventional CPU solver (a,c) and a 24-neuron photonic CTRNN (b,d). The reproduction of Lorenz attractor dynamics demonstrated the compatibility between photonic neural networks and the NEF, including all of the NEF's key principles and consequent functionality. Using this neural compiler provides a route to a variety of known applications and other [197]. The axes of Figure 18 (c,d) are scaled to have equal simulated oscillation periods, but the real-time taken to do the simulation is much faster for the photonic NEF solver.

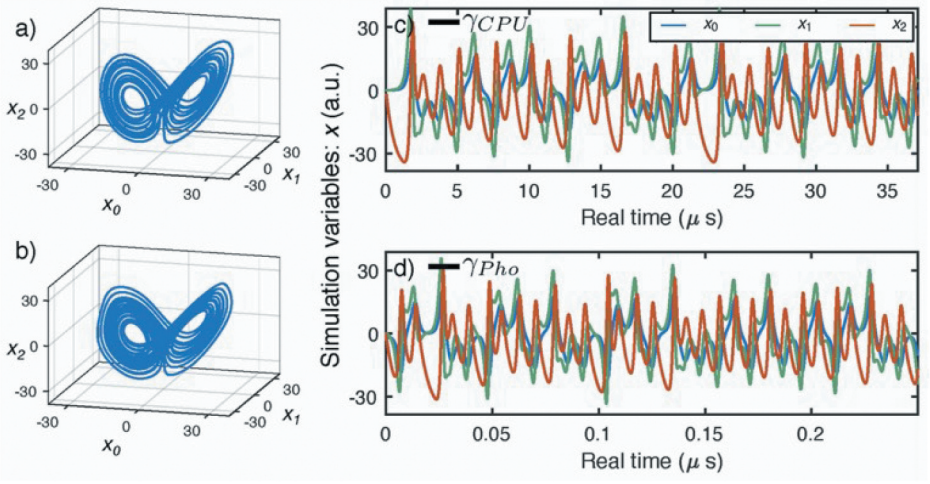

Figure 18. Photonic neural network benchmarking against a CPU. (a,b) Phase diagrams of the Lorenz attractor simulated by a conventional CPU (a) and a photonic neural network (b). (c,d) Time traces of simulation variables (blue, green, red) for a conventional CPU (c) and a photonic CTRNN [d). The horizontal axes are labeled in physical real time, and cover equal intervals of virtual simulation time, as benchmarked by γCPU and γPho. The ratio of real-time values of γ's indicates a 294-fold acceleration. From .118

6.4. Cryptography and security

6.4.1. Optical steganography (bio-inspired by marine hatchetfish camouflage strategies]

Communication systems have been integrated deeply in our daily lives, supporting applications including online banking, augmented reality experience, and telemedicine. Therefore, it is expected that the massive amount of sensitive and personal information are needed to be protected against attacks. To provide information security, sophisticated encryption schemes are used at the higher layer of the communication system, i.e. media access control (MAC) layer. However, a physical layer without proper security measures makes a communication system vulnerable to attack, resulting in total exposure of sensitive information [198].

Effective physical cryptography requires both encryption and steganography. Encryption scrambles the sensitive information so that it is unreadable without the key, while steganography hides the sensitive information in plain sight so that the attacker will not even know there is sensitive information to look for. It is like hiding valuables in a locked safe (encryption) behind a secret bookcase door (steganography). Physical encryption schemes have been intensively studied, but physical layer steganography is still underdeveloped [199, Z. 200, 201].

Turning to nature for an effective solution, Marine Hatchetfish is a master of hiding its appearance in the deep ocean using unique ocean camouflage techniques [202]. Firstly, Marine Hatcherfish has microstructured skin on the sides so that only light that is similar to its surrounding is constructively interfered, while colors that could disclose its presence are destructively interfered, this technique is called silvering (Figure 19(a)). Secondly, Marine Hatchetfish also generates and directs light to the bottom part of its body to illuminate itself so that its color and brightness is the same as its surrounding when seen from below, this technique is called counter-illumination (Figure 19(b)). The two camouflage strategies allow the Marine Hatchetfish to conceal its appearance in all directions.

Borrowing the camouflage strategies from Marine Hatchetfish and applying it in RF signal transmission in optical fiber would be a natural and an effective way to achieve steganography. Silvering can be achieved using photonic FIR to make the sensitive signal disappear in the eyes of the attacker through destructive interference [Q. 203]. Figure 19(c) shows the transformation of the FIR response at (i) the stealth transmitter, (ii) during and after signal transmission in single mode fiber, (iii) after dispersion compensation, and (iv) at the designated stealth receiver. The stealth signal is invisible at any point during the transmission and can only be retrieval with a precise stealth receiver at the designated location, Furthermore, counter-illumination can be achieved using a noise-like optical carrier that has the same spectral content and intensity as the system noise, similar to how Marine Hatchetfish illuminate itself [Q. 203]. Steganography using bio-inspired silvering and counter-illumination techniques allow the sensitive signal to be concealed in all possible domains that the attacker could be looking at. Figure 19(d)i and ii shows the measured RF spectra and constellation diagrams of the stealth signal during transmission and at the designated stealth receiver. It is proven that the stealth signal is disappeared in the eyes of the attacker.

6.4.2. Encryption and decryption

Movement of massive data traffic is rapidly growing in this big-data era, i.e. the 4th industrial revolution of 'digitalization', spurred by the emergence of machine intelligence. Since the compute capacity at any one physical location is limited by the bounds of power and cooling requirements, it is naturally inevitable that data movement will be necessary. At the same time, it is critical to maintain information security during data transit between distributed computing locations. Encryption-in-transit mechanisms protect the integrity and confidentiality of sensitive information transmitted over Internet infrastructure between physical computing locations. However, due to the scale of data volumes (~Terabits/sec), efficient (en/de) cryption mechanisms are paramount for effectiveness of encryption-in-

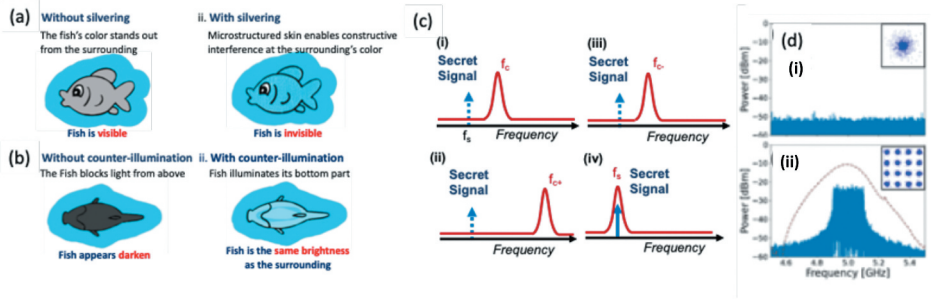

Figure 19. A) Side view (i) no camouflage – fish is visible (ii) silvering – fish is destructively interfered at colors that could indicate the presence of the fish; (b) Bottom view (i) no camouflage – fish appears darker against the bright water surface when seen from below (ii) counterillumination – fish illuminates itself to the same color and intensity as the background. (c) Silvering (i) photonic RF FIR creates destructive interference condition at the stealth signal frequency (f_s); (ii) Transmission in optical fiber will only push the constructive interference condition to a much higher frequency (f_c+); (iii) Dispersion compensation fiber at the last section of the transmission will move the constructive interference condition back to f_c; (iv) Correct dispersion at the stealth receiver allows constructive interference condition to occur at the stealth signal frequency f_s. (d) Measured RF spectra and constellation diagrams (i) during transmission without a correct stealth receiver (ii) at the designated stealth receiver with correct location and dispersion.

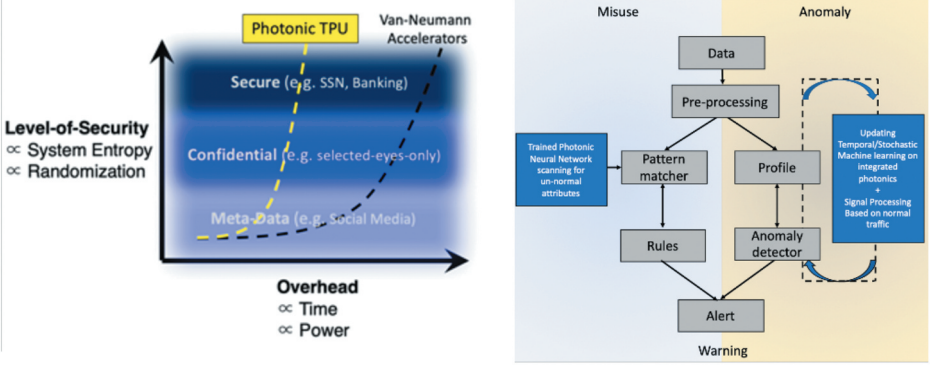

Figure 20. (left) Photonic machine-learning accelerators such as photonic tensor unit (PTU) [.138] enabled a higher level of data security with reduced overhead (e.g. time and power consumption). This opens possibilities for not only securing 'secure' and 'confidential' data, but also meta-data. [right) Flow-chart for detection of anomalies and misuses exploiting rapidly updating photonic neural network and trained photonic neural network includes programmable nonvolatile photonic memory on-chip for zero-static power consumption processing, once the kernel is written 138

transit – exactly what optical communication links provide. Since the data is already in the optical domain, thence, it is natural to consider (pre)processing information in the same optical and analog-domain (Figure 20). That is, avoiding domain crossings such as optical-to-electrical (OE, vise-versa) and eliminating digital-to-analog (and vise-versa) not only saves energy-per

-compute, but also improves system delay and reduces system complexity. The latter, is important for scaling vectors such as volume, reliability and ultimate cost of the system and hence the application.

Data security applications include three main areas: **Authentication**, which is the verification of data sources and destinations. Such functions are carried out by modules embedded within Trusted Computing Platforms, which ensure that only legitimate users can send and receive data. Modern advances in multi-factor authentication mechanisms [204] have incorporated several aspects such as knowledge factors, ownership factors and inherence factors. **Integrity**, which considers ensuring that the transmitted data arrives at the destination in an unmodified manner. Data integrity is usually guaranteed with checksum mechanisms that include error correcting codes. Such integrity-preserving mechanisms have been incorporated even inside modern hardware- for instance, some CPU architectures employ transparent checks to detect and mitigate data corruption in CPU caches, buffers and instruction pipelines as evidenced in Intel Instruction Replay mechanism in its Itanium processor family [205]. **Data Privacy** involves transformations that applied to legible data (plaintext user data) with the intent of making sure that it is only available to users that are authorized by the data owners. Typically, data encryption algorithms are used for achieving privacy guarantees, where the plaintext is transformed into cipher text before transmission and the keys needed for decrypting the encrypted cipher text are kept private. In order to secure web applications and systems, networks have to be able to promptly discern potential menaces and unwanted connections. Systems like intrusion detection (IDS) and intrusion prevention (IPS) are used for this purpose. Intrusion detection systems are divided in two groups: misuse detection (traditional IDS) and anomaly detection. Misuse detection systems are signature based, have high accuracy in detecting many kinds of known attacks but cannot detect unknown and emerging attacks. Our PTC, when properly trained according to previous knowledge of attacks, can be used as an intelligent comparator for the fast detection of misuse of the systems compared (performing convolution on string of data) to stored signatures of known exploits or attacks which are learnt in the photonic memories. This can be supplemented with anomaly-based intrusion detection as a prevention system. In fact, due to matrix multiplication and comparison performed at high-speed in optical accelerators, it can be used not just as smart pattern matcher, but as an evolving fast pre-screening of malicious activities, by collecting normal behaviors and detecting intrusion based on that, since new intrusion model can be implemented in the reprogrammable photonic memories thanks to the newly acquired and updating 'knowledge'. Optical processing of high parallelism, inherent to several cryptographic operations, is enabled by taking advantage of various attributes of light waves such as the wavelength, phase, polarization, and amplitude. As such, energy-efficient,

ultralow latency encryption-in-transit using optical accelerators can help address the grand challenges surrounding security in big data movement between computing systems.

6.5. Physics experiments

6.5.1. Intelligent prefiltering in astronomy and scientific applications

Astronomical radio observation and study on the galaxy formation have become extremely accurate thanks to the use of a plurality of telescopes arranged in an array, operating as a one single giant telescope. In this way, like all synthetic arrays, due to an enlarged equivalent aperture of 22 miles, the very large array (VLA], is sensitive and able to resolve a range of angular scales between the diffraction limit (Figure 21). A tremendous leap in this established technology is represented by the way the vast data obtained is collected and processed; data is fiber-optically fed to a supercomputer (WIDAR (*The WIDAR Supercomputer*, n.d.)) turning the VLA into a sensitive instrument. It is possible to obtain important information regarding star formation in 'interferometric' computing techniques, where the supercomputer correlates the hyper-spectral (wide band) signals from pairs of dishes obtaining a much sharper image than a single dish could produce. WIDAR uses FPGAs to perform correlations on the radiofrequency signal and only then sends relevant data to the cluster for further processing. Ultimately the cluster output is sent to an image-processing system. However, electronic data processing is limited by FPGA-setup times and fundamental electronic capacitive delay, resulting in delayed processing. To mitigate such processing limitations, photonic neural networks can be used as preprocessing unit to work synergistically with the WIDAR supercomputer on the vast data, in order to intelligently sorting and

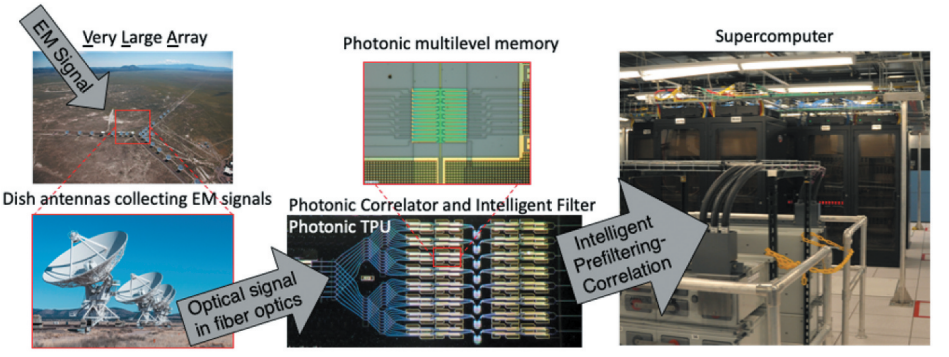

Figure 21. Photonic tensor core and neural networks enable intelligent prefiltering and correlation for scientific discovery, such as between electromagnetic signals collected by Very-Large-Array telescope systems performing intelligent pre-filtering, thus reducing computation load in supercomputers.

correlating the signal looking for specific chunks of radio-patterns (e.g. hydrogen gas moves into galaxies to fuel star formation) in near real time (~ps). Besides the increased speeds and bandwidths that can come from working directly in the optical domain, leveraging on the intrinsic optical nature of signal captured by the dish-antennas travelling in optical fibers, the advantage of using the photonic architectures consists in exploiting the wave-nature of the input signals to perform inherent correlation detection or convolution using pre-stored/programmable trained weights, without active power consumption and burdensome electro-optic conversions, as discussed above. In this way, the total amount of information to be handled by the supercomputer and consequently by the cluster unit is reduced, saving resources for useful data, favored by intelligent pre-screening, towards resolving the evolution of the universe [206].

6.5.2. Emerging applications: Quantum computer auxiliary systems and High-energy particle classification

Scalable quantum computing depends on classical auxiliary technologies for state reconstruction, calibration, and control. Neural networks have been used for quantum state reconstruction [207], tomography [208], and control [209210]. In some cases, the characteristic time constants of the state or instability are slow enough for a conventional computer to perform the task. In other cases, in particular for microwave qubits, the system is changing faster than a conventional computer can react. This means that a control and/or reconstruction task cannot occur in real-time. Photonic neural networks could reduce the latency of these operations, potentially opening up new opportunities to better monitor and stabilize quantum processing systems.

Photonic neural networks can exhibit latencies lower than electronic processors, whether neuromorphic, FPGAs, or ASICs. This low latency can make the crucial difference in certain applications where a decision must be made before the time to act passes. This critical time constraint exists in particle detectors such as CMS. Not all collisions are salient, so a trigger must classify the collision as salient or not before the next collision occurs. Hardware limitations dictate that the existing trigger uses rudimentary, non-adaptive algorithms that potentially overlook particular physics signatures. Recent work to improve the sophistication of particle classifications has adopted a neural network algorithm, implemented on FPGAs [209]. There is a potential for photonic neural networks to improve the performance of the time-critical triggering task, thus preserving physics signatures and enabling a higher collision rate.

7. Conclusion

The emergence of neural network models has highlighted the importance of interconnects and parallel processing, which is inherently advantageous to implement in photonics. The research community has built bridges between photonic device physics and neural networks. The performance improvement of photonic neural networks is expected to continue as new devices (e.g. modulators or lasers) based on new materials and nanostructures demonstrate their potential of further increasing the efficiency. The next generation of photonic devices could consume only hundreds of aJs of energy per time slot, allowing analog photonic MAC-based processors to consume even less per operation [12,53]. Meanwhile, advanced integration and fabrication techniques (e.g. silicon photonics) have provided an unprecedented platform to produce large-scale and low-cost photonic systems. The increased optical component density significantly extends the spectrum of information processing capabilities. Finally, monolithic fabrication, which integrates electronics and photonics on the same substrate, entails a tight co-integration of electronics and photonics, resulting in hybrid neuromorphic processors that can take the best advantages of both electronics and photonics depending on different applications.

In light of these developments, photonic neural networks have found places in many applications unreachable by conventional computing technology. Examples of applications explored in this paper include intelligent signal processing, high-performance computing, nonlinear programming, and control, enabling fundamental physics breakthroughs, etc. These applications particularly require low latency, high bandwidth, and low energies. To march ahead, we envisage a huge interest in developing the fundamental technologies [i.e. devices, fabrication/integration platforms, etc.) enabling large-scale photonic neural networks. We refer the interested readers to authors' recent review papers discussing a roadmap towards a large-scale photonic neural networks processor 5, 53. In parallel, more applications will be identified and demonstrated, along with the photonic platform development, promising to expand the application space of AI and information processing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the John R. Evans Leaders Fund from the Canadian Foundation for Innovation [Project number 37780]; Natural Sciences and Engineering Research Council of Canada [RGPIN-2018-05249].

ORCID

Bhavin J. Shastri http://orcid.org/0000-0001-5040-8248

References

[1] McCarthy J, Minsky ML, Rochester N, et al. A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag. 2006;27:12.
[2] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521:436–444.
[3] Berggren K, Xia Q, Likharev KK, et al. Roadmap on emerging hardware and technology for machine learning. Nanotechnology. 2020;32:012002.
[4] Schuman CD, Potok TE, Patton RM, et al. A survey of neuromorphic computing and neural networks in hardware. In arXiv preprint arXiv:1705.06963. 2017.
[5] Shastri BJ, Tait AN, De Lima TF, et al. Photonics for artificial intelligence and neuromorphic computing. Nat Photonics. 2021;15:102–114.
[6] Furber SB, Galluppi F, Temple S, et al. The spinnaker project. Proc IEEE. 2014;102:652–665.
[7] Merolla PA, Arthur JV, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. 2014;345:668–673.
[8] Schemmel J, Brüderle D, Grübl A, et al. (2010). A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *2010 ieee international symposium on circuits and systems (iscas)* (pp. 1947–1950).
[9] Govoreanu B, Kar G, Chen Y, et al. 10× 10nm 2 hf/hfo x crossbar resistive ram with excellent performance, reliability and low-energy operation. In: 2011 international electron devices meeting. 2011, Washington, DC, USA. p. 31–36.
[10] Yang JJ, Strukov DB, Stewart DR. Memristive devices for computing. Nat Nanotechnol. 2013;8:13–24.
[11] Miller DA. Device requirements for optical interconnects to silicon chips. Proc IEEE. 2009;97:1166–1185.
[12] Nahmias MA, De Lima TF, Tait AN, et al. Photonic multiply-accumulate operations for neural networks. IEEE J Sel Top Quantum Electron. 2019;26:1–18.
[13] Agarwal S, Jacobs-Gedrim RB, Bennett C, et al. Designing and modeling analog neural network training accelerators. In: 2019 international symposium on vlsi technology, systems and application (vlsi-tsa). 2019. p. 1–2).
[14] Ahmed AH, El Moznine A, Lim D, et al. A dual-polarization silicon-photonic coherent transmitter supporting 552 Gb/s/wavelength. IEEE J Solid-State Circuits. 2020;55:2597–2608.
[15] Ahmed AH, Sharkia A, Casper B, et al. Silicon-photonics microring links for datacenterschallenges and opportunities. IEEE J Sel Top Quantum Electron. 2016;22:194–203.
[16] Psaltis D, Farhat N. Optical information processing based on an associative- memory model of neural nets with thresholding and feedback. Opt Lett. 1985;10:98–100.

[17] Feldmann J, Youngblood N, Wright CD, et al. All- optical spiking neurosynaptic networks with self-learning capabilities. Nature. 2019;569:208–214.
[18] Thomson D, Zilkie A, Bowers JE, et al. Roadmap on silicon photonics. J Opt. 2016;18:073003.
[19] Shekhar S (2021). Tutorial: silicon photonics - from basics to ASICs. In *2021 ieee international solid-state circuits conference (isscc)*.
[20] Sun J, Timurdogan E, Yaacobi A, et al. Large-scale nanophotonic phased array. Nature. 2013;493:195–199.
[21] Han J, Jentzen A, Weinan E. Solving high-dimensional partial differential equations using deep learning. Proc Nat Acad Sci. 2018;115:8505–8510.
[22] Huang C, Fujisawa S, De Lima TF, et al., (2020). Demonstration of photonic neural network for fiber nonlinearity compensation in long-haul transmission systems. In *2020 optical fiber communications conference and exhibition (ofc)* (pp. 1–3), San Diego, California, USA.
[23] Khan FN, Fan Q, Lu C, et al. An optical communication's perspective on machine learning and its applications. J Lightwave Technol. 2019;37:493–516.
[24] Ma PY, Tait AN, Zhang W, et al. Blind source separation with integrated photonics and reduced dimensional statistics. Opt Lett. 2020;45:6494–6497.
[25] Prucnal PR, Shastri BJ. Neuromorphic photonics. Boca Raton: CRC Press; 2017.
[26] Cong J, Xiao B (2014). Minimizing computation in convolutional neural networks. In *International conference on artificial neural networks* (pp. 281–290), Hamburg, Germany.
[27] De Lima TF, Peng H-T, Tait AN, et al. Machine learning with neuromorphic photonics. J Lightwave Technol. 2019;37:1515–1534.
[28] De Lima TF, Tait AN, Saeidi H, et al. Noise analysis of photonic modulator neurons. IEEE J Sel Top Quantum Electron. 2019;26:1–9.
[29] Chen Y-H, Krishna T, Emer JS, et al. Eyeriss: an energy-efficient recon- figurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circuits. 2016;52:127–138.
[30] Chen Y-H, Emer J, Sze V. Eyeriss: a spatial architecture for energy-efficient dataflow for convolutional neural networks. ACM SIGARCH Comput Archit News. 2016;44:367–379.
[31] Gudaparthi S, Narayanan S, Balasubramonian R, et al. (2019). Wire-aware architecture and dataflow for cnn accelerators. In *Proceedings of the 52nd annual ieee/acm international symposium on microarchitecture* pp. (1–13), Columbus OH USA.
[32] Bankman D, Murmann B (2016). An 8-bit, 16 input, 3.2 pj/op switched-capacitor dot product circuit in 28-nm fdsoi cmos. In *2016 ieee asian solid-state circuits conference (a-sscc)* (pp. 21–24), Toyama, Japan.
[33] Skrzyniarz S, Fick L, Shah J, et al. (2016).24.3 a 36.8 2b-tops/w self-calibrating gps accelerator implemented using analog calculation in 65nm lp cmos. In *2016 ieee international solid-state circuits conference (isscc)* (pp. 420–422), San Francisco, CA, USA.
[34] Sarpeshkar R. Analog versus digital: extrapolating from electron- ics to neurobiology. Neural Comput. 1998;10:1601–1638. Retrieved from.
[35] Wattanapanitch W, Fee M, Sarpeshkar R (2007). An energy-efficient micropower neural recording amplifier. IEEE Trans Biomed Circuits Syst, 1, 136–147. *The WIDAR Supercomputer*. (n.d.). Retrieved 2021-03-07, from https://public.nrao.edu/gallery/the-widar-supercomputer/
[36] Boahen K. A neuromorph's prospectus. Comput Sci Eng. 2017;19:14–28.

[37] Brunner D, Soriano MC, Mirasso CR, et al. Parallel photonic information processing at gigabyte per second data rates using transient states. Nat Commun. 2013jan; 4: 1364Retrieved from
[38] Cao N, Chang M, Raychowdhury A. A 65-nm 8-to-3-b 1.00.36-v 9.11.1- tops/w hybrid-digital-mixed-signal computing platform for accelerating swarm robotics. IEEE J Solid-State Circuits. 2020 Jan;55:49–59.
[39] Capmany J, Mora J, Gasulla I, et al. Microwave photonic signal processing. J Lightwave Technol. 2013 feb;31:571–586.
[40] Chang G-K, Cheng L. The benefits of convergence. Philos Trans Royal Soc A. 2016 mar;374: 20140442. Retrieved from: http://rsta.royalsocietypublishing.org/lookup/doi/10.1098/rsta.2014.0442
[41] Chen H-W, Peters JD, Bowers JE (2011 jan). Forty Gb/s hybrid silicon Mach-Zehnder modulator with low chirp. Opt Express, 19, 1455. Retrieved from https://www.osapublishing.org/oe/abstract.cfm?uri=oe-19-2-1455
[42] Semenova N, Porte X, Andreoli L, et al. Fundamental aspects of noise in analog-hardware neural networks. Chaos: An Interdiscip J Nonlinear Sci. 2019;29:103128. Retrieved from.
[43] Semenova N, Larger L, Brunner D (2021). Noise in trained deep neural networks
[44] Biswas A, Chandrakasan AP (2018). Conv-ram: an energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications. In *2018 ieee international solid - state circuits conference - (isscc)* (p. 488–490), San Francisco, CA, USA.
[45] Yu S, Sun X, Peng X, et al. (2020). Compute-in-memory with emerging nonvolatile-memories: challenges and prospects. In *2020 ieee custom integrated circuits conference (cicc)* (p. 1–4), Boston, MA, United States.
[46] Liu Q, Gao B, Yao P, et al. (2020). 33.2 a fully integrated analog reram based 78.4tops/w compute-in-memory chip with fully parallel mac computing. In *2020 ieee international solid- state circuits conference - (isscc)* (p. 500–502), San Francisco, California.
[47] Marinella MJ, Agarwal S, Hsia A, et al. Multiscale co-design analysis of energy, latency, area, and accuracy of a reram analog neural training accelerator. IEEE J Emerg Selected Topics Circuits Syst. 2018;8:86–101.
[48] Hu M, Strachan JP, Li Z, et al. (2016). Dot-product engine for neuromorphic computing: programming 1t1m crossbar to accelerate matrix-vector multiplication. In *2016 53nd acm/edac/ieee design automation conference (dac)* (p. 1–6), Austin, TX, USA.
[49] Judd P, Albericio J, Hetherington T, et al. (2016). Stripes: bit-serial deep neural network computing. In *2016 49th annual ieee/acm international symposium on microarchitecture (micro)* (p. 1–12), Taipei, Taiwan.
[50] Choi J, Venkataramani S, Srinivasan V, et al. (2019). Accurate and efficient 2-bit quantized neural networks. In *Proceedings of the 2nd sysml conference* (Vol. 2019), Palo Alto, US.
[51] Miller DAB. Rationale and challenges for optical interconnects to electronic chips. Proc IEEE. 2000;88:728–749.
[52] Keyes RW. Optical logic-in the light of computer technology. *Optica Acta*. Int J Opt. 1985;32:525–535. Retrieved from.
[53] De Lima TF, Tait AN, Mehrabian A, et al. Primer on silicon neuromorphic photonic pro- cessors: architecture and compiler. Nanophotonics. 2020;9:4055–4073. Retrieved from.

[54] Hasler J, Marr H. Finding a roadmap to achieve large neur- omorphic hardware systems. Front Neurosci. 2013;7: 118. Retrieved from: https://www.frontiersin.org/article/10.3389/fnins.2013.00118
[55] Jouppi NP, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. *SIGARCH Comput*. Archit News. 2017bJune; 45: 112Retrieved from
[56] Shen Y, Harris NC, Skirlo S, et al. Deep learning with coherent nanophotonic circuits. Nat Photonics. 2017;11:441.
[57] Tait AN, Nahmias MA, Shastri BJ, et al. Broadcast and weight: an integrated network for scalable photonic spike processing. J Lightwave Technol. 2014;32:4029–4041.
[58] Bangari V, Marquez BA, Miller H, et al. Digital electronics and analog photonics for convolutional neural networks (deap-cnns). IEEE J Sel Top Quantum Electron. 2019;26:1–13.
[59] Shainline JM, Buckley SM, Mirin RP, et al. Superconducting optoelectronic circuits for neuromorphic computing. Phys Rev Appl. 2017;7:034013.
[60] Feldmann J, Youngblood N, Karpov M, et al. (2021 jan). Parallel convolutional processing using an in- tegrated photonic tensor core. Nature, 589, 52–58. Retrieved from http://www.nature.com/articles/s41586-020-03070-1
[61] Georgieva N, Glavic S, Bakr M, et al. Feasible adjoint sensitivity technique for em design optimization. IEEE Trans Microw Theory Tech. 2002;50:2751–2758.
[62] Goodman JW, Leonberger FJ, Sun-Yuan Kung, et al. Optical interconnections for vlsi systems. Proc IEEE. 1984;72:850–866.
[63] Nozaki K, Matsuo S, Fujii T, et al. Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions. Nat Photonics. 2019;13:454–459.
[64] Lin, C.H., Cheng, C.C., Tsai, Y.M., Hung, S.J., Kuo, Y.T., Wang, P. H., . . . others (2020). 7.1 a 3.4-to-13.3 tops/w 3.6 tops dual-core deep-learning accelerator for versatile ai applic- ations in 7nm 5g smartphone soc. In 2020 ieee international solid-state circuits conference- (isscc) (pp. 134–136)
[65] Xue C-X, Chen W-H, Liu J-S, et al., (2019). 24.1 a 1mb multibit reram computing-in-memory macro with 14.6 ns parallel mac computing time for cnn based ai edge processors. In *2019 ieee international solid-state circuits conference- (isscc)* (pp. 388–390), San Francisco, California.
[66] Huang C, Bilodeau S, Ferreira de Lima T, et al. Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits. APL Photonics. 2020;5:040803.
[67] Ramey C (2020). Silicon photonics for artificial intelligence acceleration: hotchips 32. In *2020 ieee hot chips 32 symposium (hcs)* (pp. 1–26), Palo Alto, CA, USA.
[68] Zhang W, Huang C, Bilodeau S, et al. Microring weight banks control beyond 8.5-bits accuracy. 2021.
[69] Amin R, George J, Sun S, et al. Ito-based electro-absorption modulator for photonic neural activation function. APL Mater. 2019;7:081112.
[70] George JK, Mehrabian A, Amin R, et al. Neuromorphic photonics with electro-absorption modulators. Opt Express. 2019;27:5181–5191.
[71] Guo X, Barrett TD, Wang ZM, et al. (2021 mar). Backpropagation through nonlinear units for the all-optical training of neural networks. Photonics Res, 9, B71. Retrieved from https://www.osapublishing.org/abstract.cfm?URI=prj-9-3-B71
[72] Tait AN, De Lima TF, Nahmias MA, et al. Silicon photonic modulator neuron. Phys Rev Appl. 2019;11:064043.
[73] Wang C, Zhang M, Chen X, et al. Integrated lithium niobate electro-optic modulators operating at cmos- compatible voltages. Nature. 2018;562:101–104.

[74] Kuo Y-H, Lee YK, Ge Y, et al. Strong quantum-confined stark effect in germanium quantum-well structures on silicon. Nature. 2005;437:1334–1336.
[75] Dong P, Liao S, Feng D, et al. Low v pp, ultralow-energy, compact, high-speed silicon electro-optic modulator. Opt Express. 2009;17:22484–22490.
[76] Xu Q, Schmidt B, Pradhan S, et al. Micrometre-scale silicon electro-optic modulator. nature. 2005;435:325–327.
[77] Amin R, Maiti R, Carfano C, et al. 0.52 v mm ito-based mach-zehnder modulator in silicon photonics. APL Photonics. 2018a;3:126104.
[78] Komljenovic T, Davenport M, Hulme J, et al. Heterogeneous silicon photonic integrated circuits. J Lightwave Technol. 2016;34:20–35.
[79] Liu M, Yin X, Ulin-Avila E, et al. A graphene-based broadband optical modulator. Nature. 2011;474:64–67.
[80] Huang C, De Lima TF, Jha A, et al. Programmable silicon photonic optical thresholder. IEEE Photonics Technol Lett. 2019;31:1834–1837.
[81] *The History of Artificial Intelligence*. (2017, August). Retrieved 2021-03-08, from https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
[82] Hughes TW, Minkov M, Shi Y, et al. (2018 jul). Training of photonic neural networks through in situ backpropagation and gradient measurement. Optica, 5, 864. Retrieved from https://www.osapublishing.org/abstract.cfm?URI=optica-5-7-864
[83] Jaeger H, Haas H (2004 apr). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science, 304, 78–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15064413
[84] Jayatilleka H, Murray K, Guillén-Torres MÁ, et al. (2015 sep). Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters. Opt Express, 23, 25084. Retrieved from https://www.osapublishing.org/abstract.cfm?URI=oe-23-19-25084
[85] Jha A, Huang C, Prucnal PR. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. Opt Lett. 2020;45:4819–4822.
[86] Zhang J, Wang Z, Verma N (2016). A machine-learning classifier implemented in a standard 6t sram array. In *2016 ieee symposium on vlsi circuits (vlsi-circuits)* (p. 1–2), Honolulu, HI, USA.
[87] Chakraborty I, Saha G, Sengupta A, et al. Toward fast neural computing using all-photonic phase change spiking neurons. Sci Rep. 2018;8:1–9.
[88] Miscuglio M, Mehrabian A, Hu Z, et al. All-optical nonlinear activation function for photonic neural networks. Opt Mater Express. 2018;8:3851–3863.
[89] Giewont K, Nummy K, Anderson FA, et al. 300-mm monolithic silicon photonics foundry technology. IEEE J Sel Top Quantum Electron. 2019;25:1–11.
[90] Amin R, Suer C, Ma Z, et al. Active material, optical mode and cavity impact on nanoscale electro-optic modulation performance. Nanophotonics. 2018;7:455–472.
[91] Amin R, Suer C, Ma Z, et al. Active material, optical mode and cavity impact on nanoscale electro-optic modulation performance. Nanophotonics. 2017;7:455–472.
[92] Larger L, Soriano MC, Brunner D, et al. (2012 jan). Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. Opt Express, 20, 3241–3249. Retrieved from http://www.opticsexpress.org/abstract.cfm?URI=oe-20-3-3241
[93] Lin C-H, Cheng -C-C, Tsai Y-M, et al., (2020). 7.1 a 3.4-to-13.3 tops/w 3.6 tops dual-core deep-learning accelerator for versatile ai applications in 7nm 5g smartphone soc. In *2020 ieee international solid-state circuits conference- (isscc)* (pp. 134–136), San Francisco, California.

[94] Lin P, Li C, Wang Z, et al. Three-dimensional memristor circuits as complex neural networks. Nat Electron. 2020;3:225–232.
[95] Lin R, Ge J, Tran P, et al. Biomimetic photonics: jamming avoidance system in eigenmannia. Opt Express. 2018;26:13349–13360.
[96] Lin X, Rivenson Y, Yardimci NT, et al. All-optical machine learning using diffractive deep neural networks. Science. 2018;361:1004–1008.
[97] Brunstein M, Yacomotti AM, Sagnes I, et al. Excitability and self-pulsing in a photonic crystal nanocavity. Phys Rev A. 2012;85:031803.
[98] Williamson IA, Hughes TW, Minkov M, et al. Repro- grammable electro-optic nonlinear activation functions for optical neural networks. IEEE J Sel Top Quantum Electron. 2019;26:1–12.
[99] Jacques M, Samani A, El-Fiky E, et al. Optimization of thermo-optic phase-shifter design and mitigation of thermal crosstalk on the SOI platform. Opt Express. 2019;27:10456.
[100] Patel D, Ghosh S, Chagnon M, et al. Design, analysis, and transmission system performance of a 41 GHz silicon photonic modulator. Opt Express. 2015;23:14263.
[101] Sorianello V, Midrio M, Contestabile G, et al. Graphene-silicon phase modulat- ors with gigahertz bandwidth. Nat Photonics. 2018;12:40–44. Retrieved from.
[102] He M, Xu M, Ren Y, et al. High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbit s −1 and beyond. Nat Photonics. 2019;13:359–364. Retrieved from.
[103] Hiraki T, Aihara T, Hasebe K, et al. Heterogeneously integrated III-V/Si MOS capacitor Mach- Zehnder modulator. Nat Photonics. 2017;11:482–485. Retrieved from.
[104] Amin R, Maiti R, Carfano C, et al. (2018b, aug). 0.52 V-mm ITO-based mach-Zehnder modulator in silicon photonics. APL Photonics, 3, 126104. Retrieved from http://aip.scitation.org/doi/10.1063/1.5052635http://arxivorg/abs/180903544
[105] Green WM, Rooks MJ, Sekaric L, et al. (2007). Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator. Opt Express, 15, 17106. Retrieved from https://www.osapublishing.org/oe/abstract.cfm?uri=oe-15-25-17106
[106] Ríos C, Youngblood N, Cheng Z, et al. In-memory computing on a photonic platform. Sci Adv. 2019;5:eaau5759.
[107] Dorren H, Lenstra D, Liu Y, et al. Nonlinear polarization rotation in semiconductor optical amplifiers: theory and application to all-optical flip-flop memories. IEEE J Quantum Electron. 2003;39:141–148.
[108] Hill MT, De Waardt H, Khoe G, et al. Fast optical flip-flop by use of mach–zehnder interferometers. Microw Opt Technol Lett. 2001b;31:411–415.
[109] Hill MT, De Waardt H, Khoe G, et al. All-optical flip-flop based on coupled laser diodes. IEEE J Quantum Electron. 2001a;37:405–413.
[110] Wang J, Zhang Y, Malacarne A, et al. Soa fiber ring laser-based three-state optical memory. IEEE Photonics Technol Lett. 2008;20:1697–1699.
[111] Alexoudi T, Kanellos GT, Pleros N. Optical ram and integrated optical memories: a survey. Light Sci Appl. 2020;9:1–16.
[112] Meng J, Miscuglio M, George JK, et al. Electronic bot- tleneck suppression in next-generation networks with integrated photonic digital-to-analog converters. In: Advanced photonics research. 2019. p. 2000033.
[113] Nahmias MA, Peng H-T, De Lima TF, et al. A laser spiking neuron in a photonic integrated circuit. In arXiv preprint arXiv:2012.08516. 2020.
[114] Peng H-T, Nahmias MA, De Lima TF, et al. Neuromorphic photonic integrated circuits. IEEE J Sel Top Quantum Electron. 2018;24:1–15.

[115] Peng G, Boxun L, Tang T, et al. (2015). Technological exploration of rram crossbar array for matrix-vector multiplication. In *The 20th asia and south pacific design automation conference* (p. 106–111), Tokyo, Japan.
[116] Artech H, Porte X, Skalli A, et al. (2021, apr). A complete, parallel and autonomous photonic neural network in a semiconductor multimode laser. J Phys, 3, 024017. Retrieved from https://iopscience.iop.org/article/10.1088/2515-7647/abf6bd
[117] Miscuglio M, Hu Z, Li S, et al. Massively parallel amplitude-only Fourier neural network. Optica. 2020;7:1812–1819.
[118] Tait AN, De Lima TF, Zhou E, et al. Neuromorphic photonic networks using silicon photonic weight banks. Sci Rep. 2017;7:1–10.
[119] Rafayelyan M, Dong J, Tan Y, et al. (2020). Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction. Phys Rev X, 10, 041037. Retrieved from https://journals.aps.org/prx/abstract/10.1103/PhysRevX.10.041037
[120] Ríos C, Stegmaier M, Hosseini P, et al. (2015, Nov). Integrated all-photonic non-volatile multi-level memory. Nat Photonics, 9, 725–732. Retrieved from http://www.nature.com/articles/nphoton.2015.182
[121] Bueno Moragues J, Maktoobi S, Froehly L, et al. (2018). Reinforcement learning in a large-scale photonic recurrent neural network.
[122] Gu J, Feng C, Zhao Z, et al. Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization. In: arXiv preprint arXiv:2012.11148. 2020.
[123] Sundstrom T, Murmann B, Svensson C. Power dissipation bounds for high-speed nyquist analog-to-digital converters. IEEE Trans Circuits Syst I: Reg Papers. 2009 march;56:509–518.
[124] Sze V, Chen Y-H, Yang T-J, et al. Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE. 2017;105:2295–2329.
[125] Horowitz M (2014). 1.1 computing's energy problem (and what we can do about it). In *2014 ieee international solid-state circuits conference digest of technical papers (isscc)* (pp. 10–14), San Francisco, CA, USA.
[126] Imani M, Patil S, Rosing T (2016). Low power data-aware stt-ram based hybrid cache architecture. In *2016 17th international symposium on quality electronic design (isqed)* pp. (88–94), Santa Clara, CA, United States.
[127] Tait AN, Jayatilleka H, Lima TFD, et al. Feedback control for microring weight banks. Opt Express. 2018 Oct;26:26422–26443.
[128] Vandoorne K, Mechet P, Van Vaerenbergh T, et al. (2014 mar). Experimental demonstration of reservoir comput- ing on a silicon photonics chip. Nat Commun, 5, 3541. Retrieved from http://www.nature.com/doifinder/10.1038/ncomms4541
[129] Walden RH. Analog-to-digital converter survey and analysis. IEEE J Sel Areas Commun. 1999 apr;17:539–550.
[130] Tran K (2016). The era of high bandwidth memory. In *2016 ieee hot chips 28 symposium (hcs)* (pp. 1–22), Cupertino, CA.
[131] Xiao TP, Bennett CH, Feinberg B, et al. Analog architectures for neural network acceleration based on non-volatile memory. Appl Phys Rev. 2020;7:031301.
[132] Xu X, Tan M, Corcoran B, et al. (2021 jan). 11 TOPS photonic convolutional accelerator for optical neural networks. Nature, 589, 44–51. Retrieved from http://www.nature.com/articles/s41586-020-03063-0
[133] Mukherjee I, Saurav K, Nair P, et al. (2021). A case for emerging memories in dnn accelerators. In *Design, automation & test in europe conference & exhibition (date)*. Murmann, B. (n.d.), Antwerp, Belgium. Retrieved from http://webstanfordedu/murmann/adcsurveyhtml

[134] Choi JW, Sohn BU, Chen GF, et al. (2017). Nonlinear optical properties of gesbs chalcogenide waveguides. *2017 Opto-Electronics and Communications Conference, OECC 2017 and Photonics Global Con- ference, PGC 2017, 2017-Novem*, 1–2, Singapore, Singapore.
[135] Miscuglio M, Meng J, Yesiliurt O, et al. (2020). Artificial synapse with mnemonic functionality using gsst-based photonic integrated memory. In *2020 international applied computational electromagnetics society symposium (aces)* pp. (1–3), Monterey, CA, USA.
[136] Li X, Youngblood N, Ros C, et al. Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. Optica. 2019;6:1.
[137] Jung Y, Jeong J, Qu Z, et al. Obser- vation of optically addressable nonvolatile memory in vo2 at room temperature. In Advanced electronic materials. 2021.
[138] Miscuglio M, Sorger VJ. Photonic tensor cores for machine learning. Appl Phys Rev. 2020;7:031404.
[139] Zhang Y, Ros C, Shalaginov MY, et al. Myths and truths about optical phase change materials: a perspective. Appl Phys Lett. 2021;118.
[140] Kim B, Cho M-H, Kim Y-G, et al. A 1 v 6-bit 2.4 gs/s nyquist cmos dac for uwb systems. 2010; 912–915.
[141] Sedighi B, Khafaji M, Scheytt JCS. 8-bit 5gs/s d/a converter for multi-gigabit wireless transceivers. 2011. p. 192–195.
[142] Lin J, Hsieh C. A 0.3 v 10-bit 1.17 f sar adc with merge and split switching in 90 nm cmos. IEEE Trans Circuits Syst I: Reg Papers. 2015;62:70–79.
[143] Liou C, Hsieh C (2013). A 2.4-to-5.2fj/conversion-step 10b 0.5-to-4ms/s sar adc with charge-average switching dac in 90nm cmos. In *2013 ieee international solid-state circuits conference digest of technical papers* p. (280–281).
[144] Zhu Y, Chan C, Chio U, et al. A 10-bit 100-ms/s reference-free sar adc in 90 nm cmos. IEEE J Solid-State Circuits. 2010;45:1111–1121.
[145] Akrout M, Wilson C, Humphreys PC, et al. (2019). Deep learning without weight transport. (NeurIPS). 07 Sept 2019. Retrieved from http://arxiv.org/abs/1904.05391
[146] Scellier B, Bengio Y. Equilibrium propagation: bridging the gap between energy-based models and backpropagation. Front Comput Neurosci. 2017;11:1–13.
[147] Psaltis D, Wagner K (1987). Multilayer optical learning networks. Applied optics, 26 (23), 5061–5076.
[148] Zhou T, Fang L, Yan T, et al. In situ optical backpropagation training of diffractive optical neural networks. Photonics Res. 2020;8:940–953.
[149] Zhou T, Lin X, Wu J, et al. (2021 may). Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat Photonics, 15, 367–373. Retrieved from http://www.nature.com/articles/s41566-021-00796-w
[150] Antonik P, Marsal N, Brunner D, et al. (2021). Bayesian optimisation of large-scale photonic reservoir computers. *Cognitive Computation*. Retrieved from 10.1007/s12559-020-09732-6
[151] Carolan J, Harrold C, Sparrow C, et al. (2015). Universal linear optics. Science, 349, 711–716. Retrieved from https://science.sciencemag.org/content/349/6249/711
[152] Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: training deep neural networks with weights and activations constrained to+ 1 or-1. In: arXiv preprint arXiv:1602.02830. 2016.
[153] Hirtzlin T, Bocquet M, Penkovsky B, et al. Digital biologically plausible implementation of binarized neural networks with differential hafnium oxide resistive memory arrays. Front Neurosci. 2020;13:1383.

[154] Bueno J, Brunner D, Soriano MC, et al. Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback. Opt Express. 2017;25:2401–2412.
[155] Song S, Miller KD, Abbott LF. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci. 2000;3:919–926.
[156] Fok MP, Tian Y, Rosenbluth D, et al. Pulse lead/lag timing detection for adaptive feedback and control based on optical spike-timing-dependent plasticity. Opt Lett. 2013;38:419–421.
[157] Toole R, Tait AN, De Lima TF, et al. Photonic implementation of spike-timing-dependent plasticity and learning algorithms of biological neural systems. J Lightwave Technol. 2015;34:470–476.
[158] Toole R, Fok MP (2015). Photonic implementation of a neuronal learning algorithm based on spike timing dependent plasticity. In *Optical fiber communication conference* (pp. W1K–6), Los Angeles, CA.
[159] Cisco. (n.d.). *Cisco annual internet report - cisco annual internet report (20182023) white pa- per*. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490html. (Accessed: 2021 Feb 17)
[160] Pillai BSG, Sedighi B, Guan K, et al. End-to-end energy modeling and analysis of long-haul coherent transmission systems. J Lightwave Technol. 2014;32:3093–3111.
[161] Agrell E, Karlsson M, Chraplyvy A, et al. Roadmap of optical communications. J Opt. 2016;18:063002.
[162] Argyris A, Bueno J, Fischer I. Photonic machine learning implementation for signal recovery in optical communications. Sci Rep. 2018;8:1–13.
[163] Zhang S, Yaman F, Nakamura K, et al. Field and lab experimental demonstration of nonlinear impairment compensation using neural networks. Nat Commun. 2019;10:1–8.
[164] Peng H-T, Lederman J, Xu L, et al. A photonic-circuits-inspired compact network: toward real-time wireless signal classification at the edge. 2021.
[165] Chagnon M. Optical communications for short reach. J Lightwave Technol. 2019;37:1779–1797.
[166] Appeltant L, Soriano MC, Van der Sande G, et al. Information processing using a single dynamical node as complex system. Nat Commun. 2011;2:1–6.
[167] Sorokina M, Sergeyev S, Turitsyn S. Fiber echo state network analogue for high-bandwidth dual-quadrature signal processing. Opt Express. 2019;27:2387–2395.
[168] Da Ros F, Ranzini SM, Bülow H, et al. Reservoir-computing based equalization with optical pre-processing for short-reach optical transmission. IEEE J Sel Top Quantum Electron. 2020;26:1–12.
[169] Li S, Pachnicke S. (2020). Photonic reservoir computing in optical transmission systems. In *2020 ieee photonics society summer topicals meeting series (sum)* (pp. 1–2), Cabo San Lucas, Mexico.
[170] Poisel R (2011). *Modern communications jamming principles and techniques*
[171] Wilhelm M, Martinovic I, Schmitt JB, et al. (2011). Short paper: reactive jamming in wireless networks: how realistic is the threat? In *Proceedings of the fourth acm conference on wireless network security* (pp. 47–52), New York, NY, United States.
[172] Bullock, Theodore H., Robert H. Hamstra, and Henning Scheich. "The jamming avoidance response of high frequency electric fish." How do Brains Work?. Birkhäuser, Boston, MA, 1972. 509-534.
[173] Scheich H. Neural basis of communication in the high frequency electric fish, eigenmannia virescens (jamming avoidance response). J Comp Physiol. 1977;113:229–255.

[174] Fok MP, Toole R (2018). *Photonic implementation of jamming avoidance response*. Google Patents. (US Patent 9,954,619)
[175] Toole R, Fok MP (2016). A photonic rf jamming avoidance response system bio-inspired by eigenmannia. In *2016 optical fiber communications conference and exhibition (ofc)* (pp. 1–3), Anaheim, CA.
[176] Capmany J, Novak D (2007, 06). Microwave photonics combines two worlds. Nat Photon, 1, 319–330. Retrieved from
[177] Tait AN, De Lima TF, Ma PY, et al. (2018). Blind source separation in the physical layer. In *2018 52nd annual conference on information sciences and systems (ciss)* pp. (1–6), Princeton, NJ, USA.
[178] Marpaung D, Roeloffzen C, Heideman R, et al. Laser photonics rev. n.d.
[179] Ambs P. Optical computing: a 60-year adventure. In: Advances in optical technologies. 2010.
[180] Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381:607–609.
[181] Li X, Zhang G, Huang HH, et al. (2016, Aug). Performance analysis of GPU-based convolutional neural networks. In *2016 45th international conference on parallel processing (icpp)* p. (67–76), Philadelphia PA , USA.
[182] Jaderberg M, Vedaldi A, Zisserman A (2014). Speeding up convolutional neural networks with low rank expansions. *CoRR, abs/1405.3866*. Retrieved from http://arxiv.org/abs/1405.3866
[183] Goodfellow I, Bengio Y, Courville A. Deep learning. The MIT Press; 2016.
[184] Jouppi NP, Young C, Patil N, et al. (2017a). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture* (p. 112). New York, NY, USA: Association for Computing Machinery. Retrieved from 10.1145/3079856.3080246
[185] Deubel M, Von Freymann G, Wegener M, et al. Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. Nat Mater. 2004;3:444–447.
[186] Moughames J, Porte X, Thiel M, et al. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks. Optica. 2020;7:640–646.
[187] Dinc NU, Lim J, Kakkava E, et al. (2020). Computer generated optical volume elements by additive manufacturing. Nanophotonics, 1 (ahead-of-print).
[188] Dinc NU, Psaltis D, Brunner D. Optical neural networks: the 3d connection. Photoniques. 2020;34–38. 10.1051/photon/202010434
[189] Hughes TW, Williamson IAD, Minkov M, et al. (2019). Wave phys- ics as an analog recurrent neural network. Sci Adv, 5 . Retrieved from https://advances.sciencemag.org/content/5/12/eaay6946
[190] Molesky S, Lin Z, Piggott AY, et al. Inverse design in nanophotonics. Nat Photonics. 2018;12:659–670. Retrieved from.
[191] Sorger VJ, Lanzillotti-Kimura ND, Ma R-M, et al. Ultra-compact silicon nanophotonic modulator with broadband response. Nanophotonics. 2012;1:17–22.
[192] Amin R, Maiti R, George JK, et al. A lateral mos-capacitor-enabled ito mach–zehnder modulator for beam steering. J Lightwave Technol. 2020;38:282–290.
[193] Amin R, Maiti R, Gui Y, et al. Sub- wavelength ghz-fast broadband ito mach–zehnder modulator on silicon photonics. Optica. 2020;7:333–335.
[194] Cichocki A, Unbehauen R, Swiniarski RW. Neural networks for optimization and signal processing. Vol. 253. wiley New York; 1993.
[195] Hopfield JJ, Tank DW. Computing with neural circuits: a model. Science. 1986;233:625–633.

[196] Stewart TC, Eliasmith C. Large-scale synthesis of functional spiking neural circuits. Proc IEEE. 2014;102:881–898.
[197] Stewart TC, DeWolf T, Kleinhans A, et al. Closed-loop neuromorphic benchmarks. Front Neurosci. 2015;9:464.
[198] Skorin-Kapov N, Furdek M, Zsigmond S, et al. Physical-layer security in evolving optical networks. IEEE Commun Mag. 2016;54:110–117.
[199] Fok MP, Wang Z, Deng Y, et al. Optical layer security in fiber-optic networks. IEEE Trans Inf Forensics Secur. 2011;6:725–736.
[200] Wang Z, Prucnal PR. Optical steganography over a public dpsk channel with asynchronous detection. IEEE Photonics Technol Lett. 2010;23:48–50.
[201] Wu B, Wang Z, Tian Y, et al. Optical steganography based on amplified spontaneous emission noise. Opt Express. 2013;21:2065–2071.
[202] Rosenthal EI, Holt AL, Sweeney AM. Three-dimensional midwater camouflage from a novel two-component photonic structure in hatchetfish skin. J Royal Soc Interface. 2017;14:20161034.
[203] Liu Q, Fok MP. Bio-inspired photonics–marine hatchetfish camouflage strategies for rf steganography. Opt Express. 2021;29:2587–2596.
[204] Bhargav-Spantzel A, Squicciarini AC, Modi S, et al. Privacy preserving multi-factor authentication with biometrics. J Com Put Secur. 2007;15:529–560.
[205] Bostian S. Rachet up reliability for mission-critical applications: intel® instruction replay technology. White Paper. 2013.
[206] *VLA Begins Huge Project of Cosmic Discovery*. (n.d.). Retrieved 2021-03-07, from https://public.nrao.edu/news/vla-begins-huge-project/
[207] Flurin E, Martin LS, Hacohen-Gourgy S, et al. Using a recurrent neural network to reconstruct quantum dynamics of a superconducting qubit from physical observations. Phys Rev X. 2020;10:011006.
[208] Torlai G, Mazzola G, Carrasquilla J, et al. Neural- network quantum state tomography. Nat Phys. 2018;14:447–450.
[209] Niu MY, Boixo S, Smelyanskiy VN, et al. Universal quantum control through deep reinforcement learning. Npj Quantum Inf. 2019;5:1–8.
[210] Duarte J, Han S, Harris P, et al. Fast inference of deep neural networks in fpgas for particle physics. J Instrum. 2018;13:P07027.