IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 5, 2019, accepted November 24, 2019, date of publication December 2, 2019,
date of current version December 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957245

Photonic Neural Networks: A Survey

LORENZO DE MARINIS[1], MARCO COCOCCIONI[2], (Senior Member, IEEE), PIERO CASTOLDI[1], (Senior Member, IEEE), AND NICOLA ANDRIOLLI[3], (Senior Member, IEEE)

[1]Scuola Superiore Sant'Anna, 56124 Pisa, Italy
[2]Department of Information Engineering, University of Pisa, 56122 Pisa, Italy
[3]National Research Council of Italy, Institute of Electronics, Computer and Telecommunication Engineering (CNR-IEIIT), 56122 Pisa, Italy

Corresponding author: Nicola Andriolli (nicola.andriolli@ieiit.cnr.it)

This work was supported in part by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab Project (Departments of Excellence program), granted to the Department of Information Engineering of the University of Pisa.

ABSTRACT Photonic solutions are today a mature industrial reality concerning high speed, high throughput data communication and switching infrastructures. It is still a matter of investigation to what extent photonics will play a role in next-generation computing architectures. In particular, due to the recent outstanding achievements of artificial neural networks, there is a big interest in trying to improve their speed and energy efficiency by exploiting photonic-based hardware instead of electronic-based hardware. In this work we review the state-of-the-art of photonic artificial neural networks. We propose a taxonomy of the existing solutions (categorized into multilayer perceptrons, convolutional neural networks, spiking neural networks, and reservoir computing) with emphasis on proof-of-concept implementations. We also survey the specific approaches developed for training photonic neural networks. Finally we discuss the open challenges and highlight the most promising future research directions in this field.

INDEX TERMS Artificial neural networks, neural network hardware, photonics, neuromorphic computing, photonic neural networks.

I. INTRODUCTION

Nowadays machine learning technology is used in an impressively large number of applications, comprising image classification, speech recognition and language translation, decision making, web searches, content filtering on social networks, recommendations on e-commerce websites, etc. [1]. Artificial Neural Networks (ANN) are useful for processing large data sets, combining and analyzing vast amounts of information quickly and without the need of explicit instructions [2].

Multiple neural network architectures have been investigated and implemented, suited to different application needs. For the implementation of massively interconnected ANN, the conventional computer architecture is fundamentally inefficient and not scalable with respect to computation, memory, and communication [3], [4].

To address the shortcomings of today's computer architecture for neural networks with the aim of increasing the computing speed and power efficiency, a growing effort from both the academia and the industry has focused on the development of specifically tailored electronic architectures. Graphics Processing Units (GPU) have been identified as particularly suitable for implementing the parallel computing tasks typical of ANN, and significantly contributed to the current success of machine learning in real application scenarios. Recently, Field-Programmable Gate Arrays (FPGA) and application-specific integrated circuits (ASIC) [5], [6] (including Google Tensor Processing Units – TPU – [7], IBM TrueNorth [4], Fujitsu's Deep Learning Unit, and Intel Nervana [8]) have been specifically designed to implement ANN computations. To this aim, these novel electronic solutions focus on advanced numerical representations, memory architectures suitable for high-speed matrix multiplications, and a very high bidirectional off-chip bandwidth (exceeding a Tb/s) to enable model and data parallelism. Further research aimed at speeding up electronic ANN by means of analog architectures based on memristors [9], [10]. Very recently, an EU-funded project has started [11], with the aim of developing a new integrated circuit technology for ANN where low-voltage field-effect transistors and non-volatile memories are tightly integrated exploiting quantum engineering

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

of heterostructures of two-dimensional materials. All these research and development activities aim to improve both speed and energy efficiency of machine learning tasks.

Over the years, photonic solutions for optical communication and processing evolved along the same lines, aiming at increasing the transmission speed and the energy efficiency [12]. For this reason, optical implementations of neural networks have been investigated since a long ago, aimed at exploiting the large parallelism (through degrees of freedom such as wavelength, polarization, and mode) and the high connectivity achievable with optics [13]–[15]. Additionally, many linear transformations can be performed with passive optics without power consumption and with minimal latency [16], and can be then detected nowadays at rates in excess of 50 Gb/s [17]. The feasibility of optical logic gates has also been demonstrated [18]–[20]. Furthermore many optical nonlinearities can in principle be used to implement the nonlinear function in each neuron [21]–[23]. These features indicate that optical implementations of neural networks can overcome electronic solutions in terms of computational speed and energy efficiency.

In this paper we present a survey of the approaches pursued in the field of Photonic Neural Networks (PNN) – alternatively called, sometimes, photonic neuromorphic computing – and we also propose a classification of the existing solutions. Previous surveys dealt with just a specific class of PNN approaches [24]–[27] or focused either on biologically inspired approaches [28] or on bottlenecks of photonic technologies and possible ways to overcome them [29].

The remainder of the paper is organized as follows: in Sec. II we present the motivations behind PNN, and introduce a taxonomy of the approaches present in the literature. In Sec. III we review the most relevant solutions categorized according to the previously proposed taxonomy, while in Sec. IV we describe the specific approaches devised for training PNN. Sec. V discusses open issues and perspectives in this field, while Sec. VI concludes the paper.

II. PHOTONIC NEUROMORPHIC COMPUTING

By following Moore's law [30], in the last decades electronic ANN achieved enormous improvements in terms of power consumption, size, and speed. In particular, the breakthrough of neural networks in the recent years was driven by the adoption of GPU/TPU, i.e., electronic hardware tailored to efficiently perform the matrix multiplications needed for inference and training in neural networks [31]. However, electronic solutions still face a main bottleneck in the interconnect problem: data transfer between processors and memories is constrained by unavoidable bandwidth limitations and is the main source of power consumption even in very optimized architectures [32].

Recently we have witnessed the rise of analog electronic circuits aimed to replace and outperform specific parts of a digital computer on specific tasks. The most notable devices in this sense are memristors, two-terminal passive elements able to "remember" the charge flow through them by a resistance change [33]. The inherent reconfigurability of memristors has been exploited mainly in crossbar array architectures to form parallel weighting units in spiking ANN [34], [35]. The main drawback of memristors concerns the high power dissipation (being resistance-based), IR drops in the array [36], the lack of accurate models for mainstream simulation tools, and the absence of process standards [33].

Photonics showed great potential at outperforming electronic ANN. A number of research efforts have been undertaken in the field of photonic devices implementing neural network functionalities. The rationale behind these studies lies in the expected enhancements in terms of computational speed and energy efficiency when carrying out training and inference tasks in optics, compared to state-of-the art electronic solutions [37]. Photonic approaches can considerably reduce the energy budget both in logic operations and data movement using passive optical circuits to perform the linear [38], [39] and in some cases nonlinear [40], [41] operations typical of a neural network. The use of passive elements in optics results in ultra-high operation speeds without energy consumption beyond transmitters and receivers. Another relevant feature of photonics, that can be suitably exploited in the context of ANN, is its inherent parallelism (underlying ANN themselves), which enables the distribution of the computing power across the network, with each neuron performing small parts of the computation in parallel [29]. While pioneering attempts [18] to replicate in photonics the classical boolean electronic logic circuits did not prove successful, the use of analog photonic computing devices is today a promising research direction especially suited for neural networks, which require fast and energy-efficient (although approximated) computations.

Early research on PNN dates back to more than thirty years ago [13], [14], [42], [43]. However, early implementations were bulky, not scalable, and suffered from the lack of an adequate technology. A relevant enabling technology that allowed to overcome some limitations of the initial solutions is integrated photonics. Among the available photonic integration platforms, efforts focused mainly on Silicon [44] and Indium Phosphide [45]–[47], being the most relevant platforms also for optical communications [48]. In the following we briefly review these material platforms.

Silicon is a great material for integrated photonics: it is transparent at communication wavelengths (i.e. 1270-1625 nm), its refractive index (3.48 at 1550 nm) guarantees a large contrast with respect to silica (1.45 at 1550 nm), and can be controlled thermally, electrically, mechanically (strain) or chemically (doping) [29]. Additionally, Silicon photonics benefits from the technology advancements brought over the years by the CMOS technology, which created a mass production platform that results both reliable and cost-effective. For these reasons, Silicon has been widely used to realize the passive elements exploited in integrated PNN, e.g., waveguides, modulators, splitters/combiners and filters.

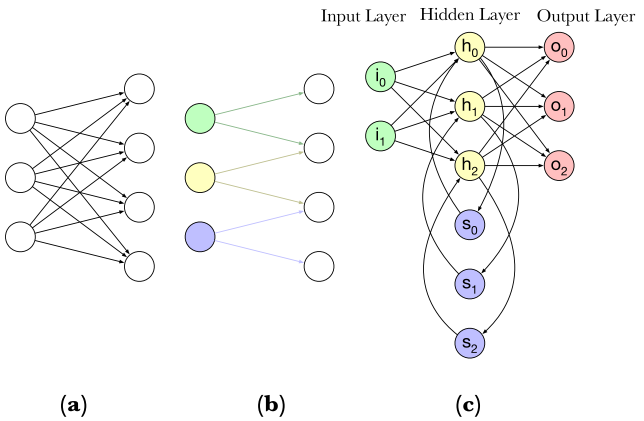


FIGURE 1. **(a) Perceptron layers (fully connected). (b) Convolutional layers: each neuron of the next layer is connected to only a small amount of neurons of the previous layer, not all of them. This sparse topology is inspired to receptive fields found in biological systems. Both (a) and (b) are stateless networks, i.e., networks without memory. In real implementations, convolutional neural networks can also contain fully connected layers, especially the ones close to the output layer. (c) Stateful networks: networks having some hidden neurons that store the past values of some other neurons. State neurons provide to the network the memory of what happened in the past.**

On the other hand, Indium Phosphide allows the integration also of active devices, namely lasers and amplifiers, enabling the realization of monolithically integrated solutions comprising both active and passive devices.

To take the best of the two photonic integration platforms, and to connect them with electronics, a number of approaches have been pursued, recently summarized in [47].

A. TAXONOMY OF CURRENT APPROACHES

In the last few years, the field of photonic neuromorphic computing has seen substantial advancements. Several solutions were developed for various kinds of ANN, both integrated and free-space optic (i.e., where light is not guided in a transmission media) approaches were investigated. In some cases, tailored training methods have also been proposed (see Sec. IV). Before proposing a taxonomy of the existing photonic neural networks, in Fig. 1 we have distinguished neural networks that are stateless (i.e., without memory), like multilayer perceptrons and convolutional neural networks, from the ones which are stateful, such as spiking neural networks and reservoir computing. Furthermore, we have distinguished neural networks having only fully connected layers, like multilayer perceptron, from those that use just local connections to reduce the number of weights, like convolutional neural networks.

In the following we provide a taxonomy of the current literature in the field of PNN, sketched in Fig. 2 and then detailed in Sec. III. Before that, in the next sub-sections we briefly recap the characteristics of the types of neural networks that have actually been implemented/prototyped in photonics.

1) MULTILAYER PERCEPTRONS

These architectures consist of several layers of neurons, where each neuron is fully connected with neurons of the previous and the successive layer. The first layer, receiving the input signals, is called the input layer, while the final layer, providing the inference task results, is the output layer. At least one (shallow nets), but typically several (deep nets) hidden layers are placed between the input and the output layer. In each layer, the information propagates through the neural network via linear combination (i.e., matrix multiplication) of the results of the previous layer with the synaptic weights. Each neuron then performs a nonlinear function on the incoming signal.

2) CONVOLUTIONAL NEURAL NETWORKS

These ANN are designed to process data in the form of multi-dimensional arrays, so they are mainly used for images and videos [1]. Convolutional neural networks (CNN) are composed of multiple layers: the first stages are a series of convolutional layers, followed by pooling layers. The former have the role to detect local conjunctions of features from previous layers, while the latter semantically merge similar features into one. Convolutional layers are connected through a sparse topology, i.e., each neuron is connected to only a small number of neurons of the previous layer. Another characteristic feature of these layers is weight sharing. Between two convolutional layers there is also a nonlinearity, typically a rectified linear unit (ReLU). Finally, to perform the classification, a fully-connected neural network is applied to the output of these layers.

3) SPIKING NEURAL NETWORKS

In these ANN the information is encoded into events, or spikes. This is an hybrid encoding with both analog and digital properties: the time at which a spike occurs is analog, while its amplitude is digital and can thus be restored [27]. The computational primitive, i.e., the spiking neuron, has generally an integrate-and-fire response behaviour. It can be considered a stateful neural network, since its neurons have internal time dynamics.

4) RESERVOIR COMPUTING

These architectures follow a different approach to neuromorphic computing. They are composed of three layers: the input, the reservoir, and the readout layer. The peculiarity is that the reservoir, which should be the most complex section, is a sparsely and randomly connected fixed network of nonlinear neurons. This means that the connectivity complexity is typically largely reduced [24]–[26]. The actual behaviour depends on the input and readout layers, so that a particular reservoir can be used for different applications by changing these layers. Training typically occurs in the readout layer.

III. LITERATURE REVIEW

Here we present a review of the current literature according to the taxonomy described in Sec. II-A and sketched in Fig. 2. Together with the proposed taxonomy, the figure categorizes a number of relevant PNN proof-of-concept implementations, indicating also their hardware design (free space optics vs.

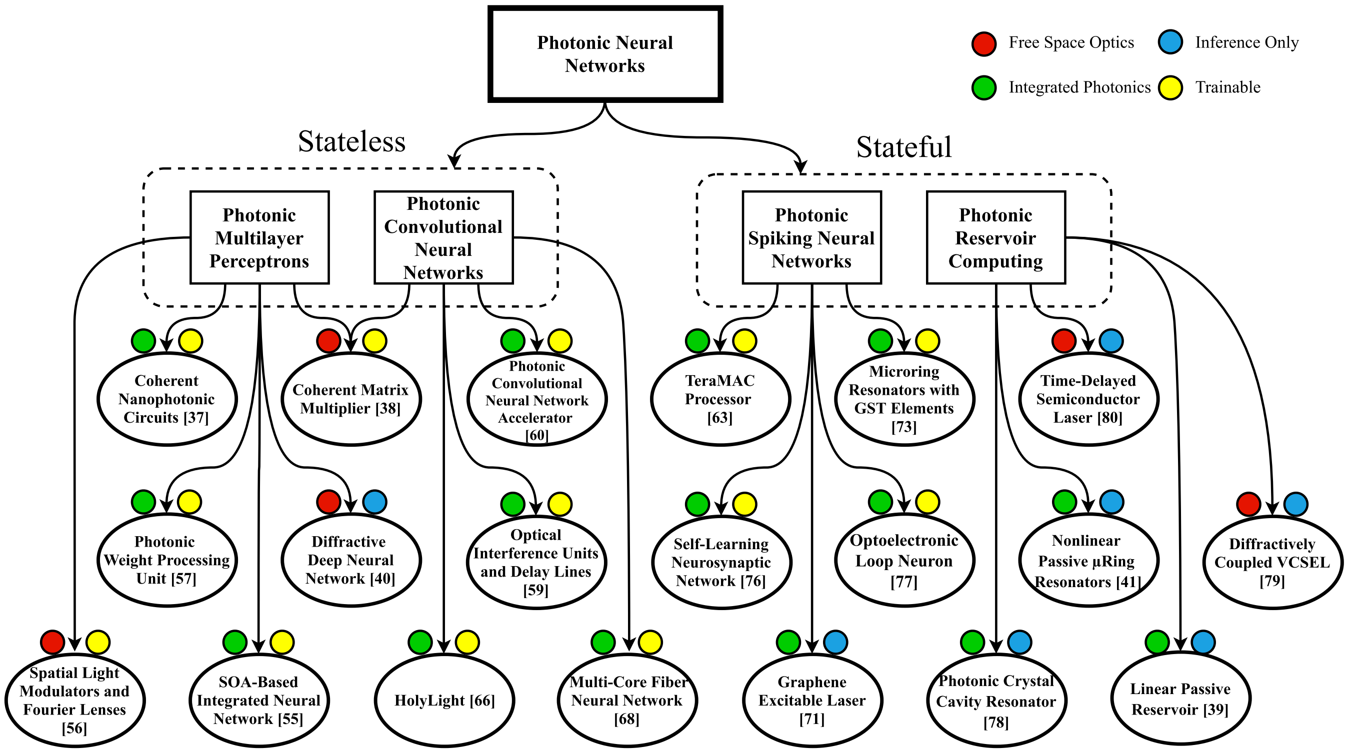


FIGURE 2. Taxonomy of PNN approaches and associated proofs of concept, indicating the hardware implementation (free space optics or integrated photonics) and the operation mode (inference only or trainable). Only the types of neural networks for which a photonic version has been demonstrated in the literature are reported.

integrated photonics) and if they can be trained in optics (a feature that is not always present, being very complex to implement). As the Coherent Matrix Multiplier CNN implementation comprises also fully connected layers, in Fig. 2 it is connected to both Photonic Multilayer Perceptron and to Photonic CNNs.

Each of the following sections describes the most relevant approaches of each category and includes a table summarizing the hardware implementation, the reported results and applications, and the training mechanisms.

A. PHOTONIC MULTILAYER PERCEPTRONS

In [37] a general deep neural network (DNN) is proposed, based on nanophotonic circuits that process coherent light. As shown in Fig. 3, layers of this architecture are composed of two parts: an Optical Interference Unit (OIU) that performs the optical matrix multiplication, and an Optical Nonlinear Unit (ONU) that implements the nonlinear activation function. The OIU consists of a cascaded array of 56 programmable Mach-Zehnder Interferometers (MZIs). The MZIs physically implement the singular value decomposition, which is a factorization of any matrix M with two unitary matrices $U, V^{\dagger}$ and a diagonal rectangular matrix Σ in the form $M = U\Sigma V^{\dagger}$ [49]. Only the OIU has been experimentally demonstrated. The ONU was emulated on a computer converting the signal from the optical to the electrical domain and vice versa. Even though a procedure for on-chip training is proposed (see description in Sec. IV), in the experimental study the training was actually performed on a computer. This implementation was tested on a vowel recognition task reaching an accuracy of 76.7%, and more recently on a MNIST digit recognition task (a widely used classification benchmark), reaching an accuracy of 95% [50] (for MNIST, the best known accuracy using DNNs trained on electronic computers is slightly below 99% [51]).

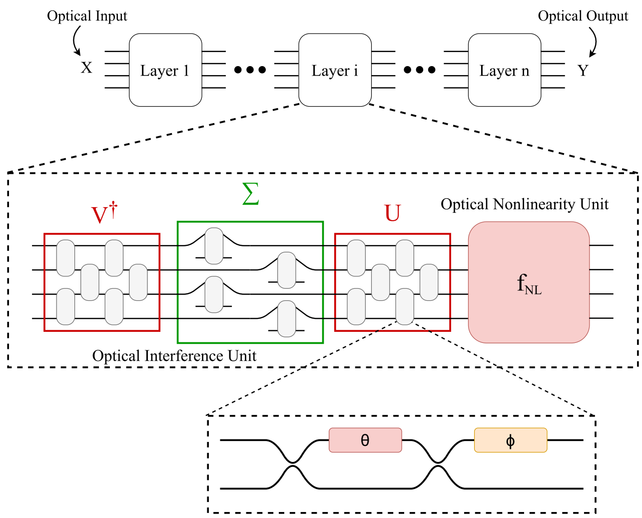


FIGURE 3. Schematic of the PNN proposed in [37]. A layer in this architecture is formed by an Optical Interference Unit (OIU) and an Optical Nonlinear Unit (ONU). The OIU consists of a mesh of programmable MZIs that implement the singular value decomposition, which is a factorization of any matrix M with two unitary matrices $U, V^{\dagger}$ and a diagonal matrix Σ in the form of $M = U\Sigma V^{\dagger}$.

One of the latest works on this MZI-based approach suggests a path towards quantum optical neural networks [52]. The goal is to directly exploit quantum optics features, such as mode mixing and optical nonlinearity, in photonic neural networks.

Back to non-quantum implementations, activity on MZI-based OIUs continued, evaluating trade-offs between expressivity and robustness in universal optical linear multipliers [53]. The study compared two types of networks based on MZI meshes, i.e., GridNet and FFTNet. The former is an universal unitary multiplier in the shape of a grid-like network, the latter is a less expressive (i.e. non truly-universal) but more compact design. The lower number of MZIs leads to reduced overall loss and noise at the expense of waveguide crossings, because components are not in a grid-like structure. Both architectures were compared on the MNIST dataset and, in the case of ideal components, classification accuracies reached 97.8% for GridNet and 94.8% for FFTNet. Fabrication inaccuracies, calibration issues and other non-idealities were then modeled by adding independent zero-mean gaussian noise to the phase of the phase shifters and to the transmittance of the beam splitters. With these non-idealities GridNet accuracy dropped below 50%, while FFTNet maintained near constant performance. This study showed how a more robust network can be greatly preferred over one with higher expressivity.

The use of Semiconductor Optical Amplifiers (SOAs) for the implementation of an Indium Phosphide-based integrated PNN has been recently proposed in [54], [55]. The implemented structure is a 3-layered neural network with four neurons per layer where the linear part, i.e., weighting and sum, is performed by an array of monolithically integrated SOAs. Four input signals are encoded at different wavelengths and multiplexed to form a single wavelength-division multiplexed (WDM) signal. The WDM signal is broadcast to the neurons of the next layer and de-multiplexed through four arrayed waveguide gratings (AWGs). The single neuron is formed by four inputs, four SOAs and a 4:1 multimode interference (MMI) combiner. Different weights are implemented by tuning SOA control currents. This chip has been used to perform a Fisher's Iris flower classification (another classification benchmark), reaching an accuracy of 95% when simulated and of 85.8% when implemented, with the decrease mainly due to distortions in E/O and O/E conversions and to the digital implementation of the nonlinear unit.

An all-optical implementation of a multilayer perceptron, including the nonlinear activation function, has been reported in [56]. The proposed architecture is based on free space optics, encoding signals with the light power. The linear part consists of a spatial light modulator placed at the back focal plane of a Fourier lens, performing the linear summation, while the nonlinear activation function is realized through electromagnetically induced transparency. The article reports the implementation of a two-layer fully-connected PNN for the classification of Ising models. The electromagnetically induced transparency is produced by laser-cooled Rb atoms in a darkline two-dimensional magneto-optical trap.

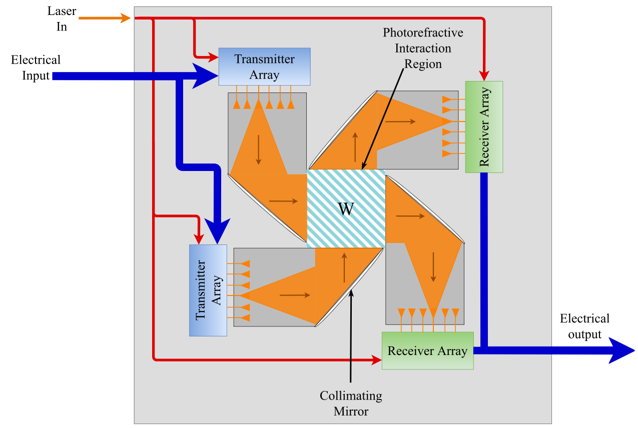


FIGURE 4. ***Photonic Weight Processing Unit*** **reported in [57]. Input electrical vectors are converted to optical signals and sent to the photorefractive interaction region, which performs a matrix-vector multiplication (weighting). Two orthogonal optical transmitters enable the parallel execution of both inference and training. Indeed, when the second input optical beam is sent through the perpendicular direction of the photorefractive interaction region, it will be subjected to the transposed weight matrix: this property is really important for an efficient implementation of the backpropagation algorithm (which requires the multiplication by the transposed matrix of weights).**

A Photonic Weight Processing Unit has been proposed in [57]. This system, depicted in Fig. 4, stores synaptic weights as refractive index gratings in a Gallium Arsenide photorefractive interaction region that forms the weighting matrix W. Gratings are written by two interfering optical beams. During inference, input electrical vectors are converted into optical signals and sent to the photorefractive region that performs the matrix-vector multiplication. The peculiarity of this implementation is that, when the optical inputs are sent through the adjacent side of the photorefractive region, they are subject to the transposed weight matrix W^T. This behaviour enables the parallel execution of both inference and training functionalities by adding a second, orthogonally placed, transmitter/receiver chain.

Finally, in [40] a PNN implementation is proposed and demonstrated using submillimiter waves. The neural network consists of a stack of diffractive surfaces realized by a 3D printing process. The realized prototype works at 0.4 THz. Each point on a layer acts as a neuron, characterized by a complex-valued response. In this case the learning phase is done through a computer, which fixes the network design, while the inference (query) phase is all-optical. This implementation has been subject to further investigations towards: (i) better accuracies (97% in the MNIST digit recognition task), (ii) integration with electronic neural networks to achieve hybrid machine learning and computer vision systems, and (iii) higher frame rates with lower power consumption [58].

B. PHOTONIC CONVOLUTIONAL NEURAL NETWORKS

An all-optical integrated solution for a convolutional neural network (CNN) is proposed in [59]. The studied solution

TABLE 1. Photonic multilayer perceptrons.

Photonic Hardware Implementation	Reported Results and Applications	Training
Coherent Nanophotonic Circuits Neural network layers are composed of an optical interference unit (OIU) and an optical nonlinear unit. The OIU is implemented using a programmable nanophotonic processor composed of 56 programmable MZI, each comprising two phase shifters [37]	Experimental implementation of a two-layer neural network. Reached an accuracy of 76.7% on vowel recognition [37] and later an accuracy of 95% on MNIST dataset [50]. Estimated total power for forward propagation $P \approx N$ mW, where N is the number of nodes per layer	Matrix parameters used in the PNN trained with the standard backpropagation algorithm using a stochastic gradient descent method (on a conventional computer). An on-chip training with forward propagation is also proposed (see Sec. IV)
SOA-based Integrated Neural Network Multilayered fully connected network that relies on WDM signals, every neuron is made of a demultiplexing stage, an SOA per input wavelength that performs weighting, and an MMI-based combiner for the output [54, 55]	Monolithical implementation of the linear part. Fisher's Iris flower classification with an accuracy of 85.8%	-
Spatial Light Modulators on Fourier Lenses Fully-connected neural network in free space optics that exploits spatial light modulators on Fourier lens for the linear part and electromagnetically induced transparency for the nonlinear part [56]	Two layer fully-connected neural network implemented for the Ising model classification	Supervised Learning
Photonic Weight Processing Unit Architecture based on a GaAs photorefractive material that stores weights as refractive index gratings. Processing functions for both inference and training performed in parallel [57]	Weight writing in the GaAs photorefractive material reported	Backpropagation
Diffractive Deep Neural Network The PNN consists of multiple layers of diffractive surfaces. Weights are based on free-space diffraction and determine the coherent interference of the secondary waves that are phase and/or amplitude modulated by the previous layers [40]	Amplitude modulation tested on the digit recognition MNIST dataset: accuracy of 93.39% reached with 5 layers. Phase modulation tested on the fashion MNIST dataset, accuracy of 86.60% reached with 10 layers. Better accuracy performance of 97% for digit recognition and 90% for the fashion MNIST developed in [58]	Training using an error backpropagation method run on a computer. The found parameters are then fixed on the surfaces during fabrication

consists of convolutional layers divided in three parts: an OIU based on MZI that performs linear operations in the kernel matrix, an optical nonlinearity and a tree of 3 dB splitters feeding variable-length delay lines forming the "repatching logic", i.e, a step required to reconstruct data processed by the CNN. The delay lines are tailored such that the outputs from a layer are synchronized in time and form a new patch for input into the next layer's kernel matrix.

In [60] a photonic accelerator for convolutional neural networks is introduced. The building blocks of this system are microring resonator (MRR) banks, proposed in [61]. These MRRs are designed to perform multiply and accumulate (MAC) operations and implement the Broadcast-and-Weight (B&W) protocol [62], [63], which relies on WDM to enable scalable interconnections among laser neuron processors. The B&W protocol is in principle agnostic to the type of neural network as long as it uses multiple wavelengths. The protocol is exploited also in spiking and perceptron-based neural networks [64], and it enables massive parallel communications among neurons. In a multinode system a WDM signal is broadcast to all nodes in parallel and each node filters a unique wavelength. The weighting is obtained by tuning the drop coefficient of a node from 0% to 100%. The accelerator consists of a weighting MRR bank repository, in which microrings tune to convolutional kernel weights. The weights are initially stored in an off-chip DRAM memory and then they are loaded into a kernel weight buffer upon arrival of a new layer request. An execution time improvement of three orders of magnitude over an electronic engine was demonstrated. This work has been further developed in [65] aiming at improved integration and better performance in terms of speed and power consumption. A speed between 2.8 and 14 times faster with respect to GPUs has been estimated, while consuming 0.75 times the energy.

Staying within the same framework, we report HolyLight, a nanophotonic accelerator meant to boost CNN inference throughput in data centers [66]. In HolyLight-M, the first version of this accelerator, the photonic hardware is exploited to perform convolutions. This architecture is based on Photonic Processing Units composed of: (i) a photonic matrix-vector multiplier that relies on an array of microdisk resonators, photodetectors and a wavelength multiplexer to perform parallel computations, (ii) a 16-bit photonic ripple carry adder consisting of 16 microdisk-based 1-bit electro-optic full adders, (iii) a Photonic Binary Shifter consisting of a square mesh of microdisk resonators. The main bottleneck of this implementation was found in the ADCs and DACs required for the matrix-vector multiplier, limiting the speed at 1.28 Gbps and consuming 85.7% of the total power. The subsequent version, called HolyLight-A, avoids the use of ADCs and DACs by recurring to power-of-2 quantized weights. The resulting accuracy degradation is less than 1%. The overall system architecture is very similar to its predecessor, except in the fact that by restricting the weight values only to exact powers of 2 the photonic matrix-vector multiplier has been removed. HolyLight-A performs CNN inference 3.1 times faster than TPU while improving the CNN inference throughput per Watt by 13 times over ISAAC electronic accelerator [67].

A multi-layer hybrid optical-electrical neural network based on an optical coherent matrix multiplier is presented in [38]. The network layers, depicted in Fig. 5, consist of a matrix-vector multiplier and an element-wise nonlinearity. The multiplier combines inputs and weights, both encoded as optical signals, and performs homodyne detection between each signal-weight pair. The resulting electronic signals are subjected to a nonlinear function, serialized, converted back to optical signals and sent to the input of the next layer. Synaptic connections are realized by the quantum multiplication process in homodyne detectors. This optical system can be used for both fully-connected and convolutional layers and it can further allow both inference and training to be performed in the same optical device.

The use of a multi-core optical fiber to realize neural networks has been proposed in [68]. Single cores inside the fiber are used to realize individual neurons, while synaptic interconnections are mimicked by means of optical coupling between the cores. The weighting behaviour is implemented through pump driven amplification in erbium-doped cores. This network can be used together with a MultiMode Fiber (MMF) for image detection as shown in [69]. Light can be preprocessed by a digital micromirror device and projected in the MMF. This projection in the MMF space reduces the number of required nodes, i.e., optical cores. We categorized this solution as CNN due to the presence of sparsely connected layers in the developed feed-forward network.

Finally, the use of diffractive surfaces tailored specifically to realize optical convolutional layers and the consequent improvements in realizing CNNs have been reported in [70].

C. PHOTONIC SPIKING NEURAL NETWORKS

Substantial research works have been conducted in this field for many years, as summarized in [27], initially exploiting ultrafast bulk optical components connected to realize large fiber-based systems. Successive photonic implementations pushed integrability aiming at greater scalability and energy efficiency, hardware cost reduction, and robustness to environmental fluctuations.

A graphene excitable laser was proposed as a spiking neuron [71], the fundamental building block for spike information processing. The embedded graphene layer was used as an optical saturable absorber to perform the nonlinear activation function. The structure was modelled as integrated, but realized as a fiber-based prototype for the demonstration of two applications: temporal pattern recognition and stable recurrence.

A more compact and faster approach was recently proposed in [63], where the neuron is based on a distributed feedback (DFB) laser instantiated on an Indium Phosphide platform. The laser has two photodetectors enabling both inhibitory and excitatory stimuli. This system is compatible with the B&W protocol [62]. The proposed device can process up to 10^{12} MACs/s. The use of semiconductor excitable lasers is a common choice for realizing a spiking neuron: many of these devices have been proposed, and a summary can be found in [72].

A new type of device for a spiking neuron, that can be also interfaced with photonic synapses, was reported in [73]. The proposed solution implements a bipolar integrate-and-fire neuron in which the integration unit consists of two double bus ring resonators with an embedded Phase Change Material (PCM), namely $Ge_2Sb_2Te_5$ (GST), controlling the propagation in the ring. The resonators' output is summed and used to excite the firing unit composed of a photonic amplifier, a circulator and a rectangular waveguide with a GST element on top. This neuron was used to simulate a fully connected network with three layers for an handwritten digit recognition task. This spiking neuron can be integrated with GST-embedded photonic synapses proposed in [74] to form an all-photonic spiking neural network. This work has been further developed with the proposal of a photonic dot product engine (i.e. matrix-vector product) in [75].

Another PCM-based approach to produce an all-optical spiking neurosynaptic network was reported in [76]. The proposed structure, depicted in Fig. 6, is a fully connected network where in each neuron the input spikes are weighted using PCM cells over waveguides (synaptic interconnections) and summed using a wavelength-division multiplexer based on an MRR arrays. The spiking mechanism is enabled by a PCM cell over a ring resonator. Spiking occurs if the integrated power of the post-synaptic spikes exceeds a certain threshold following a nonlinear response resembling the ReLU function. The study not only proposes a multilayered fully connected network with proof of supervised learning, but it also presents and demonstrates a design variation allowing unsupervised learning. This is obtained by overlapping (in time) the output pulses with the input pulses through a feedback line. The synaptic weights of all inputs that contributed to a spike generation are increased, while the others are decreased (implementing a simplified variation of the spike timing-dependent plasticity).

Finally, we report the investigation of superconducting optoelectronic spiking neurons, called Loop Neurons and described in [77] and references therein. These neurons use single photon detectors to detect incoming spikes encoded in single photons. Detection events are integrated and converted in a supercurrent by a Josephson circuit that ultimately stores it in a superconductive loop. The threshold behaviour is performed by a Josephson junction: when the stored current exceeds its crucial current a fluxon is produced, which triggers an amplification sequence. The result of this amplification is light emission from an LED, i.e., the firing event. Synaptic weights are dynamically managed by a parallel circuit that combines another single photon detector and another Josephson junction. A network of these neurons spanning a whole 300 mm Si wafer, i.e., one million neurons and 200 million synapses, would dissipate only 1 W of power.

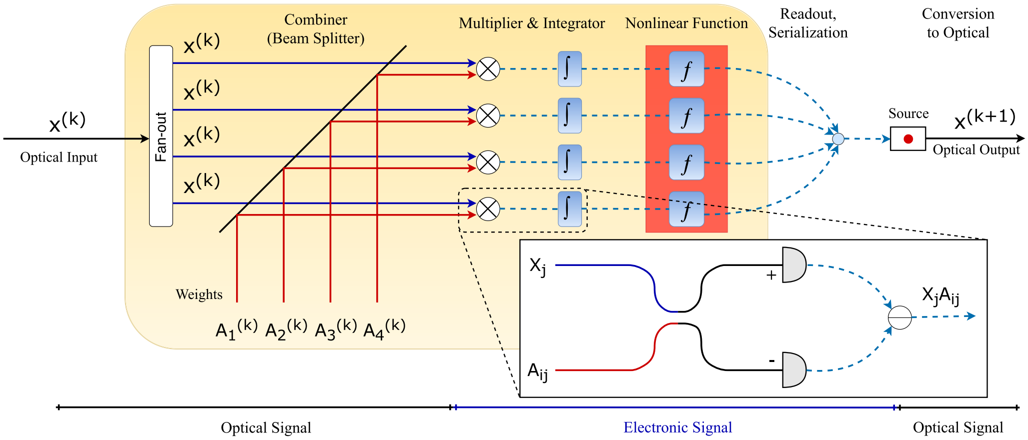


FIGURE 5. **Layer diagram of the architecture based on optical coherent matrix multiplication proposed in [38]. In this solution both inputs and weights are encoded as serial optical signals. Input vectors are replicated and combined with the weights through a multiplier that performs homodyne detection between each signal-weight pair. The nonlinear activation function is thus performed in the electrical domain. Finally, the resulting electrical vectors are serialized, converted to optical signals and sent to the input of the next layer.**

TABLE 2. **Photonic convolutional neural networks.**

Photonic Hardware Implementation	Reported Results and Applications	Training
OIU and Delay Lines CNN with convolutional layers consisting of three parts: 1) MZI-based OIU, 2) optical nonlinearity, 3) tree of 3 dB splitters feeding variable length delay lines as repatching logic [59]	Inference demonstrated. The proposed architecture is supposed to perform million inferences per second at as low power as 2 mJ per inference: it should outperform state-of-the-art ASIC by 30 times in speed while using the same power	-
Photonic Convolutional Neural Network Accelerator (PCNNA) [60] Based on photonic microring resonator weight banks and the Broadcast-and-Weigth protocol [61, 65]	The optical accelerator showed a 3 order of magnitude execution time improvement over electronic engines	-
HolyLight Photonic accelerator for CNNs: the photonic hardware based on 16-bit photonic ripple-carry adders and photonic binary shifters is exploited to perform convolutions [66]	The optimized version HolyLight-A performs CNN inference 3.1 times faster than TPU and improves the CNN inference throughput per Watt by 13 times over ISAAC electronic accelerator	-
Coherent Matrix Multiplier Architecture based on coherent detection where both inputs and weights are encoded in optical signals. The system passively performs general matrix-vector products [38]	Inference simulated. Network used for the MNIST digit recognition dataset. ImageNet considered as a benchmark problem with a reported accuracy of 84.9%. Predicted energy consumption in the tens of fJ for moderately large networks (100s neurons), i.e., 2 to 3 orders of magnitude smaller than state-of-the-art CMOS solutions	Gradient descent and backpropagation
Multi-Core Fiber Neural Network Individual silica cores inside a multi-core fiber exploited as neurons and synapses. Optical signals transferred through optical coupling between the cores. Weighing behaviour achieved through pump driven amplification in erbium-doped cores [68]	Pattern recognition tasks successfully simulated	-

However, the working temperature of 4.3 K and the subsequent need of a cryogenic cooler result to be the main bottleneck of this architecture. To achieve the superconducting behaviour a cryogenic system, consisting of liquid Helium, is required consuming 1 kW without power being dissipated on-chip and 1 kW per on-chip dissipated W.

D. PHOTONIC RESERVOIR COMPUTING

Reservoir computing has attracted a lot of attention in the last decades because of its potential versatility. The reservoir is a highly dynamical system used to perform signal-mixing of the inputs, and the resulting reservoir states are then processed by an application-specific readout. The peculiarity of this approach is that the most complex part of the PNN, i.e., the reservoir, is fixed and can be used for multiple tasks by changing the readout. Two papers provide a good introduction to this topic, surveying different physical realizations of reservoirs (from electronic through photonic to biological) [26], and focusing on the most promising applications of photonic reservoirs [25].

A reservoir computing architecture realized with only passive photonic elements was demonstrated in [39]. It consists of a 16-node square mesh of splitters, waveguides, and optical combiners, resulting in a complex and random interferometer [24]. Being a fully passive reservoir, it is ideal in terms of energy efficiency, but it is an inherently linear

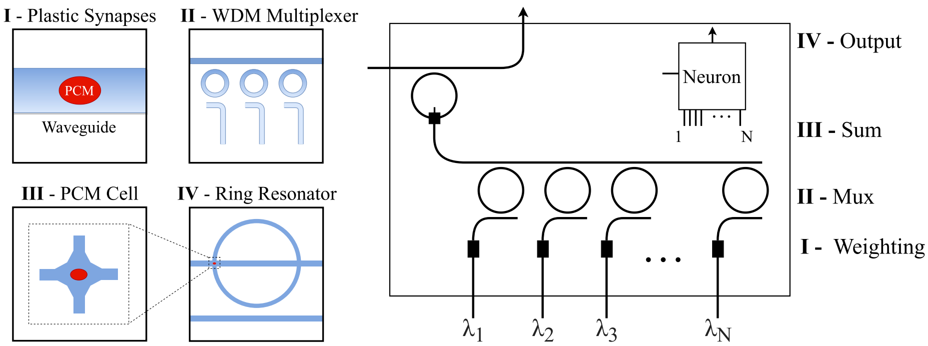


FIGURE 6. Scheme of the Spiking Neuron proposed in [76]. The structure exploit a Phase Change Material (PCM) placed over the waveguides. The input spikes are weighted using PCM cells (I) and summed through a wavelength division multiplexer based on an MRR array (II). The spiking mechanism is enabled by a PCM cell (III) over a ring resonator (IV): if the integrated input power exceeds a certain threshold, a spike occurs with a nonlinear response resembling the ReLU function.

TABLE 3. Photonic spiking neural networks.

Photonic Hardware Implementation	Reported Results and Applications	Training
Graphene Excitable Laser Graphene is used as a saturable absorber. Modelled as integrated, but fabricated as a fiber-based prototype [71]	Two applications reported: temporal pattern recognition and stable recurrence	Synaptic time dependent plasticity proposed for spike-pattern cluster analysis
TeraMAC Processor Processor based on a spiking laser neuron, designed to be compatible with the Broadcast-and-Weight (B&W) protocol [62, 63]	The single device can process $\sim 10^{12}$ MAC/s with an energy efficiency of $\sim$ 270 fJ per MAC operation	-
Microring Resonators with GST Elements GST embedded ring resonator to form a bipolar integrate-fire neuron [73]. GST tapered silicon nitride waveguides as on-chip weighted interconnections [74]	Fully connected ANN with 3 layers for standard handwritten digit recognition task based on the MNIST dataset. Power consumption of 14-19 pJ per write operation	Synaptic weights trained using the backpropagation algorithm
Self-Learning Neurosynaptic Network Input spikes are weighted using PCM cells over waveguides (synapses) and summed using a WDM based on an array of microring resonators. Spiking mechanism is enabled by a PCM cell over a ring resonator [76]	Simple image recognition tasks (binary patterns) under supervised learning demonstrated. Simulation for digit recognition and language identification provided pattern recognition on 15 pixel images with unsupervised learning. Energy needed for a neuron's cycle operation is 1.5 nJ	Supervised and unsupervised learning demonstrated
Optoelectronic Loop Neuron Neurons are constituted by single photon detectors, Josephson junctions and firing LEDs. The firing circuit is an analog photon-to-fluxon transducer, input single-photon signals are converted in a supercurrent that eventually triggers light emission from an LED [77]	Neuron cascadability and fan-out discussed. Synaptic time dependent plasticity reported	The circuit performs Synaptic time dependent plasticity based on temporal correlation between photons from the pre-synaptic and post-synaptic neurons

device. The lack of nonlinearity is compensated in the readout part, exploiting the nonlinear response of a fast photodiode.

Recently a photonic reservoir has been proposed, based on silicon all-pass MRRs as nodes in a 4×4 swirl topology [41]. This reservoir architecture, depicted in Fig. 7, results to be both passive and nonlinear thanks to the ring resonator nodes, with the advantage that the readout can be linear. The network has been numerically tested for a boolean XOR task, training the linear readout using ridge regression.

Another photonic reservoir architecture suitable for integration on a silicon chip has been recently proposed, made of a photonic crystal cavity resonator with a quarter-stadium shape [78]. The cavity consists of holes with a radius of 155 nm etched from a 220 nm silicon slab, and is connected to external waveguides through W1-defects in the photonic crystal cavity walls. Light is sent inside through one W1-defect, while the other waveguides are used for the readout. This structure was numerically tested in a header recognition task and in boolean XOR and AND tasks.

The use of vertical cavity surface emitting lasers (VCSELs) in the field of PNN is mainly investigated as a base for spiking neurons [72]. Nevertheless, a photonic reservoir based on a diffractively coupled VCSEL matrix has also been reported in [79]. This reservoir is based on diffractive imaging. Light emission from the VCSEL matrix crosses a diffractive optical element and is imaged onto a reflective spatial light modulator that controls the network coupling weights. This coupling introduces nonlinear dynamics in the system.

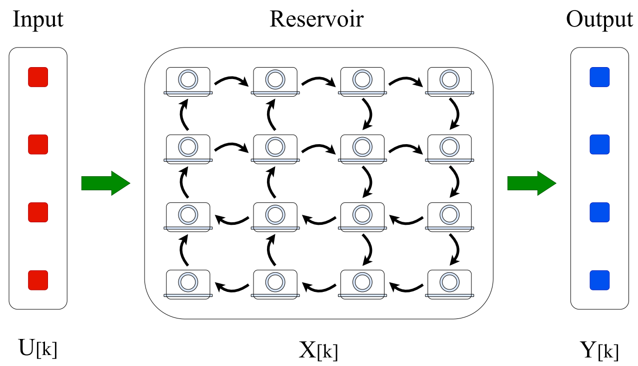


FIGURE 7. **Reservoir based on Nonlinear Passive Microring Resonators proposed in [41]. The reservoir is both passive and nonlinear thanks to the use of MRRs in a 4 × 4 swirl topology. The nonlinear behaviour of the reservoir relaxes the readout complexity.**

Finally, time-delayed reservoir architectures have been proposed, where a more complex recurrent network is emulated with a single nonlinear node by means of a delay line in a closed loop. The realization presented in [80] exploits a semiconductor laser coupled to an optical fiber delay loop to form the reservoir. The loop comprises an optical circulator that defines the propagation direction and suppresses unwanted reflections. Feedback conditions are controlled through a polarization controller and an attenuator.

IV. TRAINING

Training is a nontrivial aspect of neural networks as it does not only influence the network behavior, but it also affects its overall performance. In supervised learning, the training procedure makes use of an objective function that measures the distance (or error) between the desired output and the actual one. This function is used to adjust the neural network internal parameters, i.e., the weights in synapses. In order to minimize the objective function, a gradient vector is computed, to assess the way the error is affected by the change of any weight [1].

Whenever there is a change in the nature of the data being treated by the network, it needs to be retrained. This is called learning. This retraining can be done gradually as the network performs the inference (online learning), or it can be done by a separate machine that adjusts the network according to a new batch of training data (offline learning).

Since the training involves gradient computation or even more complex computations, it is a resource and time consuming phase. On the contrary, the inference (i.e., querying the neural network) is a much simpler process, since the weights are assumed known at this stage. For this reason, many PNN implementations support only the inference phase, and the weights are obtained by using software implementations. Some implementations cannot be trained at all, like the deep diffractive neural network presented in [40]. These architectures are very fast and power efficient, but application-specific, since their weights are frozen in hardware during fabrication. However, training in the electrical domain has two main drawbacks: (i) a dependence of the physical system on the model accuracy is added, and (ii) the speed and power efficiency improvements associated with the use of a photonic architecture are likely to be lost [81].

To take full advantage of photonic technology, the training, although complex, should be specifically tailored for optical architectures. Below the most relevant training methods proposed for PNN are reviewed and summarized in Table 5.

A. ON-CHIP FORWARD PROPAGATION

In [37], together with the coherent nanophotonic circuit, a on-chip training method tailored for this PNN was presented, even though the neural network used in the experimental demonstration was trained on a conventional computer.

This PNN is suitable for obtaining every parameter's gradient by forward propagation and finite difference method: two forward propagations are carried out on-chip to relate the cost function to a parameter's actual value and a perturbed version of it. The gradient associated with that parameter is then calculated by finite difference.

This training was proven through simulations. Thanks to the on-chip implementation, this method follows the PNN forward propagation speed and its power consumption scales with the number of distinct parameters, instead of the number of total parameters as in a conventional backpropagation [37].

B. IN SITU BACKPROPAGATION AND GRADIENT MEASUREMENT

The method, proposed in [81], develops backpropagation on a physical realization of the adjoint variable method, which is typically implemented for the optimization and inverse design of photonic structures. The procedure implements this method by physically propagating the adjoint field and interfering its time-reversed copy with the original field. Gradient terms, expressed as the solution to an electromagnetic adjoint problem, can be directly retrieved through in situ intensity measurement.

In [81], the discussion focuses on the well-known coherent nanophotonic circuit [37], that is used as the basis for a simulation of the training procedure on a simple XOR problem.

The most significant aspect of this method is that it has been developed starting from Maxwell's equations and not over a particular network, so it can be potentially extended to different photonic platforms. The method scales in constant time with respect to the number of parameters, allowing backpropagation to be efficiently implemented in a hybrid optoelectronic network. The main limitation of the procedure is that it is exact only in the limit of a lossless, feed-forward and reciprocal system. Uniform losses are not an issue, but they must be taken into account by adding a step for scaling the measured gradients.

C. TRAINING PHOTONIC CONVOLUTIONAL NEURAL NETWORK WITH SINUSOIDAL ACTIVATIONS

The problem of training actual hardware realizations of PNN is discussed in [82]. The chosen basis architecture for a fully

TABLE 4. Photonic reservoir computing.

Photonic Hardware Implementation	Reported Results and Applications	Training
Linear Passive Reservoir Passive silicon photonic chip containing only waveguides, splitters and combiners as reservoir in a swirl topology with 16 nodes. The nonlinearity is implemented in the readout (electronically and offline) [39]	Three applications demonstrated: 2-bit XOR task, header recognition and classification of spoken digits (the latter not experimentally tested). Reached speed of 12.5 Gbps for the above-mentioned tasks	-
Nonlinear Passive Microring Resonators Microring resonators as nodes in a 4×4 swirl topology reservoir [41]	Performance simulated on the delayed XOR task, evaluating the impact of data rate, injected power, and optical detuning. For a BER $< 10^{-3}$ a data rate of 20 Gbps with a power consumption of 2.4 mW is reported	Readout trained with ridge regression
Photonic Crystal Cavity Resonator Resonator in the shape of a quarter-stadium resonator as a reservoir [78]	Performance simulated on several tasks: boolean XOR & AND, header recognition up to 6-bit headers, working for a very wide range of bit rates. Simulations suggest that the system should work at a speed of the order of 50 Gbps	Readout trained with ridge regression and linear discriminant analysis
Diffractively Coupled VCSEL Optical emission from a VCSEL matrix reaches a spatial light modulator located in the imaging plane to control the network's coupling weights [79]	Multiple nonlinear functions were implemented. Global network state updated at a rate of 0.5 GHz (reservoir dynamical timescale is set by the external coupling delay time).	-
Time-Delayed Semiconductor Laser Semiconductor laser exploited as nonlinear node in a time-delayed reservoir. Feedback conditions in delay loop controlled through a polarization controller and an attenuator [80]	System trained to predict the chaotic Mackey-Glass time series. Reported the effect of frequency detuning between the injection laser and the reservoir laser	-

optical neural network is a variation of the one proposed in [61], where positive and negative weights are implemented through couples of WDM signals, and the activation function is sinusoidal (which should ensure the highest expressivity).

At first, the problem of PNN initialization, i.e., the choice of initial weights, is addressed. These values are derived from the constraint of constant variance of input signals through the various layers, in order to allow the information to arrive to the employed training error function and to ensure the smooth flow of gradients.

The second tackled issue is the bounding of the neurons' response, i.e., the response of the activation function, since physical systems work better within a specific power range. This problem is addressed together with the weight regularization, a method used in ''classical'' contexts to reduce overfitting phenomena, here employed with the aim to impose constraints dictated by the PNN hardware.

The third non-ideality discussed is Noise-Aware Training, which is of paramount importance when dealing with physical implementations. The developed training method aims to significantly increase noise robustness by allowing the network to adapt to the actual characteristics of the hardware implementation.

Finally, the proposed methods have been tested using three different CNNs in four different recognition problems: MNIST dataset, MNIST fashion dataset, CIFAR10 dataset and the FI-2010 dataset. The methods have been separately evaluated and finally a proof of concept has demonstrated that they can also tackle all aforementioned issues (activation limits, weights limits, and noise) at the same time.

D. NONLINEARITY INVERSION

In photonic reservoir computing, training concerns the readout layer. Recent research in photonic reservoir mainly focused on reservoir layer development, with several structures proposed (see Sec. III-D). Nevertheless the readout layer is of fundamental importance because it defines the network behavior and, unlike the reservoir, it has to be properly trained. Training and signal mixing in the readout has been so far developed in the electrical domain [39], but this results in a limitation of the speed and of the power consumption reduction achievable with an all-optical reservoir.

Implementing an all-optical readout requires only a single photodetector that receives the weighted sum of all optical signals. This approach has a relevant drawback: the direct observability of the photonic reservoir states is lost. Observation of the internal states is required in classical linear readout training algorithms, such as ridge regression.

In [83] a training procedure, called nonlinearity inversion, is proposed, that overcomes the observability problem of an all-optical readout. The method solves this issue by estimating the amplitude and phase of the reservoir states through a single photodetector (the absolute phase is lost, but relative phases can be obtained). Complex states are observed by appropriately setting the readout weights while iterating over a predefined input sequence. The training procedure has been numerically tested on a 3-bit header recognition task over a wide range of input signal bit rates. The baseline approach still exhibits a slightly better task performance, but it needs a more complex hardware implementation and is applicable in a smaller range of bit rates.

TABLE 5. Training methods for PNN.

Name	Training Methodology	Application
On-chip training with forward propagation	The PNN based on coherent nanophotonic circuit [37] can be used to obtain the gradient for each distinct parameter using forward propagation and finite difference method	Performed full optical simulation of the on-chip training for the vowel recognition problem. The execution time scales with the number of distinct parameters instead of total parameters, as in conventional backpropagation
In situ backpropagation and gradient measurement	The procedure implements the adjoint variable method by physically propagating the adjoint field and interfering its time-reversed copy with the original field. Gradient terms, expressed as the solution to an electromagnetic adjoint problem, can be directly retrieved through in situ intensity measurement. The method scales constantly in time with respect to the number of parameters [81]	The method is applicable to different photonic systems. Demonstrated on the PNN based on coherent nanophotonic circuit [37]. The procedure has been used to numerically train the network for a XOR gate task
Training Photonic CNNs with Sinusoidal Activations	The work in [82] focuses on training procedures aware of photonic hardware implementations concerning: initialization, activation functions limits, weight limits and noise robustness.	Four recognition tasks (MNIST, MNIST fashion, CIFAR10, and FI-2010 datasets) on three different CNNs implementations
Nonlinearity inversion	The method solves the oservability issue of reservoir states by estimating their amplitude and phase through a single photodetector. Complex states of the reservoir are observed by appropriately setting the readout weights, while iterating over a predefined input sequence [83]	Training procedure for passive photonic reservoirs with an optical readout. Simulated readout trained to perform a 3-bit header recognition task

V. DISCUSSIONS: THE LONG MARCH TOWARDS TRULY DEEP PNN

From the presented survey, it becomes apparent that the challenge of developing truly deep neural networks with photonics is still open. Photonic multilayer perceptrons and photonic spiking neural networks seem to have a lot of potential towards the realization of all-optical ANNs. While waiting for these long-term breakthroughs, photonic accelerators for CNNs seems to be, in the short term, the most promising photonic solutions in order to enhance inference speed and to reduce power consumption.

Both prominent companies [53], [57] and fast-growing start-ups [50], [84]–[86] are actively working in the field, which indicates a growing industrial interest and possibly practical applications in the foreseeable future.

However, there are still many under-investigated opportunities to improve the implementation of PNN. Furthermore, some types of deep neural networks (Long-Short-Term Memory Neural Networks, Generative Adversarial Nets, Geometric Deep Neural Networks, Deep Belief Networks, Deep Boltzmann Machines, etc.) have not yet been implemented using photonics.

In fact, more research is needed to assess whether a specific type of deep neural network can be implemented optically in an efficient manner, i.e., in a way that provides advantages with respect to fully electronic implementations. For instance, an open question is whether or not this is the case with respect to Hyperdimensional Learning (HL), proposed in [87], a really promising approach to neural networks, which is still in its infancy. Here the problem of a photonic implementation mainly resides in the very large size of the internal representation of objects used in HL.

Regarding the actual realization of PNN, the ultimate goal is to demonstrate large networks with thousands of nodes and interconnections across many hidden layers, i.e., truly deep architectures. With this in mind, it is evident how essential are the PNN cascadability (enabled by low propagation losses, crosstalk, and noise [54]) and robustness not only to fabrication imperfections, but also to parameter drifts over time [29]. Resonant structures like microring resonators are particularly sensitive to manufacturing deviations: this issue needs to be properly addressed in relation to their use in PNN, as discussed in [88]. On the other hand, linear optical processors based on MZI appear to be more robust to process inaccuracies, thanks also to their reconfigurability: some studies discuss how to achieve reliable photonic computations even with imperfect components [89], [90].

Furthermore, the photonic implementation of the nonlinear activation function is a relevant aspect that still requires adequate investigation. Indeed many demonstrations still emulate the nonlinearities in software, as the integration of nonlinear elements is still challenging. Several approaches to address this issue have been reported, including devices based on: MZIs [91], [92], graphene and quantum well electro-optic absorption modulators [93], and photonic crystals [94].

Some technological breakthroughs would be beneficial to PNN, in particular the implementation of an integrable, non-volatile and energy-efficient photonic memory element. In this scenario the use of Phase Change Materials (PCM) seems the most promising approach in order to achieve such photonic memories, since they have also shown the potentiality for multi-level storage [95]. PCM cells have been recently exploited in PNN, mainly for spiking neural networks [73], [74], [76].

VI. CONCLUSION

In this work we have reviewed the most relevant implementations of ANN with photonic hardware. In particular, we have proposed a taxonomy of both stateless and stateful PNN, allowing to group the demonstrated architectures and highlight similarities, technological issues and trends. We have then reviewed specific approaches for the training of PNN, which is the most challenging task to accomplish even in photonics.

The progress in the field of PNN is indeed remarkable, especially in the latest years, and new promising research directions are emerging, with potentially disruptive impact on developing faster and/or less energy-hungry PNN implementations. Many challenges are still to be overcome: many diverse material platforms are being investigated without a clear winner, the maturity of the developed hardware must improve (in many demonstrators critical elements are still emulated in electronics), the scalability and the robustness of the proposed PNN architectures should grow in order to meet the requirements of typical machine learning problems.

Nevertheless the growing research efforts both from academia, prominent companies, and fast-growing start-ups indicate that the coming years will witness a huge expansion of this field, with strong potentials to address real-world applications.

ACKNOWLEDGMENT

N. Andriolli was with the Scuola Superiore Sant'Anna when this work was carried out.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, Jan. 2016.

[3] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[4] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, "Convolutional networks for fast, energy-efficient neuromorphic computing," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 41, pp. 11441–11446, 2016.

[5] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, nos. 1–3, pp. 239–255, 2010.

[6] (Aug. 2017). *Intel Delivers 'Real-Time AI' in Microsoft's New Accelerated Deep Learning Platform*. [Online]. Available: https://newsroom.intel.com/news/intel-delivers-real-time-ai-microsofts-accelerated-deep-learning-platform/

[7] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, Oct. 2016.

[8] *Intel Nervana Neural Network Processor*. Accessed: Dec. 7, 2019. [Online]. Available: https://www.intel.ai/nervana-nnp/

[9] H. Jeong and L. Shi, "Memristor devices for neural networks," *J. Phys. D, Appl. Phys.*, vol. 52, no. 2, 2018, Art. no. 023003.

[10] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature Commun.*, vol. 9, no. 1, p. 2385, Jun. 2018.

[11] *QUEFORMAL, Quantum Engineering for Machine Learning*. Accessed: Dec. 7, 2019. [Online]. Available: https://www.queformal.eu

[12] E. Agrell, M. Karlsson, A. R. Chraplyvy, D. J. Richardson, P. M. Krummrich, P. Winzer, K. Roberts, J. K. Fischer, S. J. Savory, B. J. Eggleton, M. Secondini, F. R. Kschischang, A. Lord, J. Prat, I. Tomkos, J. E. Bowers, S. Srinivasan, M. Brandt-Pearce, and N. Gisin, "Roadmap of optical communications," *J. Opt.*, vol. 18, no. 6, 2016, Art. no. 063002.

[13] Y. Abu-Mostafa and D. Psaltis, "Optical neural computers," *Sci. Amer.*, vol. 256, no. 3, pp. 88–95, Mar. 1987.

[14] H. M. Stoll and L.-S. Lee, "A continuous-time optical neural network," in *Proc. IEEE Int. Conf. Neural Netw.*, Jul. 1988, pp. 373–384.

[15] S. Jutamulia and F. Yu, "Overview of hybrid optical neural networks," *Opt. Laser Technol.*, vol. 28, no. 2, pp. 59–72, 1996.

[16] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, pp. 58–61, 1994.

[17] Europractice. *Imec's Si-Photonics iSiPP50G*. Accessed: Dec. 7, 2019. [Online]. Available: http://europractice-ic.com/mpw-prototyping/siphotonics/imec/

[18] F. Ramos, E. Kehayas, J. M. Martinez, R. Clavero, J. Marti, L. Stampoulidis, D. Tsiokos, H. Avramopoulos, J. Zhang, P. V. Holm-Nielsen, N. Chi, P. Jeppesen, N. Yan, I. T. Monroy, A. M. J. Koonen, M. T. Hill, Y. Liu, H. J. S. Dorren, R. Van Caenegem, D. Colle, M. Pickavet, and B. Riposati, "IST-LASAGNE: Towards all-optical label swapping employing optical logic gates and optical flip-flops," *J. Lightw. Technol.*, vol. 23, no. 10, pp. 2993–3011, Oct. 2005.

[19] F. Bontempi, S. Pinna, N. Andriolli, A. Bogoni, X. J. M. Leijtens, J. Bolk, and G. Contestabile, "Multifunctional current-controlled InP photonic integrated delay interferometer," *IEEE J. Quantum Electron.*, vol. 48, no. 11, pp. 1453–1461, Nov. 2012.

[20] J.-Y. Kim, J.-M. Kang, T.-Y. Kim, and S.-K. Han, "All-optical multiple logic gates with XOR, NOR, OR, and NAND functions using parallel SOA-MZI structures: Theory and experiment," *J. Lightw. Technol.*, vol. 24, no. 9, pp. 3392–3399, Sep. 2006.

[21] Q. Bao, H. Zhang, Z. Ni, Y. Wang, L. Polavarapu, Z. Shen, Q.-H. Xu, D. Tang, and K. P. Loh, "Monolayer graphene as a saturable absorber in a mode-locked laser," *Nano Res.*, vol. 4, no. 3, pp. 297–307, 2011.

[22] K. Nozaki, T. Tanabe, A. Shinya, S. Matsuo, T. Sato, H. Taniyama, and M. Notomi, "Sub-femtojoule all-optical switching using a photonic-crystal nanocavity," *Nature Photon.*, vol. 4, Jul. 2010.

[23] S. M. Hendrickson, A. C. Foster, R. M. Camacho, and B. D. Clader, "Integrated nonlinear photonics: Emerging applications and ongoing challenges [Invited]," *J. Opt. Soc. Amer. B, Opt. Phys.*, vol. 31, no. 12, pp. 3193–3203, Dec. 2014.

[24] G. van der Sande, D. Brunner, and M. C. Soriano, "Advances in photonic reservoir computing," *Nanophotonics*, vol. 6, no. 3, pp. 561–576, 2017.

[25] A. Katumba, M. Freiberger, F. Laporte, A. Lugnan, S. Sackesyn, C. Ma, J. Dambre, and P. Bienstman, "Neuromorphic computing based on silicon photonics and reservoir computing," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 8300310.

[26] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, "Recent advances in physical reservoir computing: A review," *Neural Netw.*, to be published.

[27] A. N. Tait, M. A. Nahmias, Y. Tian, B. J. Shastri, and P. R. Prucnal, "Photonic neuromorphic signal processing and computing," in *Nanophotonic Information Physics*. Berlin, Germany: Springer, 2014, pp. 183–222.

[28] Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," *Light Sci. Appl.*, vol. 8, no. 42, pp. 1–14, 2019.

[29] T. F. de Lima, H. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, "Machine learning with neuromorphic photonics," *J. Lightw. Technol.*, vol. 37, no. 5, pp. 1515–1534, Mar. 1, 2019.

[30] M. M. Waldrop, "More than moore," *Nature*, vol. 530, no. 7589, pp. 144–148, 2016.

[31] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[32] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 1, 2017.

[33] A. K. Maan, D. A. Jayadevi, and A. P. James, "A survey of memristive threshold logic circuits," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1734–1746, Aug. 2017.

[34] S. Sun, J. Li, Z. Li, H. Liu, Q. Li, and H. Xu, "Low-consumption neuromorphic memristor architecture based on convolutional neural networks," in *Proc. Internat. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.
[35] P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, and W. D. Lu, "Sparse coding with memristor networks," *Nature Nanotechnol.*, vol. 12, no. 8, p. 784, 2017.
[36] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *Proc. IEEE/ACM Internat. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2014, pp. 63–70.
[37] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, pp. 441–446, Jun. 2017.
[38] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, vol. 9, no. 2, 2019, Art. no. 021032.
[39] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature Commun.*, vol. 5, p. 3541, Mar. 2014.
[40] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, to be published.
[41] F. Denis-Le Coarer, M. Sciamanna, A. Katumba, M. Freiberger, J. Dambre, P. Bienstman, and D. Rontani, "All-optical reservoir computing on a photonic chip using silicon-based ring resonators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 7600108.
[42] H. J. Caulfield, J. Kinser, and S. K. Rogers, "Optical neural networks," *Proc. IEEE*, vol. 77, no. 10, pp. 1573–1583, Oct. 1989.
[43] D. Casasent, "Optical processing in neural networks," *IEEE Expert*, vol. 7, no. 5, pp. 55–61, Oct. 1992.
[44] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[45] L. A. Coldren, S. C. Nicholes, L. Johansson, S. Ristic, R. S. Guzzon, E. J. Norberg, and U. Krishnamachari, "High Performance InP-Based Photonic ICs-A Tutorial," *J. Lightw. Technol.*, vol. 29, no. 4, pp. 554–570, 2011.
[46] M. Smit *et al.*, "An introduction to InP-based generic integration technology," *Semicond. Sci. Technol.*, vol. 29, no. 8, 2014, Art. no. 083001.
[47] M. Smit, K. Williams, and J. van der Tol, "Past, present, and future of InP-based photonic integration," *APL Photon.*, vol. 4, no. 5, 2019, Art. no. 050901.
[48] A. Y. Liu and J. Bowers, "Photonic integration with epitaxial III–V on silicon," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 6000412.
[49] N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. Baehr-Jones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing, and D. Englund, "Linear programmable nanophotonic processors," *Optica*, vol. 5, no. 12, pp. 1623–1631, Dec. 2018.
[50] Y. Shen and Y. Bai, "Statistical computing with integrated photonics system," in *Proc. 24th OptoElectron. Commun. Conf. (OECC) Int. Conf. Photon. Switching Comput. (PSC)*, Jul. 2019, p. 1.
[51] M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "A fast approximation of the hyperbolic tangent when using posit numbers and its application to deep neural networks," in *Proc. Conf. Appl. Electron. Pervading Ind., Environ. Soc. (ApplePies)*, Sep. 2019, pp. 1–8.
[52] G. R. Steinbrecher, J. P. Olson, D. Englund, and J. Carolan, "Quantum optical neural networks," *NPJ Quantum Inf.*, vol. 5, no. 1, p. 60, 2019.
[53] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," *Opt. Express*, vol. 27, no. 10, pp. 14009–14029, 2019.
[54] B. Shi, N. Calabretta, D. Bunandar, D. Englund, and R. Stabile, "WDM weighted sum in an 8×8 SOA-based InP cross-connect for photonic deep neural networks," in *Proc. Photon. Switching Comput. (PSC)*, 2018, pp. 1–3.
[55] B. Shi, N. Calabretta, and R. Stabile, "Image classification with a 3-layer SOA-based photonic integrated neural network," in *Proc. 24th OptoElectron. Commun. Conf. (OECC) Int. Conf. Photon. Switching Comput. (PSC)*, Jul. 2019, pp. 1–3.
[56] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," *Optica*, vol. 6, no. 9, pp. 1132–1137, Sep. 2019.
[57] A. Abel, F. Horst, P. Stark, R. Dangel, F. Eltes, Y. Baumgartner, J. Fompeyrine, and B. Offrein, "Silicon photonics integration technologies for future computing systems," in *Proc. 24th OptoElectron. Commun. Conf. (OECC) Int. Conf. Photon. Switching Comput. (PSC)*, Jul. 2019, pp. 1–3.
[58] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 3700114.
[59] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-chip optical convolutional neural networks," 2018, *arXiv:1808.03303*. Accessed: Dec. 7, 2019. [Online]. Available: https://arxiv.org/abs/1808.03303
[60] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, "Pcnna: A photonic convolutional neural network accelerator," in *Proc. 31st IEEE Int. Syst. Chip Conf. (SOCC)*, Sep. 2018, pp. 169–173.
[61] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 7430, p. 8, 2017.
[62] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightw. Technol.*, vol. 32, no. 21, pp. 3427–3439, Nov. 1, 2014.
[63] M. A. Nahmias, H.-T. Peng, T. F. de Lima, C. Huang, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "A TeraMAC neuromorphic photonic processor," in *Proc. IEEE Photon. Conf. (IPC)*, Sep. 2018, pp. 1–2.
[64] A. N. Tait, T. F. de Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Silicon photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, 2019, Art. no. 064043.
[65] V. Bangari, B. A. Marquez, H. B. Miller, A. N. Tait, M. A. Nahmias, T. F. de Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," 2019, *arXiv:1907.01525*. Accessed: Dec. 7, 2019. [Online]. Available: https://arxiv.org/abs/1907.01525
[66] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "HolyLight: A nanophotonic accelerator for deep learning in data centers," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 1483–1488.
[67] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comp. Archit. (ISCA)*, 2016, pp. 14–26.
[68] E. Cohen, D. Malka, A. Shemer, A. Shahmoon, Z. Zalevsky, and M. London, "Neural networks within multi-core optic fibers," *Sci. Rep.*, vol. 6, 2016, Art. no. 029080.
[69] N. Shabairou, E. Cohen, O. Wagner, D. Malka, and Z. Zalevsky, "Color image identification and reconstruction using artificial neural networks on multimode fiber images: Towards an all-optical design," *Opt. Lett.*, vol. 43, no. 22, pp. 5603–5606, 2018.
[70] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Sci. Rep.*, vol. 8, no. 1, Aug. 2018, Art. no. 012324.
[71] B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," *Sci. Rep.*, vol. 6, Jan. 2016, Art. no. 019126.
[72] P. R. Prucnal, B. J. Shastri, T. F. de Lima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," *Adv. Opt. Photon.*, vol. 8, no. 2, pp. 228–299, May 2016.
[73] I. Chakraborty, G. Saha, A. Sengupta, and K. Roy, "Toward fast neural computing using all-photonic phase change spiking neurons," *Scient. Rep.*, vol. 8, no. 1, 2018, Art. no. 012980.
[74] Z. Cheng, C. Ríos, W. H. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," *Sci. Adv.*, vol. 3, no. 9, 2017, Art. no. e1700160.
[75] I. Chakraborty, G. Saha, and K. Roy, "Photonic in-memory computing primitive for spiking neural networks using phase-change materials," *Phys. Rev. Appl.*, vol. 11, no. 1, 2019, Art. no. 014063.
[76] J. Feldmann, N. Youngblood, C. Wright, H. Bhaskaran, and W. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, p. 208, 2019.
[77] J. M. Shainline, S. M. Buckley, A. N. McCaughan, J. Chiles, A. Jafari-Salim, R. P. Mirin, and S. W. Nam, "Circuit designs for superconducting optoelectronic loop neurons," *J. Appl. Phys.*, vol. 124, no. 15, 2018, Art. no. 152130.

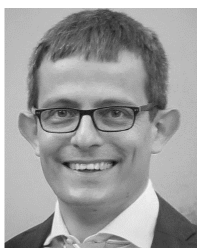
[78] F. Laporte, A. Katumba, J. Dambre, and P. Bienstman, ''Numerical demonstration of neuromorphic computing with photonic crystal cavities,'' *Opt. Express*, vol. 26, no. 7, pp. 7955–7964, 2018.
[79] D. Brunner and I. Fischer, ''Reconfigurable semiconductor laser networks based on diffractive coupling,'' *Opt. Lett.*, vol. 40, no. 16, pp. 3854–3857, 2015.
[80] J. Bueno, D. Brunner, M. C. Soriano, and L. Fischer, ''Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback,'' *Opt. Express*, vol. 25, no. 3, pp. 2401–2412, Feb. 2017.
[81] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, ''Training of photonic neural networks through *in situ* backpropagation and gradient measurement,'' *Optica*, vol. 5, no. 7, pp. 864–871, Jul. 2018.
[82] N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, ''Training deep photonic convolutional neural networks with sinusoidal activations,'' *IEEE Trans. Emerg. Topics Comput. Intell.*, to be published.
[83] M. Freiberger, A. Katumba, P. Bienstman, and J. Dambre, ''Training passive photonic reservoirs with integrated optical readout,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1943–1953, Jul. 2019.
[84] *Lightelligence*. Accessed: Dec. 7, 2019. [Online]. Available: https://www.lightelligence.ai/
[85] *Lightmatter*. Accessed: Dec. 7, 2019. [Online]. Available: https://lightmatter.co/
[86] *Luminous Computing*. Accessed: Dec. 7, 2019. [Online]. Available: https://www.luminouscomputing.com/
[87] P. Kanerva, ''Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors,'' *Cognit. Comput.*, vol. 1, no. 2, pp. 139–159, Oct. 2009.
[88] A. N. Tait, A. X. Wu, T. F. De Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, ''Microring weight bank designs with improved channel density and tolerance,'' in *Proc. IEEE Photon. Conf. (IPC)*, Oct. 2017, pp. 101–102.
[89] D. A. Miller, ''Perfect optics with imperfect components,'' *Optica*, vol. 2, no. 8, pp. 747–750, 2015.
[90] R. Burgwal, W. R. Clements, D. H. Smith, J. C. Gates, W. S. Kolthammer, J. J. Renema, and I. A. Walmsley, ''Using an imperfect photonic network to implement random unitaries,'' *Opt. Express*, vol. 25, no. 23, pp. 28236–28245, 2017.
[91] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, ''An all-optical neuron with sigmoid activation function,'' *Opt. Express*, vol. 27, no. 7, pp. 9620–9630, 2019.
[92] G. Cong, M. Okano, Y. Maegami, M. Ohno, N. Yamamoto, and K. Yamada, ''Method to generate sigmoid-like function in silicon photonic devices towards applications in photonic neutral network,'' in *Proc. IEEE 15th Int. Conf. Group IV Photon. (GFP)*, Aug. 2018, pp. 1–2.
[93] J. George, A. Mehrabian, R. Amin, P. R. Prucnal, T. El-Ghazawi, and V. J. Sorger, ''Neural network activation functions with electro-optic absorption modulators,'' in *Proc. IEEE Internat. Conf. Rebooting Comput. (ICRC)*, Nov. 2018, pp. 1–5.
[94] H. Sharifi, S. M. Hamidi, and K. Navi, ''A new design procedure for all-optical photonic crystal logic gates and functions based on threshold logic,'' *Opt. Commun.*, vol. 370, pp. 231–238, Jul. 2016.
[95] M. Wuttig, H. Bhaskaran, and T. Taubner, ''Phase-change materials for non-volatile photonic applications,'' *Nature Photon.*, vol. 11, no. 8, p. 465, 2017.

LORENZO DE MARINIS was born in Pisa, Italy, in 1993. He received the B.S. and M.S. degrees *(magna cum laude)* in electronic engineering from the University of Pisa, in 2017 and 2019, respectively. He was a Research Scholar with the Scuola Superiore Sant'Anna, Pisa, from May to September 2019. He is currently a Ph.D. student with the Scuola Superiore Sant'Anna, working on photonic integrated circuit design for neuromorphic computing.

MARCO COCOCCIONI received the Laurea degree *(magna cum laude)* from the University of Pisa, in 2000, and the Diploma degree *(magna cum laude)* in computer engineering from Scuola Superiore S. Anna, in 2001, and the Ph.D. degree in computer engineering from the University of Pisa, in 2004. From 2010 to 2011, he spent two years as Senior Visiting Scientist with the NATO Undersea Research Centre (now CMRE), La Spezia, Italy. For his collaboration with CMRE, he obtained the NATO Scientific Achievement Award, in 2014. Since 2016, he has been an Associate Professor with the Department of Information Engineering, University of Pisa. His main research interests include artificial intelligence, decision support systems, fuzzy logic, multiobjective evolutionary optimization, genetic fuzzy systems, pattern recognition, neural networks, and operations research. He has been the General Chair of three IEEE international conferences. In addition, he has co-organized six special sessions at international conferences and has been program committee member of over 50 international conferences. He is in the editorial board of four journals indexed by Scopus. He is a member of three IEEE task forces: Genetic Fuzzy Systems, Computational Intelligence in Security and Defense, and Intelligent System Application. He has coauthored 90 contributions to international journals and conferences. He is a Senior Member of the Association for Computing Machinery (ACM).

PIERO CASTOLDI received the Laurea degree *(cum laude)* in electrical engineering from the University of Bologna, and the Ph.D. degree from the University of Parma, in 1991 and 1996, respectively.

He is currently a Full Professor in telecommunications, and is responsible for the ''Networks and Services'' area of the TeCIP Institute at Scuola Superiore Sant'Anna, Pisa, Italy. He has also been coordinator of two Erasmus Mundus Master programs. He has served as the Project Manager for the Italian National Consortium for Telecommunications (CNIT). His most recent research interest cover protocols and switching techniques for optical networks, optical interconnection networks for Data Centers, photonic integrated networks, networks for industrial applications, 5G network architectures. He is author of more than 400 technical articles published in international journals and international conference proceedings. He has also filed a total of 20 patents.



NICOLA ANDRIOLLI received the Laurea degree in telecommunications engineering from the University of Pisa, in 2002, and the Diploma and Ph.D. degrees from Scuola Superiore Sant'Anna, Pisa, in 2003 and 2006, respectively.

He was a Visiting Student at DTU, Copenhagen, Denmark, and a Guest Researcher at NICT, Tokyo, Japan. He has been an Assistant Professor with the Scuola Superiore Sant'Anna, from 2007 to 2019. Since November 2019, he has been a Researcher with the Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy. His research interests span the optical networks, systems, and technology fields. They include the design and the performance analysis of optical circuit-switched and packet-switched networks and nodes, optical interconnection network architectures and scheduling, as well as photonic integration technologies for datacom and telecom applications. He has authored more than 170 publications in international journals and conferences, contributed to one IETF RFC, and filed 11 patents. He has been a TPC Member in several international conferences (ECOC, EuCNC, IPR, PSC) and serves as an Associate Editor of the IEEE Access.