Reconfigurable Activation Functions in Integrated Optical Neural Networks

José Roberto Rausell Campo and Daniel Pérez-López

***Abstract*—The implementation of nonlinear activation functions is one of the key challenges that optical neural networks face. To the date, different approaches have been proposed, including switching to digital implementations, electro-optical or all optical. In this article, we compare the response of different electro-optic architectures where part of the input optical signal is converted into the electrical domain and used to self-phase modulate the intensity of the remaining optical signal. These architectures are made up of Mach Zehnder Interferometers (MZI) and microring resonators (MRR). We have compared the corresponding transfer functions with commonly used activation functions in state-of-the-art machine learning models and carried out an in-depth analysis of the capabilities of those architectures to generate the proposed activation functions. We demonstrate that a ring assisted MZI and a two-ring assisted MZI present the highest expressivity among the proposed structures. To the best of our knowledge, this is the first time that a quantified analysis of the capabilities of optical devices to mimic state-of-the-art activation functions is presented. The obtained activation functions are benchmarked on two machine learning examples: classification task using the Iris dataset, and image recognition using the MNIST dataset. We use complex-valued feed-forward neural networks and get test accuracies of 97% and 95% respectively.**

***Index Terms*—Complex-valued neural networks, electro-optic modulation, machine learning, nonlinear optics, optical activation functions, optical neural networks.**

I. Introduction

ARTIFICIAL neural networks (ANN) have seen an exponential growth in recent years thanks to the development of algorithms with a wide range of applications such as image recognition [1], [2], speech recognition [3], mastering board games [4] and audio signal processing [5]. In turn, the computational requirements have escalated rapidly and future progress may be endangered if there is not an increase in the efficiency of the algorithms, or a movement to more computationally-efficient architectures [6]–[8].

In this regard, specialized electronic processors like graphical processing units (GPUs) [9], tensor processing units (TPUs) [10] and analog-based neuromorphic devices [11] have been investigated, improving the efficiency of generic CPUs by some orders of magnitude. However, these solutions rely on electronic systems and thus, have some fundamental limits in bandwidth and energy efficiency [12].

To overcome these hurdles, photonic processors have emerged as a promising technology for machine learning accelerators, capable to provide high bandwidths, high parallelism, low latencies and low crosstalk [13]–[15]. Different approaches based on free space optics have been proposed in the literature, using spatial light modulators (SLM), diffractive elements or coherent detection [16]–[19]. Recent progress in integrated photonic technologies have open a new platform for optical artificial neural networks (OANN) [20], [21]. A solution based on the broadcast-and-weight protocol uses microring resonators to introduce weights, and photodetectors to perform the optical matrix multiplication [6]. This architecture has the advantage of potential high-density integration and parallelization exploiting spectral diversity. However, it requires opto-electronic conversions in each neural network layer. Furthermore, MRR based systems present significant limitations for scaling up. They require high-finesse MRR, which limits the number of frequency channels available, and they present a quadratic scaling of the path length which negatively affects the computational speed and energy consumption [22]. Another approach performs optical matrix multiplications with meshes of Mach Zehnder Interferometers (MZIs), capable of implementing unitary transformations [23], [24]. Arbitrary linear matrix multiplications can also be performed using SVD decomposition [25].

Beyond the multiply and accumulate section, one of the remaining challenges in OANN is the implementation of the nonlinear activation function (NAF). Fig. 1 illustrates alternative ways to perform this functionality. NAF are essential for ANN because they allow to model target variables that have complex relations with their explanatory variables. To tackle this problem, ONN proposed in [17], [20], [21], photodetect the optical signal at the end of each layer an add the nonlinear stage using digital processors (Fig. 1a). The outputs are sent to a modulator to produce the input of the next layer. This solution takes advantage of the expressivity and programmability of digital processors and re-uses the same physical layer to feature deeper neural

Manuscript received October 30, 2021; revised February 28, 2022; accepted April 12, 2022. Date of publication April 26, 2022; date of current version May 9, 2022. This work was supported in part by FPI-UPV Grant Program under Grant PAID-01-20-24 from the Universitat Politècnica de València, through the Spanish MINECO Juan de la Cierva Program and in part by the H2020-ICT-2019-2 Neoteric 871330 Project. *(Corresponding author: Daniel Pérez-López.)*

José Roberto Rausell Campo is with the Institute of Telecommunications and Multimedia Applications (ITEAM), Universitat Politècnica de València, 46022 Valencia, Spain (e-mail: joraucam@upv.es).

Daniel Pérez-López is with the Institute of Telecommunications and Multimedia Applications (ITEAM), Universitat Politècnica de València, 46022 Valencia, Spain, and also with the iPronics Programmable Photonics S.L, Ed. 9B, Universitat Politècnica de València 46022, Spain (e-mail: dperez@iteam.upv.es;daniel.perez@ipronics.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2022.3169833.

Digital Object Identifier 10.1109/JSTQE.2022.3169833

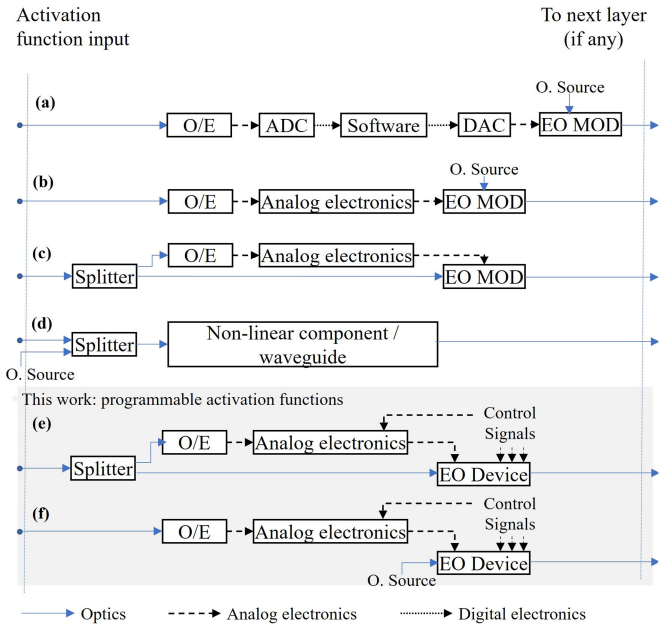

Fig. 1. Schematic non-linear activation function architectures for optical neural networks employing Optical (O), Electrical (E) and digital conversions and processing. (a) OEDEO approach, (b) Direct OEO, (c) self-modulated OEO, (d) Optical approach, (e) This work: programmable activation function based on self-modulated OEO. OEO Part of the input optical signal is divided and feed into and O/E circuit where it is photodetected and conditioned. (f) programmable activation function based on OEO link.

networks virtually; however, the analog-to-digital conversion reduces the computational speed. To avoid delays introduced by the digital signal processing, an alternative architecture attempts to introduce the nonlinear function optically (ONAF). On the one hand, optoelectronic activations convert part, or all, the optical signal into the electrical domain and use it to modulate the remaining, or a new, optical signal with systems that give nonlinear responses. In [26], absorption modulators with different active materials are studied. A variety of ONAF are synthetized using a doped MRR modulator in [27]. Reference [28], introduce a novel scheme where the incoming optical signal is sent into a directional coupler in which a small portion is separated and converted into the electrical domain and used for modulating the remaining optical signal using a MZI, obtaining a non-linear device (Fig. 1c). The reconfigurability of the nonlinear activation function is limited by the interferometric response and the electrical circuitry settings. Optical bistability using optoelectronic feedback have also been reported in [29] in a silicon photonics platform based on ring resonators.

Finally, the third solution for the non-linear layer is employing all-optical layer, (Fig. 1d). Without exiting the optical domain, these have the potential to offer the highest bandwidths and speed, but, optical nonlinearities are weak, requiring high power thresholds or high response time. A semiconductor optical amplifier (SOA) based scheme have been experimentally demonstrated to obtain sigmoid like activation function [30]. Moreover, the bistability of an injection-locked Fabry-Perot semiconductor laser [31], induced transparency and reversed saturated absorption [32] have been proposed. Finally, A. Jha *et al.*, demonstrated in [33] an all-optical reconfigurable NAF via the free-carrier dispersion effect in ring assisted Mach-Zehnder interferometers.

Since there is no universal activation function that can be employed for every learning process, there is a strong interest on achieving a reconfigurable hardware element that enables the programming of arbitrary nonlinear responses. In this work, we make an analysis of a collection of 9 electro-optic interferometric systems made up of combinations of MZIs and MRRs to generate reconfigurable optical nonlinear activation functions. The studied architectures are based on the previously described working principle of [28], where part of the incoming signal is converted into the electronic domain and used to feed one of the phase shifters of the system, producing a self-phase modulation (Fig. 1e). They are designed to work with MZI-based matrix multipliers employing a single wavelength. The transfer functions of the proposed reconfigurable systems are compared with 14 NAFs commonly used in state-of-the-art machine learning models. The ability of the optical systems to mimic these NAFs is measured using the root mean squared (*rms*) between both functions. Previous cited works rely on two different types of architectures. First, those which can only generate one optical nonlinearity, which although interesting lack programmability, and those which present programmable structures but have not carried out an in-depth analysis of the expressivity of their architectures. Our work presents a broad and measurable analysis of the expressivity of the different proposed O/E/O systems comparing their response to the most common state-of-the-art activation functions and allows to determine which electro-optic scheme is most suitable in terms of programmability for feed-forward ONN.

The remainder of the article is structured as follows. In Section II, the typical NAFs used to compare with the response of our optical architectures are presented. In Section III, we describe the proposed optical architectures. In Section III-A, we present the transfer functions of the systems. In Section III-B, the transfer functions of the proposed systems are fitted to the data generated by the equations in Section II, using a method based on the Levenberg-Marquardt algorithm. We used the *root mean squared error (rmse)* to evaluate the expressivity of the devices as nonlinear activation functions generators. In Section III-C, we simulate fabrication errors and component imperfections of the optical systems and analyze its consequences in the performance of the architectures. In Section IV, the ONAFs generated by a ring assisted MZI and a two-ring assisted MZI are tested on two machine learning benchmarks, the classification problem based on the classification of flowers using the IRIS dataset and the image recognition problem using the MNIST dataset.

II. Digital Nonlinear Activation Functions

In this section, we describe the nonlinear activation functions that are used as targets for our optical interferometric architectures. These functions, correspond to the more common nonlinear activations implemented in digital processors. Extra variables are incorporated into the equations to scale the output

and displace the origin of the functions because the ONAFs are studied in the power domain, i.e., only positive inputs and outputs are permitted. Moreover, the added parameters allow to calculate different representations of the same function in the optical domain, giving the possibility to test the expressivity and limits of our optical devices.

Hereafter, the collection of 14 functions used in our work is presented:

- ReLU [34]:

$$f\ (\mathrm{z}) = \max\left(m\left(z-\beta\right),\ 0\right) \tag{1}$$

- Clipped ReLU:

$$f\ (z) = \min\left(\beta z,\ \alpha\right)\ ,\ \beta \leq 1 \tag{2}$$

- ELU [35]:

$$f\ (z) = \begin{cases} \alpha[\exp\left(m\left(z-\beta\right)-1\right]+c,\ z \leq \beta \\ m\left(z-\beta\right)+c,\ z \geq \beta \end{cases} \tag{3}$$

- GeLU [36]:

$$f\left(z\right) = \left[\frac{1}{2}\alpha\left(z-\beta\right) \times \left(1+\mathrm{erf}\left(\frac{\alpha\left(z-\beta\right)}{\sqrt{2}}\right)\right)+c\right]\ /scale \tag{4}$$

- Parametric ReLU [37]:

$$f\ (z) = [\max\left(\alpha\left(z-\beta\right),\ \left(z-\beta\right)\right)+c]/\ scale\ ,\ a \leq 1 \tag{5}$$

- SiLU [38]:

$$f\ (z) = \left[\frac{\alpha\left(z-\beta\right)}{1+\exp\left(-\alpha\left(z-\beta\right)\right)}+c\right]\ /scale \tag{6}$$

- Gaussian:

$$f\ (z) = \exp\left(\frac{-\left(z-\beta\right)^2}{2\alpha^2}\right)\ /scale \tag{7}$$

- Quadratic:

$$f\ (z) = \ z^2 \tag{8}$$

- Sigmoid:

$$f\ (z) = \frac{scale}{1+\exp\left(-\alpha\left(z-\beta\right)\right)} \tag{9}$$

- Sine:

$$f\left(z\right)\ =\ \left[\sin\left(\alpha z\ +\beta\right)\ +\ 1\right]/2 \tag{10}$$

- Softplus [39]:

$$f\ (z) = \ \log(1+\exp\left(\alpha\left(z-\beta\right)\right)/scale \tag{11}$$

- Tanh:

$$f\ (z) = \left[\tanh\left(\alpha\left(z-\beta\right)\right)\ +c\right]\times scale \tag{12}$$

- Softsign [40]:

$$f\ (z) = \left[\frac{\alpha\left(z-\beta\right)}{1+\ \left|\alpha\left(z-\beta\right)\right|}+c\right]\ /scale \tag{13}$$

TABLE I
SUMMARY OF PARAMETER VALUES FOR THE ACTIVATION FUNCTIONS

	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
ReLU	-	0.15, 0.25, 0.35, 0.55
Clipped ReLU	0.6, 0.8, 1.0	0.55, 0.65, 0.75, 0.85
ELU	0.1, 0.2, 0.3	0.15, 0.25, 0.35, 0.55
GeLU	6.0, 8.0, 10.0	0.25, 0.35, 0.45, 0.55
Parametric ReLU	0.1, 0.15, 0.2	0.15, 0.25, 0.35, 0.55
SiLU	10.0, 15.0, 20.0	0.25, 0.35, 0.45, 0.55
Gaussian	0.1, 0.2, 0.3	0.5, 0.6, 0.7
Quadratic	-	-
Sigmoid*	10.0, 20.0	0.5, 0.7
Sine	1.0, π, 2π	-π/2
Softplus	10.0, 15.0, 20.0	0.2, 0.3, 0.4
Tanh	5.0, 10.0, 15.0	0.5, 0.7
Softsign	5.0, 10.0, 15.0	0.5, 0.7
Exponential	-	3.0, 6.0, 9.0

*Sigmoid is divided into two sets, Sigmoid 1 and Sigmoid 2. For Sigmoid 1, *scale* is variable in (9) and for Sigmoid 2, *scale = 1*.

- Exponential:

$$f\ (z) = \ \exp\left(\beta\left(z-1\right)\right) \tag{14}$$

α and β are variable parameters used to display different representations of the same function. The selected values are shown in Table I. Each of the α values is combined with all the β values, resulting in a total of 108 representations. Parameters c and $scale$ are calculated automatically depending on α and β to keep the output of the function between 0 and 1. This feature is necessary because we have worked with normalized input power, that is, our inputs are bounded between 0 and 1, and the studied devices are passive, which forces the outputs to be equal or lower than the inputs.

III. PROPOSED ARCHITECTURES

In this section, we present the proposed hardware architectures to generate nonlinear functions for optical neural networks. Our work is based in the combination of two basic integrated optical devices that have been extensively studied in the literature, Mach Zehnder interferometers [41], and microring resonators [42]. MZIs include the combination of 2 3-dB couplers and 2 phase shifters. The control of each phase shifter enables the creation of tunable optical couplers with independent power ratio and phase shift tuning. Regarding the ring resonators, these are close-path waveguides (cavities) that are accessed through optical couplers. Moreover, if we integrate phase actuators within the cavity and/or couplers, their response can be tuned using external electrical signals, giving the possibility to obtain a programmable device [43], [44].

The use of MZIs to generate activation functions in ONN using E/O conversion was first reported by [45]. In this work, I. Williamson *et al.*, presented an architecture where a small part

of the incoming signal power, around 10%, is separated from the input signal by passing it through a directional coupler. This signal power is then converted into the electrical domain using an optical-to-electrical conversion circuit. First, the optical signal is photodetected and turned into an electrical current. Second, this current is converted into a voltage that goes through a nonlinear signal conditioner that adds and additional nonlinear conversion to the voltage signal. Finally, the output of the conditioner is combined with a static bias voltage.

The resulting signal serves as a voltage supply of an optical phase shifter that induces a phase shift on the remaining optical signal that travels through one of the arms of an MZI. The applied phase shift corresponds to

$$\Delta\phi \;=\; \frac{\pi}{V_\pi}\left[V_b + H\left(G\Re\alpha|z|^2\right)\right], \tag{15}$$

where $|z|^2$ is the normalized optical power extracted in the optical monitor, α is the ratio of optical power divided in the directional coupler, $\Re$ is the responsivity of the photodetector, G is the gain of the transimpedance amplifier, $H(.)$ is the transfer function of the nonlinear conditioner, V_b is the static bias voltage and V_π is the necessary voltage to induce a phase shift of π in the phase shifter.

The optical signal at the output of the MZI is then a self-phase modulation of the input signal, that is, the output signal depends nonlinearly on the input signal. The resulting architecture is limited in the number and type of nonlinear responses it can synthetize.

In our work, with the aim of proposing a flexible hardware to enable arbitrary nonlinear functions, we follow a similar approach. The studied architectures are a combination of MZIs and ORRs where some of its parameters, coupling coefficients and phase shifts, can be variable. All the resulting devices are built with a phase shifter working as described in (15). In our case, we use a more general form in which the correspondent induced phase shift is

$$\Delta\phi \;=\; a + b|z|^2, \tag{16}$$

where a and b are tunable and represent the O/E circuit parameters: voltage, gain, responsivity …. A general scheme of the described system is presented in Fig. 1e, where the aforementioned phase shift is embedded in the programmable interferometric device.

For the interferometric device analysis, we selected a total number of 9 architectures, Fig. 2., to compute and compare their capability to program the aforementioned non-linear responses. These architectures are (a) an MZI, (b) an ORR, (c) an ORR in add-drop configuration, (d) an ORR in add-drop configuration with the output port coupled, (e) a ring-assisted MZI, (f) a double ring-assisted MZI, (g) a double SCISSORS, (h) a double CROW, (i) a double CROW in add-drop configuration with the output ports coupled. The k_i coefficients represent the coupling coefficients of the waveguides. Coupling can be made using directional couplers (DC), multimode interferometers (MMI), or MZIs if the k_i is defined to be tunable. The ϕ_i represent phase shifts in the waveguide. These shifts can be implemented using thermo-optic phase shifters, micro electromechanical (MEMS) phase shifters or phase shifters based in phase change materials (PCM). Subindex *r* is used to indicate that the correspondent element is part of a ring resonator.

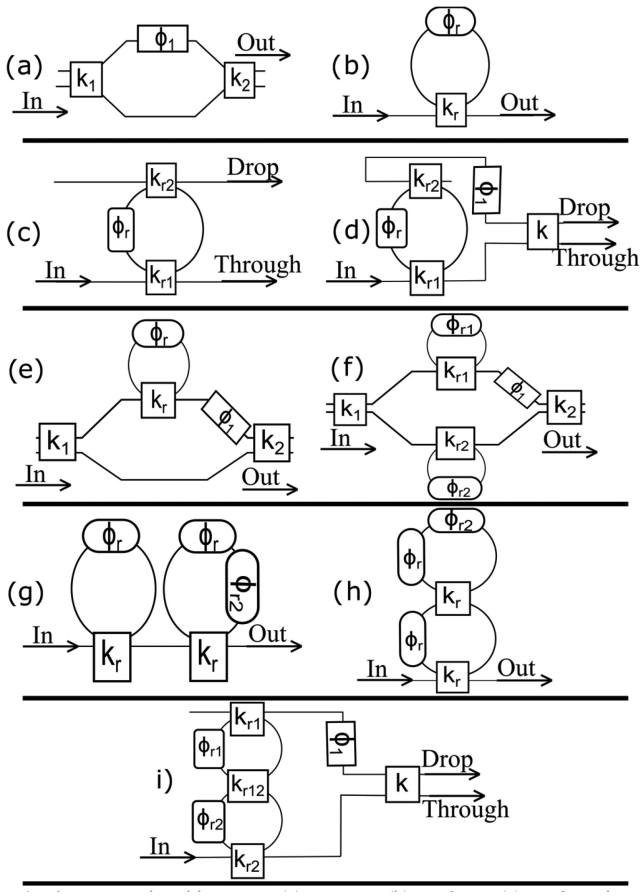

Fig. 2. Proposed architectures: (a) an MZI, (b) an ORR, (c) an ORR in add-drop configuration, (d) an ORR in add-drop configuration with the output port coupled, (e) a ring-assisted MZI, (f) a double ring-assisted MZI, (g) a double SCISSORS, (h) a double CROW, (i) a double CROW in add-drop configuration with the output ports coupled. Parameters $\mathrm{k_i}$ represent coupling coefficients and ϕ_i represent phase shifts.

The number of tunable parameters in the system is a collection of the optical system parameters: coupling and phase shifts, and the O/E circuit parameters: a and b. To limit the possible final configurations and degrees of freedom in each architecture, we allow a maximum of 4 tunable variables, or free parameters. Depending on these free parameters, we have different subsystems for the same architecture. The different combination of parameters for each architecture is summarized in Table II. The elements presented as $\phi_i^{a,b}$ denote the phase shifters that perform the self-phase modulation as described in (16), and therefore, consider the two parameters a and b of the O/E circuit.

A. Transfer Functions

The generated output field in an optical device is a combination of the incoming optical field and the transfer function of the system as follows,

$$z_{out} \;=\; H_{transfer} z_{in}, \tag{17}$$

TABLE II
SUMMARY OF FREE PARAMETER COMBINATIONS FOR THE OPTICAL DEVICES

	1	2	3
MZI	$\phi_1^{a,b}$	$\phi_1^{a,b}, k_1$	-
ORR	$\phi_r^{a,b}$	$\phi_r^{a,b}, k_r$	-
ORR – AD drop port	$\phi_r^{a,b}$	$\phi_r^{a,b}, k_{r1}, k_{r2}$	-
ORR – AD through port	$\phi_r^{a,b}$	$\phi_r^{a,b}, k_{r1}, k_{r2}$	-
ORR – AD coupled	$\phi_r^{a,b}, \phi_1$	$\phi_r^{a,b}, \phi_1, k$	-
MZI – ORR	$\phi_r^{a,b}, \phi_1$	$\phi_r^{a,b}, \phi_1, k_r$	$\phi_r^{a,b}, k_r$
MZI – 2 ORR	$\phi_{r1}^{a,b}, k_{r1}, k_{r2}$	$\phi_{r1}^{a,b}, k_{r1}, \phi_1$	$\phi_{r1}^{a,b}, k_{r1}, \phi_{r2}$
Double SCISSORS	$\phi_r^{a,b}, k_r, \phi_{r2}$	-	-
Double CROW	$\phi_r^{a,b}, k_r, \phi_{r2}$	-	-
Double CROW -AD coupled	$\phi_{r1}^{a,b}, k, \phi_1$	$\phi_{r1,r2}^{a,b}, k, \phi_1$	$\phi_{r1,r2}^{a,b}, k_r, \phi_1$

where $H_{transfer}$ is the transfer function and describes the complex response of the whole optical system.

The comparison between the digital and the optical generated activation functions is made using the optical power as an input. Therefore, our optical nonlinearities will have the form:

$$f\left(|z_{in}|^2\right) = |H_{transfer}|^2 |z_{in}|^2. \tag{18}$$

Next, we introduce the corresponding transfer functions of the aforementioned optical systems:

- MZI:

$$H_{MZI} = -j\left(s_2 c_1 \exp\left(-j\phi_1\right) + c_2 s_1\right) \tag{19}$$

- ORR:

$$H_R = \frac{c_r - \gamma x \exp\left(-j\phi_r\right)}{1 - c_r \gamma x \exp\left(-j\phi_r\right)} \tag{20}$$

- ORR – AD drop port:

$$H_{Od} = \frac{-s_1 s_2 \sqrt{\gamma x \exp\left(-j\phi_r\right)}}{1 - c_1 c_2 \gamma x \exp\left(-j\phi_r\right)} \tag{21}$$

- ORR – AD through port:

$$H_{Ot} = \frac{c_{r1} - c_{r2} \gamma x \exp\left(-j\phi_r\right)}{1 - c_{r1} c_{r2} \gamma x \exp\left(-j\phi_r\right)} \tag{22}$$

- ORR – AD coupled:

$$H_{ORR-cpl} = -jsH_{Ot} \exp\left(-j\phi_1\right) + cH_{Od} \tag{23}$$

- MZI – ORR:

$$H_{MO} = -j\left(s_2 c_1 H_R \exp\left(-j\phi_1\right) + c_2 s_1\right) \tag{24}$$

- MZI – 2ORR:

$$H_{MO} = -j\left(s_2 c_1 H_{R1} \exp\left(-j\phi_1\right) + c_2 s_1 H_{R2}\right) \tag{25}$$

- Double SCISSORS:

$$H_{sc} = H_{R1} H_{R2} \tag{26}$$

- Double CROW:

$$H_{CR} = \frac{c_1 - c_{12} a_1 - c_1 c_{12} a_2 + a_1 a_2}{1 - c_1 c_{12} a_1 - c_{12} a_2 + c_1 a_1 a_2} \tag{27}$$

- Double CROW – AD coupled:

$$H_{CRt} = \frac{c_1 - c_{12} a_1 - c_1 c_2 c_{12} a_2 + c_2 a_1 a_2}{1 - c_1 c_{12} a_1 - c_{12} c_2 a_2 + c_1 c_2 a_1 a_2}$$

$$H_{CRd} = \frac{j s_1 s_{12} s_2 \sqrt{a_1 a_2}}{1 - c_1 c_{12} a_1 - c_{12} c_2 a_2 + c_1 c_2 a_1 a_2}$$

$$H_{CR-cpl} = -jsH_{CRt} \exp\left(-j\phi_1\right) + cH_{CRd} \tag{28}$$

where $c_i = \sqrt{1 - k_i}$, $s_i = \sqrt{k_i}$ and γ is the loss coefficient. $x = \exp(-j\beta L)$, where $\beta = 2\pi n_{eff}/\lambda$ is the propagation constant, L is the total length of the correspondent ring, n_{eff} is the effective index of the material and λ is the wavelength of the optical signal. The terms a_i in (27) and (28) have been introduced to simplify the equations and correspond to $a_i = \gamma_i x_i \exp(-j\phi_i)$.

B. Generation of Specific Functions Through Computational Optimization

To understand the expressivity of the optical devices under study we map the optical transfer functions build as (18) with the standard activation functions described in Section II. The process is built as follows.

First, we choose one of the activation functions described in (1) – (14). Using the correspondent equation, a 500 points array is generated. The input of the function is a set of evenly spaced numbers over the interval [0,1].

Second, we select the optical system to study and its response. The optical devices are simulated using state-of-the-art fabrication processes parameters. We use the standard wavelength for telecommunications systems, $\lambda = 1.55\ \mu\mathrm{m}$. We consider silicon waveguides of width and height of $500\ nm$ and $220\ \mathrm{nm}$ respectively, with propagation losses of $\alpha = 2\ dB/cm$ and an effective index of $n_{eff} = 2.394$ [46]. All the ring resonators have the same radius of $r = 300\ \mu m$ to anticipate the need of phase shifters with length in the range of several tens or hundreds microns. Insertion losses have been neglected in our simulations. Overall, although the impact of the insertion loss of the activation function can be compensated through additional optical power in the optical source, this power penalty is expected to decrease with optimized components like low loss couplers [47], [48] and tunable cavities [49].

The response $f(|z_{in}|^2)$ is then fitted to the data previously generated using a non-linear least square method based on the Levenberg-Marquardt algorithm. The parameters to be optimized during the fitting process are those shown in Table II.

Finally, using the optimized parameters, we calculate the response of the system for the same input points of the first step. We choose the *root mean square error (rmse)* between the optical response and the activation function as a figure of merit

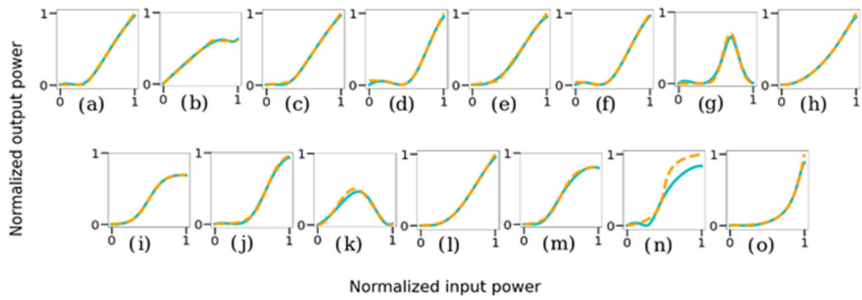

Fig. 3. Activation functions generated by an MZI-ORR: (a) ReLU, (b) Clipped ReLU, (c) ELU, (d) GeLU, (e) PReLU, (f) SiLU, (g) Gaussian, (h) Quadratic, (i) Sigmoid 1, (j) Sigmoid 2, (k) Sine, (l) Softplus, (m) Tanh, (n) Softsign, (o) Exponential. In dotted yellow, the target activation function and in solid blue, the system' s response.

of the quality of the optimization. The *rmse* formula is

$$\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_{i,o}-y_{i,a})^2}, \tag{29}$$

where $y_{i,o}$ represents the optical output power and $y_{i,a}$ the corresponding activation function value.

The rms results for combinations of photonic circuits and targeted activation functions are presented in Fig. 12. using a logarithmic scale, that is, $rmse_{dB} = 10log_{10}(rmse)$. Each color represents an interval. From worst to best, values of $rmse_{dB}$ between 0 and -10 are represented using grey boxes. Values between -10 and -15 use orange boxes. Light green boxes depict values between -15 and -20, and dark green boxes are used for values between -20 and -25, which is the lowest achieved number.

The heatmap shows that the interferometric architectures that include MZIs achieve better results than those build only with ring resonators. In particular, the response of the drop port of a ring resonator performs extremely poor, being incapable to achieve $rmse$ lower than -11 dB.

Moreover, interferometric architectures where MZIs and ORRs are combined perform better than those with only one of those devices. It is also implied that more complex architectures give higher degree of expressivity incurring in a complexity vs expressivity trade-off. Even though, when comparing the ring assisted MZI (MZI-ORR) and the two-ring assisted MZI (MZI-2ORR), the results show only slightly differences for their best cases. An example of the obtained nonlinearities using the MZI-ORR is shown in Fig. 3.

It is important to notice that some of the activation functions cannot be achieved with any of the proposed architectures, specifically some Clipped ReLU, Sigmoid and Tanh functions. This is because for some values the output is higher than the input and that cannot be replicated by our passive devices. Systems made of active materials would be necessary to introduce those activation functions [50].

On the other hand, along with the Softplus, the ReLU and its variations are easier to reproduce by all the optical architectures than the other activation function whose response presents more variation in their working domain.

C. Sensitivity to Non-Idealities and Parasitic Effects

The aforementioned comparison assumed accurate precision for the driving conditions. However, real devices have a non-perfect behavior due to component imprecision that come from fabrication errors, waveguide roughness, temperature variations and crosstalk (electrical, optical, thermal). When errors of different components accumulate, the performance of the system can vary significantly from the desired behavior, and thus, degrade the final results. Furthermore, an optical circuit for deep learning is composed of different stages of matrix multiplication and activation function arrays, and it is important to maintain a common response of the optical devices to ensure circuit scalability.

In this subsection, we study the variation of the proposed circuit architectures responses when random noise is added to the ideal value of the parameters of the system after the optimization process. To do so, we add a gaussian noise to the coupling coefficients k_i, to the phase shifts ϕ_i, and to the a and b parameters that control the self-phase modulation as follows

$$x \rightarrow x + \mathcal{N}(o,\sigma) \tag{30}$$

where $\mathcal{N}(o,\sigma)$ refers to a normal distribution with mean equal to 0 and standard deviation σ. Then, the new distorted parameters serve to calculate the response of the system when affected by noise. We are aware that some of the components show correlated variations during the manufacturing processes [50], [51]. However, we have decided to maintain non-correlated Gaussian distributions to deal with the worst-case scenario and ensure that the experimentally expected results are equal to or better than those simulated.

Figs. 4-5 show the variation of the $rmse$ when noise is added to the MZI-ORR with free parameters $\phi_r^{a,b}$, ϕ_1, k_r, and to the MZI-2ORR with free parameters $\phi_{r1}^{a,b}$, k_{r1}, ϕ_1, respectively. For the statistical analysis, we consider standard deviations of the normal distribution from coupling factors and phase shifters that represent incremental non-ideal operation within practical values [53]: a) $\sigma = 0.5\%$, b) $\sigma = 1\%$, c) $\sigma = 1.5\%$ and d) $\sigma = 2\%$. 100 points have been used in each of the simulations.

Boxes extend from the Q1 and the Q3 quartile values of the $rmse_{dB}$. The line inside the boxes represents the median, Q2 quartile. The whiskers extend from the boxes to the maximum and minimum value of the data but limited to a maximum size of $1.5 * IQR$, where $IQR = Q3 - Q1$. Data points beyond those

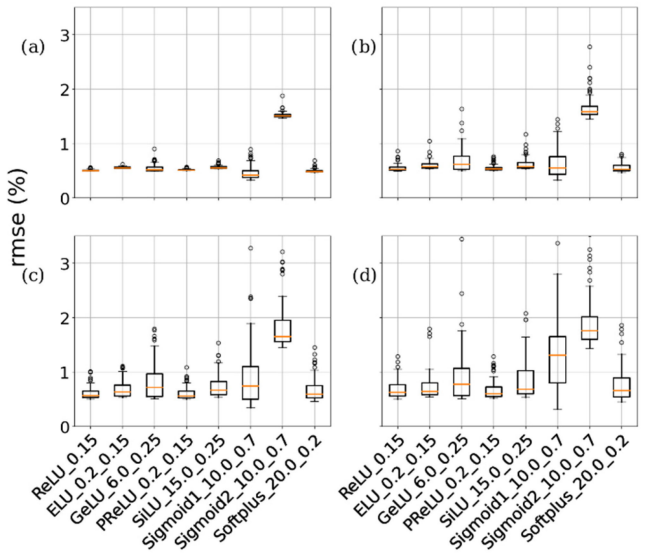

Fig. 4. RMSE (%) variation of the MZI-ORR generated activation functions for different values of standard deviation: (a) 0.5%, (b) 1.0%, (c) 1.5% and (d) 2%.

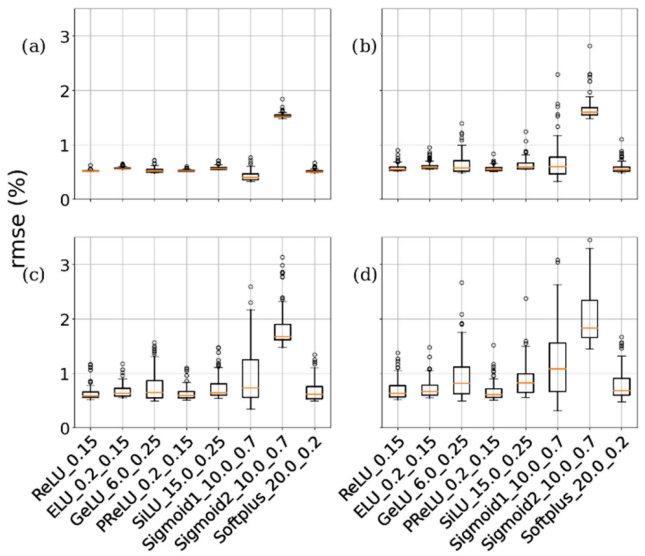

Fig. 5. RMSE (%) variation of the MZI- 2 ORR generated activation functions for different values of standard deviation: a) 0.5%, b) 1.0%, c) 1.5% and d) 2%.

limits are considered outliers and plotted with circles. The lower whisker cannot extend further than the optimal value obtained when noise is equal to zero.

We focus our analysis on the $\sigma = 1\%$ case, that correspond to the state-of-the-art precision. In the functions presented, the ReLU, ELU, Parametric ReLU and Softplus are the most robust activation with respect to noise. In the worst scenario, they go from percentage errors of 0.50, 0.54, 0.50, 0.47 to 0.86, 1.04, 0.76 and 0.80 respectively for the MZI - ORR architecture and from 0.51, 0.55, 0.51, 0.49 to 0.90, 0.95, 0.83, 1.10 for the MZ I- 2 ORR architecture.

SiLU and GeLU have slightly higher deviations from the noise-zero case. They go from percentage errors of 0.53 and 0.50 to 1.16, 1.63 for the MZI -ORR and from 0.55, 0.49 to 1.34, 1.29.

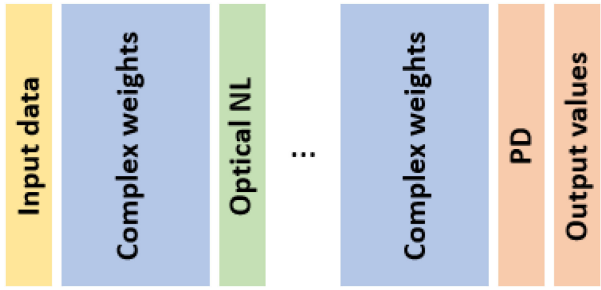

Fig. 6. General scheme of the neural networks model.

Finally, the Sigmoid 1 and Sigmoid 2 present more sensitivity to driving and coupling fluctuations. They go from percentage errors of 0.32 and 1.45 to 1.64 and 2.77 for the MZI - ORR and from 0.32 and 1.48 to 2.29 and 2.82 for the MZI – 2 ORR architecture. The reason behind this sensitivity is that, to match the sigmoid functions, the interferometric devices are configured in a working region presenting higher slope (gradient) versus the deviation of the parameters under study.

It is also important to highlight that for the studied functions, the results given by the MZI – ORR are more robust in terms of noise than those given by the MZI – 2 ORR architecture. The scalability of the proposed E/O architectures in terms of optical losses, bandwidth and SNR is addressed in Appendix B.

IV. Machine Learning Examples

In this section, we use the best E/O activation functions obtained from the MZI-ORR and the MZI-2 ORR architecture to solve two machine learning problems. In Section IV-A we train a neural network to solve the classification problem of the IRIS dataset. In Section IV-B we benchmark our optical nonlinearities with the MNIST image classification problem.

The neural network is implemented with the *TensorFlow* framework [54]. To simulate the response of an MZI mesh-based matrix multiplier, we use complex-valued weights and biases. Our optical nonlinear activations are programmed to work in the optical field domain, $f\left(z_{in}\right) = H_{transfer}.\, z_{in}$. In the final stage of the neural network, we introduce a quadratic function $|.|^2$ that mimics the response of a photodetector. When using one-hot encoding, the output of the photodetector is normalized by the sum of the output vector to create a probability distribution. Fig. 6 shows a general scheme of the neural networks. Note that, although some portion of the signal is converted to the electrical domain in the nonlinear layers, the optical signal can complete the full neural network in the optical domain till arriving to the photodetection layer at the end.

A. Iris Dataset

The Iris dataset is one of the most common used datasets for classification task benchmarking. It was first introduced by R.A. Fischer in 1936 [55]. The objective of the problem is to classify a given iris plant into one of the three subspecies: *setosa,*

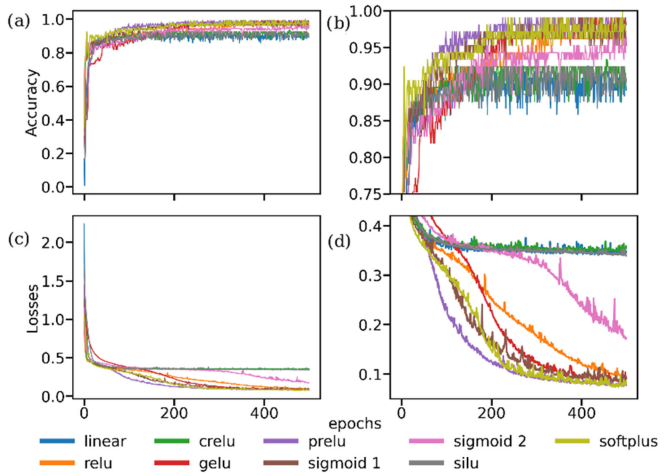

Fig. 7. (a) Accuracy, (b) Accuracy zoom view. (c) Losses and (d) Losses zoom view of the MZI – ORR activation functions for the IRIS dataset.

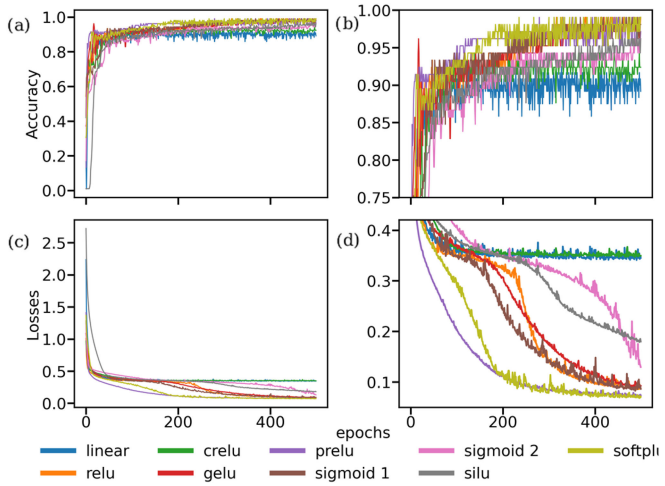

Fig. 8. (a) Accuracy, (b) Accuracy zoom view. (c) Losses and (d) Losses zoom view of the MZI – 2 ORR activation functions for the IRIS dataset.

versicolour or virginica. The dataset consists of 150 instances, 50 per each of the three classes, containing information of four attributes: *sepal length, sepal width, petal length and petal width.* One class is linearly separable from the other two, that are not separable from each other.

We use a feed-forward architecture made up of an input layer of 4 units, one hidden fully connected complex layer with 10 units and one output layer with 3 units.

The dataset is split into training and test samples using a 70:30 ratio. Input data is normalized in the [0,1] interval and encoded in the real part of the input tensors. Imaginary part of the input tensors is set to zero for all the examples. The output is a 3-dimension vector with one-hot encoding. We use the Adam optimizer with a learning rate of 0.0015 and a batch size of 32.

Figs. 7 and 8 represent the evolution of the a) accuracy and c) losses during the training of the model using the optical nonlinearities generated by the MZI – ORR and the MZI – 2ORR respectively. For both architectures, except from the Clipped ReLU activation (green line), the rest of the benchmarked optical

TABLE III
TEST RESULTS FOR THE IRIS DATASET MODEL

	MZI - ORR		MZI – 2 ORR	
	Accuracy (%)	Loss	Accuracy (%)	Loss
ReLU	97.78	0.102	97.78	0.095
Clipped ReLU	86.67	0.331	88.89	0.344
GeLU	97.78	0.091	97.78	0.097
PReLU	97.78	0.078	97.78	0.078
Sigmoid 1	97.78	0.114	97.78	0.111
Sigmoid 2	95.56	0.144	95.56	0.157
SiLU	97.78	0.153	97.78	0.176
Softplus	97.78	0.086	97.78	0.084
Linear	**86.67**	**0.382**		

nonlinearities achieve high accuracies during the training. Parametric ReLU (purple), Softplus (yellow), ReLU (orange), GeLU (red), and Sigmoid 1 (brown) present the best performance with accuracies over 98%. Sigmoid 2 (pink) and SiLU (grey) results are poorer, giving training accuracies around 95%. Clipped ReLU results are similar to the results given by the linear (blue) activation functions. This is because Clipped ReLU has a linear response for input values close to the origin of the function, and thus, has higher difficulties to segment the space in nonlinear clusters.

Analyzing the losses curves, we can see that the Parametric ReLU and the Softplus reach lower values in the least epochs. They are followed by the ReLU, Sigmoid 1 and GeLU. Finally, Sigmoid 2 and SiLU have more difficulties to achieve comparable results.

The model evaluation results using the test data are presented in Table III. Our model achieves high accuracy results for all the optical nonlinearities except from Clipped ReLU, as expected by the training curves, outperforming the linear model by a 10%. These results are comparable to those obtained in previous works [28], [56].

B. *MNIST Dataset*

We choose the MNIST classification problem as a second benchmark for our optical nonlinear activations. The MNIST dataset is a collection of 60000 training examples and 10000 test examples of grayscale handwritten digits images with 28 x 28 digits.

Each 28 x 28 image is flattened into 728 units to be the input of the neural network as shown in Fig. 9. The magnitude of each pixel is normalized between 0 and 1, and it is encoded in the real part of the complex input vector. After the input layer there is a 25 units hidden layer with complex weights and finally there is a 10 units output layer. The output of the neural network is one-hot encoded.

During the training, we have used the Adam optimizer with a learning rate of 0.001 and a batch size of 256 with 500 epochs. We use as a loss the cross-entropy between the output of the neural network and the one-hot output vector.

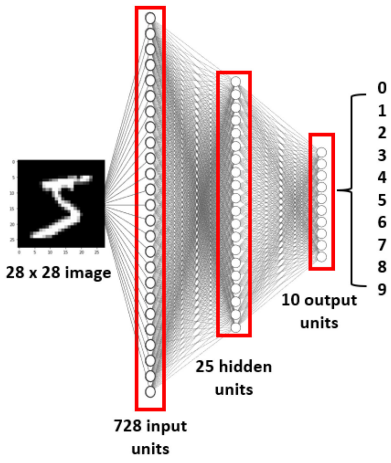

Fig. 9. Scheme of the model used to solve the image recognition MNIST problem.

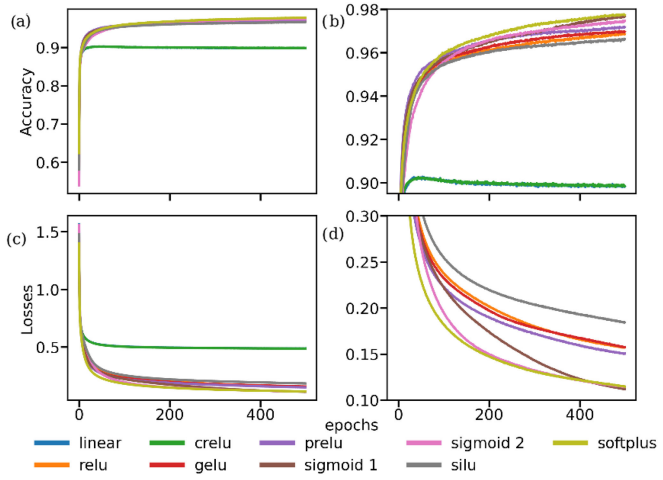

Fig. 10. (a) Accuracy, (b) Accuracy zoom view. (c) Losses and (d) Losses zoom view of the MZI – ORR activation functions for the MNIST dataset..

Accuracy and loss evolution during training are presented in Figs. 10 and 11 for the MZI – ORR and the MZI – 2 ORR respectively. As in the IRIS problem, optical Clipped ReLU results are indistinguishable from the lineal, or no linearity, case. The remaining optical activations reach training accuracies between 96-98%, outperforming the linear case almost a 10%. In this task, Softplus activations achieves the higher training accuracy, followed by the Sigmoid 1, Sigmoid 2 and PReLU optical nonlinearities. Finally, ReLU, GeLU and SiLU give training accuracies that are 1-2% lower than the previous ones.

In Table IV, we calculate and present test accuracies and losses. Final performance is similar for all the optical nonlinearities for both architectures as obtained for the IRIS dataset. It is important to point out that the optical activations functions that give better results are not all the same for the IRIS and the MNIST. This is clear example of the necessity of programmable devices that can generate a wide variety of optical activation functions for different learning tasks.

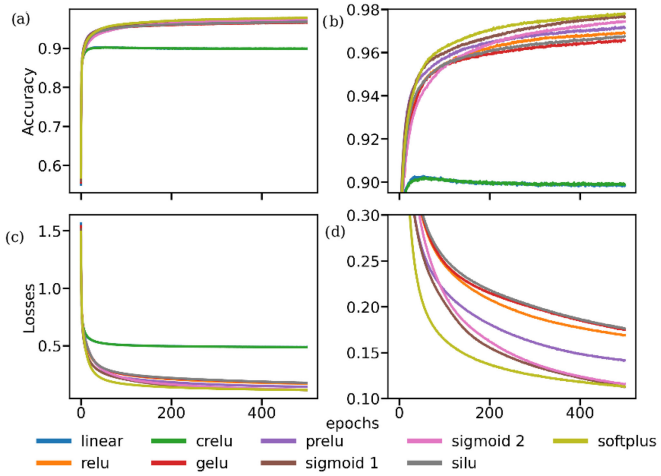

Fig. 11. (a) Accuracy, (b) Accuracy zoom view. (c) Losses and (d) Losses zoom view of the MZI – 2 ORR activation functions for the MNIST dataset.

TABLE IV
TEST RESULTS FOR THE MNIST DATASET MODEL

	MZI - ORR		MZI – 2 ORR	
	Accuracy (%)	Loss	Accuracy (%)	Loss
ReLU	95.57	0.195	95.63	0.206
Clipped ReLU	89.61	0.511	89.70	0.505
GeLU	95.76	0.196	95.46	0.210
PReLU	95.57	0.181	95.91	0.185
Sigmoid 1	95.95	0.170	96.25	0.152
Sigmoid 2	96.06	0.155	95.93	0.162
SiLU	95.70	0.191	95.52	0.211
Softplus	95.86	0.163	96.02	0.163
Linear	**89.60**	**0.510**		

V. CONCLUSION

The selection of the non-linear activation function is critical for the accurate learning process in machine learning. Most of the implementations today relay on its digital implementation degrading the benefits of the optical stage. Alternative EO or all-optical solutions have been demonstrated constrained to either single or a small set of activations functions. In this work, we have studied the capabilities of 9 optical architectures made up of MZIs and ORRs to generate a set of optical nonlinear activation functions with a common hardware. The general architecture is based in a O/E/O circuit with a self-phase modulation scheme, where a little amount of the optical input power is taken and photodetected to use the generated current to, together with an electrical circuit, feed one of the phase shifters of the optical architecture.

For each of the optical systems we have defined subsystems depending on the device parameters that are selected as tunable variables during the activation function reconfiguration. We analyzed up to 14 nonlinear activation functions commonly used in digital processors. The *rmse* between the fitted function and the sought activation function have been used as a measure of the quality of the optical nonlinearities. Among all the given results

we have selected two architectures that provide a high expressivity (i.e., capability to synthetize more activations functions with less error): the ring assisted MZI and the two ring assisted MZI.

To evaluate the robustness of the obtained functions to component variations we have studied the deviation of the rmse from the optimal value by adding a gaussian noise to the coupling coefficients and the phase shifters. We have used standard deviations of 0.5, 1, 1.5 and 2 percent. We see that the ReLU, ELU, Parametric ReLU and Softplus are the most robust functions in terms of noise, with deviations between 0.3 and 0.4%. They are followed by the GeLU, SiLU that give slightly higher deviations of 0.6 and 1.1%. Finally, the Sigmoid 1 and Sigmoid 2 give the worst results with deviations of 1.9 and 1.4%. We also highlight that the MZI – ORR is less affected by noise than the MZI – 2 ORR architecture.

The activations that achieved better results, that is, the ReLU, Clipped ReLU, GeLU, Parametric ReLU, SiLU, Softplus, Sigmoid 1 and Sigmoid 2 have been used to solve two standard machine learning tasks for benchmarking, IRIS dataset and MNIST dataset. In both problems, the optical activation functions achieve accuracies higher than 95%, which is comparable to other results obtained in previous similar works. It has been shown that the activation functions resulting in higher accuracies change depending on the addressed problem, highlighting the necessity of programmable optical devices for the implementation of the non-linear layer.

Although the future of integrated ONN seems promising, alternative approaches relying on free-space optics offer interesting trade-offs in terms of footprint, parallelization, and scalability. For example, whereas integrated solutions scale linearly with the number of outputs at each layer, O(N_{out}). free space architectures have introduced activation functions that do not scale with the number of outputs, preserving the full parallelism of free-space optics [57], [58]. All in all, future efforts in the field should be directed towards practical demonstrations of ONN and benchmarking with state-of-the-art electronic-based solutions.

APPENDIX A

See Fig. 12.

APPENDIX B
SCALABILITY

The presented electro-optical systems avoid the need for digital processing of the optical signal to apply the correspondent nonlinear activation function. However, these types of architectures present a trade-off between the bandwidth of the electronic loop and the achievable signal to noise ratio, limiting the system scalability and computational speed.

In the presented work, the analog electronic loop consists of a photodetector, that detects part of the incoming optical signal, an amplifier, and a bias signal. Noise sources will come mainly from the photodetector and the amplifier. Photodetectors major noise sources can be classified as thermal noise, shot noise and dark current noise. Thermal noise mean squared noise current can be modeled as $\langle i \rangle_{th}^2 = \frac{4kT}{R_L} B$, where k is the Boltzmann's constant, T the temperature, R_L the load resistance and B is the electrical bandwidth. Shot noise mean squared noise current can be modeled as $\langle i \rangle_{sh}^2 = 2q\Re P_s B$, where q is the electron charge, $\Re$ the responsivity of the photodetector and P_s the received optical power. Finally, dark current noise can be treated as $\langle i \rangle_{dk}^2 = 2qI_d B$, where I_d is the dark current of the photodetector.

The SNR of the O/E system results in

$$SNR = \frac{\langle I_s^2(t) \rangle}{\langle i \rangle_{th}^2 + \langle i \rangle_{sh}^2 + \langle i \rangle_{dk}^2 + \langle i \rangle_{amp}^2} \tag{A.1}$$

Using this expression, we can calculate the minimum detected power necessary to have a SNR = 10 dB, which is a conservative number. Using commercially available photodetectors [59] with $\Re = 0.85$ A/W, $I_d = 30$ nA, $B = 10$ GHz, $R_L = 50\ \Omega$ and reported TIAs [60] with input referred noise density of 16.7 pA/$\sqrt{\text{Hz}}$ we obtain that the minimum optical power that has to be photodetected is -21.6 dBm, or $6.96{*}10^{-3}$ mW. Using these values, we can estimate the number of layers that can be concatenated.

We consider two different scenarios:

Tunable tap: In this case, only the minimum amount of optical power necessary to achieve the desired value of the SNR at the nonlinear E/O system is separated from the incoming optical power. As the input optical power at each layer is different due to the losses at previous layers the corresponding coupling coefficient that divides the optical signal is tuned. The difference between the input optical power and the output optical power is equal to the amount of optical power separated at each layer:

$$L\ P_{act} = \left(\frac{P_{in}}{N} IL^L - P_{out} \right) \tag{A.2}$$

where P_{act} is the minimum optical power needed at the PD to achieve a certain SNR, P_{in} is the total input optical power at the system, it is divided by N as we assume that the optical power is equally distributed among all the outputs, P_{out} is the optical power at one output after L layers and IL are the insertion losses of the nonlinear E/O system in linear units.

Fix tap: In this case, the coupling coefficient is the same for all the layers. After L layers, the optical power will be

$$\frac{P_{in}}{N}(1 - k^L) = \frac{P_{in}}{N} IL^L - P_{out} \tag{A.3}$$

where k corresponds to the fixed coupling coefficient of the through port, that is, $(1 - k)\ P_i$ is separated at each E/O system, with P_i being the input optical power at each nonlinear stage. Equation A.3 is not sufficient to describe this option because we have to guarantee that the separated optical power is always equal or higher than the necessary to achieve the correspondent SNR. To satisfy this condition, we impose that the separated optical power at the L layer is equal to P_{act}

$$\frac{P_{in}}{N}(1-k)\,k^{L-1}\,IL^{L-1} = P_{act} \tag{A.4}$$

The final achievable number of layers is set as the minimum between the values obtained using (A.3) and (A.4).

In Fig. 13. it is shown the maximum number of layers that can be concatenated for the different scenarios for matrices of 32×32 and an input optical power of 40 mW. The horizontal axis represents the SNR at the output of the photonic neural network

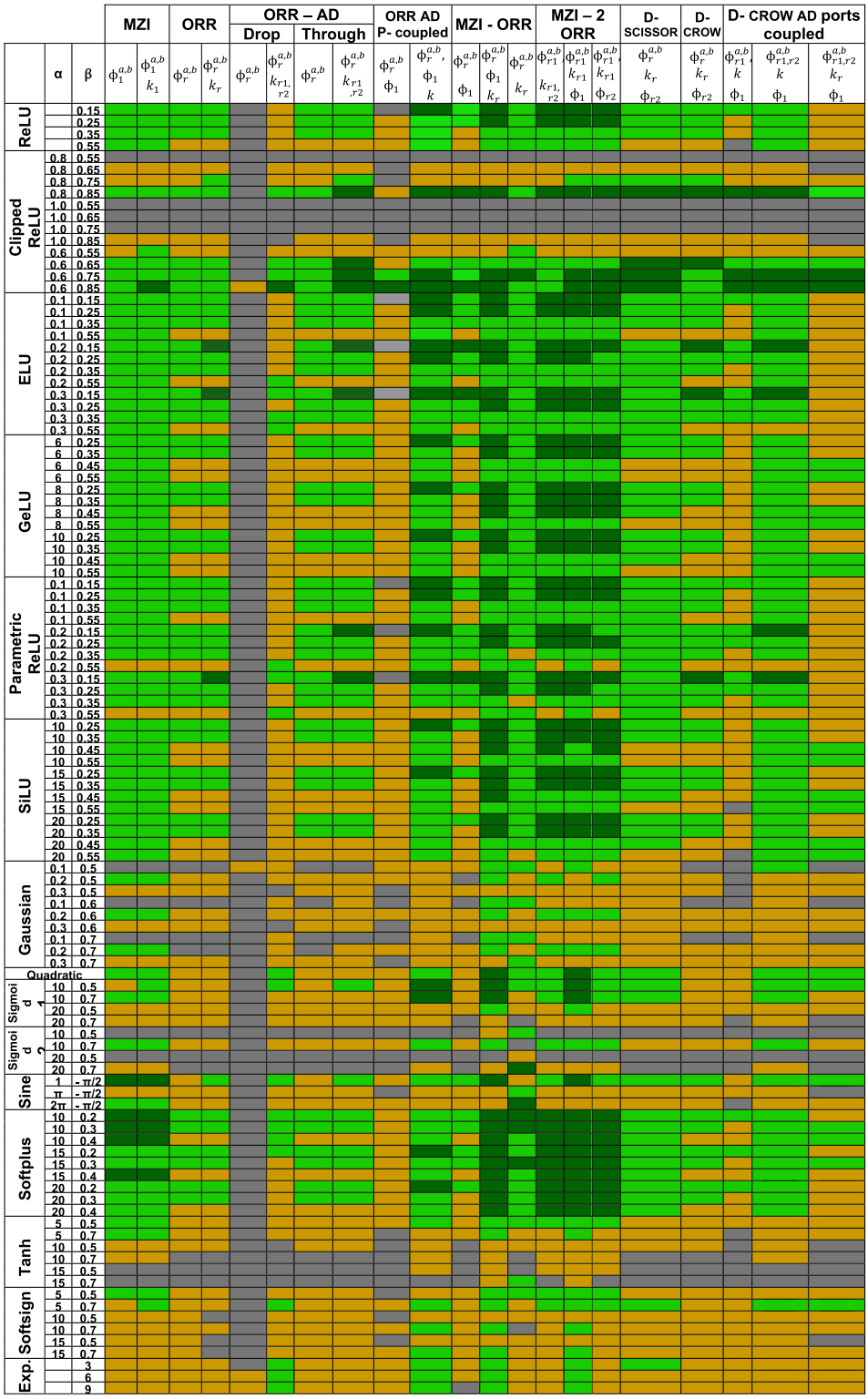

Fig. 12. Root mean square (rmse) between the response of the optical architectures and the activation functions. Values between 0 and −10 (grey), beetwen −10 and −15 (orange), −15 and −20 (light-green) and −20 to −25 (dark green) Results are given in dB.

which does not coincide with the SNR = 10 dB required at the nonlinear E/O system. We have assumed at the receiver an analog circuit similar to that used in the nonlinear architecture and the value of the necessary power at the output, P_{out}, has been calculated using the same expression relating power and SNR presented in this section. The vertical dotted green lines indicate the minimum SNR to achieve a certain bit resolution. We have used n = (SNR [dB] – 1.761)/6.02, where n is bit resolution. The IL of the E/O architecture has been assumed to be 0.3 dB

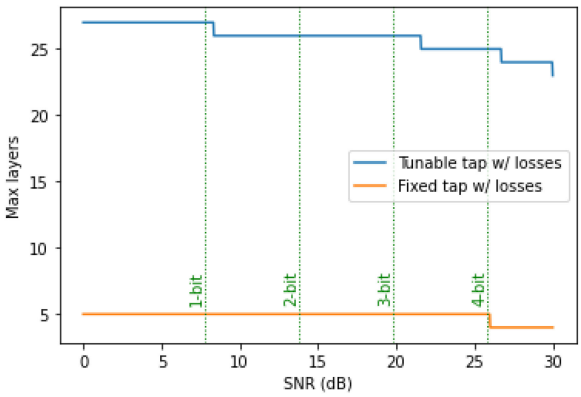

Fig. 13. Number of maximum layers achievable for the tunable tap (blue) and the fixed tap (orange) cases. It has been considered a matrix of size 32 x 32 with input power of 40 mW. The nonlinear E/O system has been assumed to have 0.3 dB of insertion losses. Vertical dotted green lines represent the SNR necessary to achieve the specified bit resolution. Loss-less components assumed in the matrix multipliers.

TABLE V
MAXIMUM NUMBER OF LAYERS. WE CONSIDER INPUT POWER OF 40 MW, ACTIVATION FUNCTION LOSS OF 0.3 DB, 0.3-DB MZI LOSS IN THE UNITARY SECTION, N = 1

	N = 4	N = 8	N = 16	N = 32
Activation function tap – Tunable	48	41	33	27
Activation function tap – Fixed	5	5	5	5
Unitary + activation – Tunable	13	7	3	1
Unitary + activation - Fixed	2	1	1	0

which is compatible with the commercially available photonic systems.

In the previous example, only the losses corresponding to the nonlinear E/O have been considered. We can compare these results with the case where the losses of the MZI based optical matrix multiplier are taken into account. We use a simple rule where each output has gone through N interferometers with insertion losses (IL_{MZI}) of 0.3 dB. After L layer, insertion losses have an impact of $N{\cdot}L{\cdot}IL_{MZI}$ on the input optical power. Table V compares the maximum number of layers achievable for different matrix sizes when only the nonlinear E/O stage loss is considered and when the nonlinear E/O system and the MZI based matrix multiplier loss are considered. The results show that the limiting part of the scalability of the system is the interferometric mesh.

The optical sensitivity of the system can be further improved using PDs with higher responsivities like avalanche photodetectors (APD) which increases the generated photocurrent. Thermal and dark current noise remains the same for equal load resistance and dark current but shot noise is enhanced. In our case, we are working on the thermal noise dominated regime (low incident optical power) so there is a benefit in using APD. High-speed and low noise APD have been reported in silicon foundry compatible systems [61]. Another alternative is the use of non-invasive detectors [62] which will eliminate the necessity to take a portion of the incoming light to trigger the self-made modulation.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[3] W. Xiong *et al.*, "Toward human parity in conversational speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2410–2423, Dec. 2017, doi: 10.1109/TASLP.2017.2756440.
[4] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016, doi: 10.1038/nature16961.
[5] H. Purwins, B. Sturm, B. Li, J. Nam, and A. Alwan, "Introduction to the issue on data science: Machine learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 203–205, Feb. 2019, doi: 10.1109/JSTSP.2019.2914321.
[6] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 7430, doi: 10.1038/s41598-017-07754-z.
[7] P. Stark, F. Horst, R. Dangel, J. Weiss, and B. J. Offrein, "Opportunities for integrated photonic neural networks," *Nanophotonics*, vol. 9, no. 13, pp. 4221–4232, 2020, doi: 10.1515/nanoph-2020-0297.
[8] N. C. Thompson, K. H. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," vol. abs/2007.0, 2020, *arXiv:2007.05558*.
[9] Z. Luo, H. Liu, and X. Wu, "Artificial neural network computation on graphic process unit," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2005, vol. 1, pp. 622–626, doi: 10.1109/IJCNN.2005.1555903.
[10] N. Jouppi, C. Young, N. Patil, and D. Patterson, "Motivation for and evaluation of the first tensor processing unit," *IEEE Micro*, vol. 38, no. 3, pp. 10–19, May./Jun. 2018, doi: 10.1109/MM.2018.032271057.
[11] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, Oct. 1990, doi: 10.1109/5.58356.
[12] D. A. B. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 2017, doi: 10.1109/JLT.2017.2647779.
[13] F.-C. F. Tsai, C. J. O'Brien, N. S. Petrović, and A. D. Rakić, "Analysis of optical channel cross talk for free-space optical interconnects in the presence of higher-order transverse modes," *Appl. Opt.*, vol. 44, no. 30, pp. 6380–6387, Oct. 2005, doi: 10.1364/AO.44.006380.
[14] J. W. Goodman, A. R. Dias, and L. M. Woody, "Fully parallel, high-speed incoherent optical method for performing discrete fourier transforms," *Opt. Lett.*, vol. 2, no. 1, pp. 1–3, 1978, doi: 10.1364/OL.2.000001.
[15] Y. Xie, L. Zhuang, and A. J. Lowery, "Picosecond optical pulse processing using a terahertz-bandwidth reconfigurable photonic integrated circuit," *Nanophotonics*, vol. 7, no. 5, pp. 837–852, 2018, doi: 10.1515/nanoph-2017-0113.
[16] D. Psaltis, D. Brady, X. - G. Gu, and S. Lin, "Holography in artificial neural networks," *Nature*, vol. 343, no. 6256, pp. 325–330, 1990, doi: 10.1038/343325a0.
[17] X. Lin *et al.*, "All-optical machine learning using diffractive deep neural networks.," *Science*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018, doi: 10.1126/science.aat8084.
[18] T. Zhou *et al.*, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nature Photon.*, vol. 15, no. 5, pp. 367–373, 2021, doi: 10.1038/s41566-021-00796-w.
[19] 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, pp. 21032, May 2019, doi: 10.1103/PhysRevX.9.021032.
[20] J. Wang, F. Sciarrino, A. Laing, and M. G. Thompson, "Integrated photonic quantum technologies," *Nature Photon.*, vol. 14, no. 5, pp. 273–284, 2020, doi: 10.1038/s41566-019-0532-1.
[21] D. Marpaung, J. Yao, and J. Capmany, "Integrated microwave photonics," *Nature Photon.*, vol. 13, no. 2, pp. 80–90, 2019, doi: 10.1038/s41566-018-0310-5.
[22] C. M. V. Burgos and N. Vamivakas, "Challenges in the path toward a scalable silicon photonics implementation of deep neural networks,"

IEEE J. Quantum Electron., vol. 55, no. 5, Oct. 2019, Art. no. 8400110, doi: 10.1109/JQE.2019.2934758.
[23] L. De Marinis, *et al.*, "Photonic integrated reconfigurable linear processors as neural network accelerators," *Appl. Sci.*, vol. 11, no. 13. 2021, Art. no. 6232, doi: 10.3390/app11136232.
[24] D. Perez *et al.*, "Silicon photonics rectangular universal interferometer," *Laser Photon. Rev.*, vol. 11, no. 6, Nov. 2017, Art. no. 1700219. [Online]. Available: https://doi.org/10.1002/lpor.201700219
[25] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, 2017, doi: 10.1038/nphoton.2017.93.
[26] J. K. George *et al.*, "Neuromorphic photonics with electro-absorption modulators," *Opt. Exp.*, vol. 27, no. 4, pp. 5181–5191, Feb. 2019, doi: 10.1364/OE.27.005181.
[27] A. N. Tait *et al.*, "Silicon photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, Jun. 2019, Art. no. 64043, doi: 10.1103/PhysRevApplied.11.064043.
[28] I. A. D. Williamson, *et al.*, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7700412, doi: 10.1109/JSTQE.2019.2930455.
[29] A. Majumdar and A. Rundquist, "Cavity-enabled self-electro-optic bistability in silicon photonics," *Opt. Lett.*, vol. 39, no. 13, pp. 3864–3867, 2014, doi: 10.1364/OL.39.003864.
[30] G. Mourgias-Alexandris, *et al.*, "An all-optical neuron with sigmoid activation function," *Opt. Exp.*, vol. 27, no. 7, pp. 9620–9630, 2019, doi: 10.1364/OE.27.009620.
[31] J. Crnjanski, M. Krstić, A. Totović, N. Pleros, and D. Gvozdić, "Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron," *Opt. Lett.*, vol. 46, no. 9, pp. 2003–2006, May 2021, doi: 10.1364/OL.422930.
[32] M. Miscuglio *et al.*, "All-optical nonlinear activation function for photonic neural networks," *Opt. Mater. Exp.*, vol. 8, no. 12, pp. 3851–3863, 2018, doi: 10.1364/OME.8.003851.
[33] A. Jha, C. Huang, and P. R. Prucnal, "Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics," *Opt. Lett.*, vol. 45, no. 17, pp. 4819–4822, Sep. 2020, doi: 10.1364/OL.398234.
[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[35] D. - A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. 4th Int. Conf. Learn. Representations*, 2016, *arXiv preprint arXiv:1511.07289*.
[36] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis .*, 2015, pp. 1026–1034.
[38] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, 2018.
[39] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–4, doi: 10.1109/IJCNN.2015.7280459.
[40] J. Turian, J. Bergstra, and Y. Bengio, "Quadratic features and deep architectures for chunking," in *Proc. Hum. Lang. Technol.: 2009 Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Companion Volume: Short Papers*, 2009, pp. 245–248, [Online]. Available: https://aclanthology.org/N09-2062
[41] A. I. Stanley, G. Singh, J. Eke, and H. Tsuda, "Mach–Zehnder interferometer: A review of a perfect all-optical switching structure," in *Proc. Int. Conf. Recent Cognizance Wireless Commun. Image Process.*, 2016, pp. 415–425.
[42] W. Bogaerts *et al.*, "Silicon microring resonators," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, Jan. 2012. [Online]. Available: https://doi.org/10.1002/lpor.201100017
[43] D. Pérez *et al.*, "Multipurpose silicon photonics signal processor core," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 636, doi: 10.1038/s41467-017-00714-1.
[44] D. Pérez-López, A. López, P. DasMahapatra, and J. Capmany, "Multipurpose self-configuration of programmable photonic circuits," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 6359, doi: 10.1038/s41467-020-19608-w.
[45] I. A. D. Williamson, *et al.*, "Tunable nonlinear activation functions for optical neural networks," in *Proc. Conf. Lasers Electro-Opt.*, 2020, pp. 1–2, doi: 10.1364/CLEO_SI.2020.SM1E.2.
[46] R. Boeck, *Silicon Ring Resonator Add-Drop Multiplexers*, Vancouver, Canada: The Univ. of British Columbia, 2011.
[47] J. M. Fargas Cabanillas *et al.*, "Broadband repeatable <0.025 dB average loss rapid adiabatic based 3-dB coupler in a 45 nm SOI CMOS process," in *Proc. OSA Adv. Photon. Congr.*, 2021, Art. no. JTh3A.3, [Online]. Available: http://opg.optica.org/abstract.cfm?URI=NOMA-2021-JTh3A.3
[48] Z. Sheng *et al.*, "A compact and low-loss MMI coupler fabricated with CMOS technology," *IEEE Photon. J,* vol. 4, no. 6, pp. 2272–2277, Dec. 2012, doi: 10.1109/JPHOT.2012.2230320.
[49] Y. Zhang *et al.*, "Design and demonstration of ultra-high-Q silicon microring resonator based on a multi-mode ridge waveguide," *Opt. Lett.*, vol. 43, no. 7, pp. 1586–1589, 2018, doi: 10.1364/OL.43.001586.
[50] B. Haq *et al.*, "Micro-transfer-printed III-V-on-Silicon C-Band semiconductor optical amplifiers," *Laser Photon. Rev.*, vol. 14, no. 7, Jul. 2020, Art. no. 1900364, [Online]. Available: https://doi.org/10.1002/lpor.201900364
[51] Y. Xing, D. Spina, A. Li, T. Dhaene, and W. Bogaerts, "Stochastic collocation for device-level variability analysis in integrated photonics," *Photon. Res.*, vol. 4, no. 2, pp. 93–100, 2016, doi: 10.1364/PRJ.4.000093.
[52] T. - W. Weng, *et al.*, "Uncertainty quantification of silicon photonic devices with correlated and non-Gaussian random parameters," *Opt. Exp.*, vol. 23, no. 4, pp. 4242–4254, 2015, doi: 10.1364/OE.23.004242.
[53] D. Pérez and J. Capmany, "Scalable analysis for arbitrary photonic integrated waveguide meshes," *Optica*, vol. 6, no. 1, pp. 19–27, 2019, doi: 10.1364/OPTICA.6.000019.
[54] M. Abadi *et al.*, "{TensorFlow}: Large-scale machine learning on heterogeneous systems." 2015 [Online]. Available: https://www.tensorflow.org/
[55] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.
[56] H. Zhang *et al.*, "An optical neural chip for implementing complex-valued neural network," *Nature Commun.*, vol. 12, no. 1, 2021, Art. no. 457, doi: 10.1038/s41467-020-20719-7.
[57] A. Ryou *et al.*, "Free-space optical neural network based on thermal atomic nonlinearity," *Photon. Res.*, vol. 9, no. 4, pp. B128–B134, 2021, doi: 10.1364/PRJ.415964.
[58] Y. Zuo *et al.*, "All-optical neural network with nonlinear activation functions," *Optica*, vol. 6, no. 9, pp. 1132–1137, 2019, doi: 10.1364/OPTICA.6.001132.
[59] S. Y. Siew *et al.*, "Review of silicon photonics technology and platform development," *J. Lightw. Technol.*, vol. 39, no. 13, pp. 4374–4389, Jul. 2021, doi: 10.1109/JLT.2021.3066203.
[60] K. R. Lakshmikumar *et al.*, "A process and temperature insensitive CMOS linear TIA for 100 Gb/s/PAM-4 optical links," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3180–3190, Nov. 2019, doi: 10.1109/JSSC.2019.2939652.
[61] D. Benedikovic *et al.*, "Silicon-Germanium avalanche receivers with fJ/bit energy consumption," *IEEE J. Sel. Top. Quantum Electron.*, vol. 28, no. 2, Mar./Apr. 2022, Art. no. 3802508, doi: 10.1109/JSTQE.2021.3112494.
[62] F. Morichetti *et al.*, "Non-Invasive on-chip light observation by contactless waveguide conductivity monitoring," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 4, pp. 292–301, Jul./Aug. 2014, doi: 10.1109/JSTQE.2014.2300046.

José Roberto Rausell Campo received the B.S. and M.S. degrees in physics from the Universitat de València, Valencia, Spain, in 2018 and 2019, respectively. He is currently working toward the Ph.D. degree in telecommunications with the Universitat Politècnica de València, Valencia, Spain. His research interests include nonlinear optics, optical information processing, and machine learning.

Daniel Pérez-López (M.Sc., Ph.D.), has spent his academic life on the creation, development of a programmable photonics: such as, giving to large-scale photonic integrated circuits flexibility through run-time software programming. His career started with his PhD and a two-year postdoc research with the Universitat Politècnica de València's Photonics Research Labs, where he investigated novel photonic integrated circuit (PIC) designs and architectures that can be reconfigured to perform multipurpose functionalities. These works led to the 2017 Graduate Student Fellowship by the IEEE Photonics Society and 2018 Extraordinary PhD Thesis Award from UPVLC. During 2019—2020, he was with Xanadu Quantum Computing, Canada, and founded iPronics Programmable Photonics S.L, Spain. In 2019 and 2021, he was the recipient of the prestigious Juan de la Cierva Fellowship grants that allowed him to initiate his research on high-dense large-scale programmable photonic integrated circuits and its applications to signal processing and computing.