Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells

George Mourgias-Alexandris, Angelina Totović, Apostolos Tsakyridis, Nikolaos Passalis,
Konstantinos Vyrsokinos, Anastasios Tefas, and Nikos Pleros

***Abstract*—Neuromorphic photonics aims to transfer the high-bandwidth and low-energy credentials of optics into neuromorphic computing architectures. In this effort, photonic neurons are trying to combine the optical interconnect segments with optics that can realize all critical constituent neuromorphic functions, including the linear neuron stage and the activation function. However, aligning this new platform with well-established neural network training models in order to allow for the synergy of the photonic hardware with the best-in-class training algorithms, the following requirements should apply: i) the linear photonic neuron has to be able to handle both positive and negative weight values, ii) the activation function has to closely follow the widely used mathematical activation functions that have already shown an enormous performance in demonstrated neural networks so far. Herein, we demonstrate a coherent linear neuron architecture that relies on a dual-IQ modulation cell as its basic neuron element, introducing distinct optical elements for weight amplitude and weight sign representation and exploiting binary optical carrier phase-encoding for positive/negative number representation. We present experimental results of a typical IQ modulator performing as an elementary two-input linear neuron cell and successfully implementing all-optical linear algebraic operations with 104-ps long optical pulses. We also provide the theoretical proof and formulation of how to extend a dual-IQ modulation cell into a complete *N*-input coherent linear neuron stage that requires only a single-wavelength optical input and avoids the resource-consuming Wavelength Division Multiplexing (WDM) weighting schemes. An 8-input coherent linear neuron is then combined with an experimentally validated optical sigmoid activation function into a physical layer simulation environment, with respective training and physical layer simulation results for the MNIST dataset revealing an average accuracy of 97.24% and 94.37%, respectively.**

***Index Terms*—All-optical signal processing, neural networks, neuromorphic computing, neuromorphic photonics, optical neural network accelerators.**

I. Introduction

With Moore's and Koomey's laws starting to decline [1], neuromorphic computing appears as a highly promising candidate for sustaining computational advances and overcoming the digital energy efficiency wall of Von-Neumann architectures. Recent state-of-the-art neuromorphic deployments confirm the huge potential of non-Von-Neumann brain-inspired layouts like IBM's TrueNorth [2], SpiNNaker [3] and Intel's Loihi [4] to rapidly increase the energy efficiency by means of million neurons. Migrating from von-Neumann into brain-inspired computing paradigms are still carried out by investigating and investing in alternative enabling technologies, aiming to optimally synergize performance and energy benefits offered by both the architectural and technological fields. Inspired by the well-known speed and energy benefits of photonics that are gradually turning interconnection into the stronghold of optical technologies [5]–[8], recent research efforts are already attempting to transfer the neuromorphic computing principles over optics [9], [10]. This has led to the introduction of neuromorphic photonics [9]–[11] as a new scientific area, indicating already its huge energy and footprint efficiency perspectives in case of successfully transferring the large bandwidth advantages of onto a neural network operational platform [11]. First training attempts of photonic neural networks highlight the credentials to form reliable neuromorphic machines [12]–[14], revealing negligible accuracy degradation by using photonic activation functions with fabrication tolerances.

This transfer requires, however, the deployment of all necessary neuromorphic functions as optically-enabled building blocks, ensuring at the same time that these can yield a photonic circuit infrastructure compatible with the well-established neural network training framework. As such, the linear neuron stage that is responsible for carrying out the linear algebra operations has to support both positive and negative weight representations, while the activation stage should ideally conform to widely used mathematical functions like ReLU, sigmoid, tanh etc. [15]–[17]. To cope with the need for both positive/negative weights, the weighting layouts that have been proposed so far have mainly relied on WDM schemes [10], [18], [19], encoding every input

Manuscript received June 18, 2019; revised September 19, 2019; accepted October 18, 2019. Date of publication October 23, 2019; date of current version February 12, 2020. This work was supported in part by the EC through H2020 Projects ICT-MASSTART under Contract 825109 and in part by the ICT-PLASMONIAC under Contract 871391. *(Corresponding author: George Mourgias-Alexandris.)*

G. Mourgias-Alexandris, A. Totović, A. Tsakyridis, A. Tefas, and N. Pleros are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54621, Greece (e-mail: mourgias@csd.auth.gr; angelina@auth.gr; atsakyrid@csd.auth.gr; tefas@csd.auth.gr; npleros@csd.auth.gr).

N. Passalis is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54621, Greece. He is now with the Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland (e-mail: passalis@csd.auth.gr).

K. Vyrsokinos is with the Department of Physics, Aristotle University of Thessaloniki, Thessaloniki 54621, Greece (e-mail: kv@auth.gr).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2019.2949133

signal onto a different wavelength and using a pair of balanced photodiodes (PDs) to realize the summation of positive and negative weights. However, besides requiring a high number of wavelengths as the neuron fan-in increases, this scheme necessitates the use of optoelectronic conversion and a balanced PD followed by an optical modulator [10] for carrying out the linear algebraic summation after converting the optical signals in the electrical domain. This impedes the employment of alternative all-optical non-linear [20] or sigmoid [21], [22] activation functions, requiring the deployment of new training frameworks to accommodate non-typical activation functions, like the modulator's sinusoidal response [14]. An alternative linear photonic neuron scheme that is potentially compatible with non-sinusoidal all-optical activation functions has been presented in [18], employing complementary data onto different wavelengths to realize the negative weights. Hence, this wavelength-encoding scheme can in principle eliminate the need for optical-to-electrical-to-optical conversion, but the scaling of fan-in still requires at least two different lasers for positive/negative weight representation [18]. Coherent layouts that exploit the phase of the optical carrier electric field for sign encoding purposes can yield single-wavelength and single-laser linear neuron deployments, but have been demonstrated so far only in a rather complex spatial layout for matrix multiplication purposes with multiple cascaded Mach-Zehnder Interferometers (MZIs) [9]. This design follows the Reck-proposal [9] and requires a N^2 number of MZIs for an N-input configuration, scaling quadratically with the fan-in value.

In this paper, we propose a single-wavelength Coherent Optical Linear Neuron (COLN) that relies on a dual-IQ modulator as its elementary 2-input cell, scales linearly with the number of inputs and can be effectively synergized with all-optical activation functions that closely follow typical mathematical functions used in deep learning models. The dual-IQ modulator cell comprises a typical IQ modulator structure where, however, every interferometric arm includes both the I- and Q-modulation stage, yielding a layout that almost merges the two IQ modulators into the same MZI structure. In this way, every MZI branch acts as a complete axon that offers electrical-to-optical input signal conversion via the I-modulation stage and weighting via the Q-modulation stage, separating the weight amplitude and weight sign encoding processes. We demonstrate experimentally the underlying linear algebra principles of a typical IQ modulator in weighted optical addition and subtraction operations between two optical signals at 10 Gbaud/s. As a result, the mathematical framework is formulated towards the mathematical confirmation of its scaling capabilities to N-input coherent arrangement. Finally, a complete and properly trained 8-input coherent optical linear neuron combined with an experimentally verified all-optical sigmoid activation function recently proposed by our group in [21] is then implemented in a VPIphotonics-based physical layer simulation framework. Successful demonstration of the complete neuron at 10 Gbaud/s for the MNIST dataset reveals an average accuracy of 94.37%, i.e. only 2.97% lower compared to the ideal software-based training and performance evaluation process.

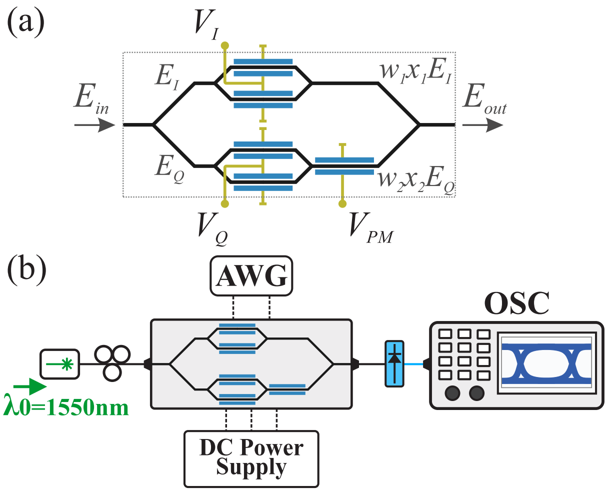

Fig. 1. (a) Typical layout of an optical IQ modulator. (b) The experimental setup for the emulation of the 2-input COLN.

II. The IQ Modulator as a 2-Input Optical Linear Algebra Unit

Fig. 1 depicts a typical IQ modulator that employs two push-pull MZI structures and a phase shifter for the I and Q component modulation, with the device being fed by an optical input signal E_{in}. The I and Q MZIs are driven by the V_I and V_Q differential voltages imprinting on the real part of E_I and E_Q the I and Q signals, respectively, with the phase shifter at its Q branch being controlled by a V_{PM} voltage level to achieve the orthogonality between I and Q via a $\pi/2$ phase shift. Taking advantage of the strong electro-optic synthesis portfolio of IQ modulators [23] and enforcing via the phase shifter a phase shift of $\phi_s = 0$ or $\phi_s = \pi$ instead of $\pi/2$, the IQ modulator can be transformed into an elementary linear algebra cell capable of performing algebraic operations between two numbers. Assuming that the V_I and V_Q modulator voltages result in multiplying the respective modulator's incoming field with a factor of w_1x_1 and w_2x_2, respectively, where $w_1x_1 = \sin(\frac{\pi V_I}{2V_\pi})$ and $w_2x_2 = \sin(\frac{\pi V_Q}{2V_\pi})$ the IQ modulator output can be expressed as:

$$E_{out} = \frac{1}{2}E_{in}\left[w_1x_1 + w_2x_2e^{j\phi_s}\right] = \frac{1}{2}E_{in}\left[w_1x_1 \pm w_2x_2\right] \tag{1}$$

where w_i, x_i denote the weight and the input value, respectively, of each input. (1) shows that a typical IQ-modulator can perform either addition or subtraction between the two numbers w_1x_1 and w_2x_2, respectively, or, alternatively, can perform the algebraic addition with ϕ_s denoting whether the w_2x_2 value is positive or negative.

To validate experimentally the ability of a typical IQ modulator to act as an elementary linear algebra module between two numbers w_1x_1 and w_2x_2, as expressed in (1), we have used the experimental setup shown in Fig. 1(b). A laser beam at $\lambda_0 = 1550\,\text{nm}$ is injected into a LiNbO$_3$ IQ modulator where both the I and Q MZI structures are driven by two electrical

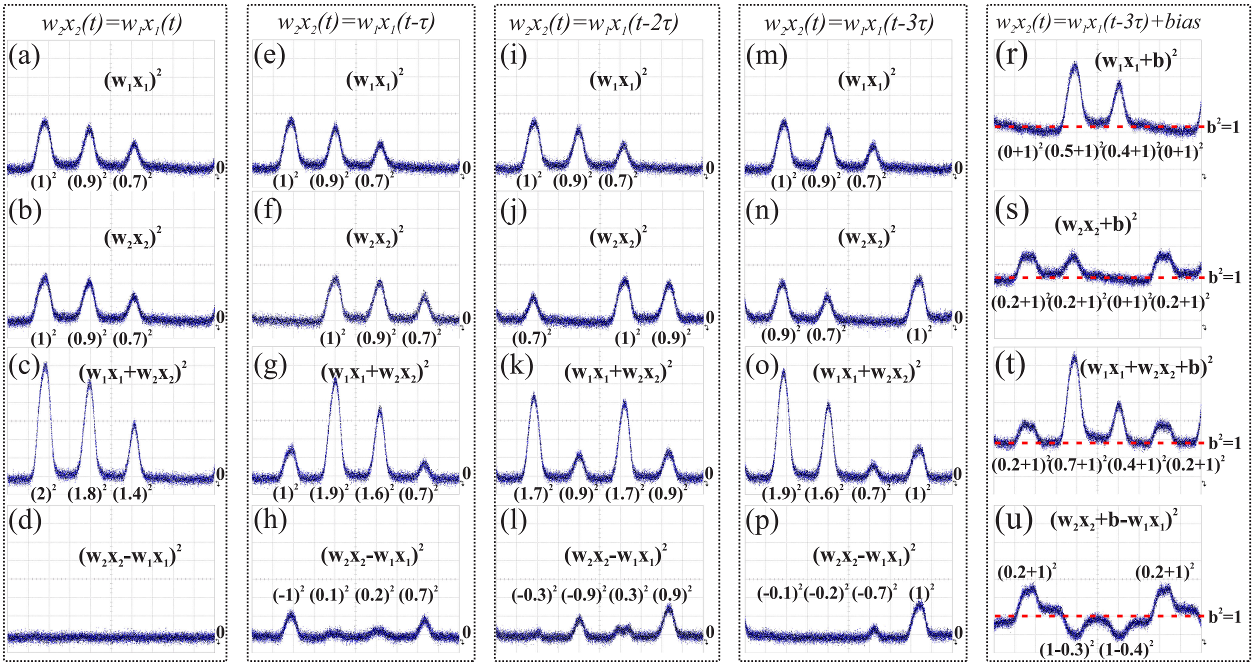

Fig. 2. Time traces of the initial w_1x_1 and w_2x_2 signals generated by I and Q MZIs, summed and subtracted when: (a)–(d) $w_2x_2(t) = w_1x_1(t)$, (e)–(h) $w_2x_2(t) = w_1x_1(t - \tau)$, (i)–(l) $w_2x_2(t) = w_1x_1(t - 2\tau)$, (m)–(p) $w_2x_2(t) = w_1x_1(t - 3\tau)$, and (r)–(u) $w_2x_2(t) = w_1x_1(t - 3\tau) + bias$. ($x$-axis scale: 100 psec/div, y-axis scale: 2.5 mV/div).

Gaussian pulse-shaped sequences with a 104-ps pulse width, generated by an Arbitrary Waveform Generator (AWG) to emulate the w_ix_i products. The respective DC power supplies are employed in order to properly bias the I and Q MZIs as well as to define the 0 or π phase shift at the phase shifter, determining in this way whether addition or subtraction will be carried out between w_1x_1 and w_2x_2. Finally, the IQ modulator output signal is fed into a PD and is monitored via an oscilloscope for signal evaluation purposes.

The pulse traces that were captured at the PD electrical output for different cases of I and Q modulated pulse sequences are illustrated in Fig. 2. Given that the PD output provides the equivalent optical power of the respective pulse while the linear algebra operations are carried out on the electrical field, the square root of the monitored intensity has to be used to estimate the electrical field value and validating the result of the algebraic operation. Every column of Fig. 2 illustrates a different set of two input signals w_1x_1 and w_2x_2 corresponding to the $(w_1x_1)^2$ and $(w_2x_2)^2$ optical powers that are shown in the first and second row and are generated at the output of the I and Q MZI structures, respectively, depicting at the third and fourth row within every column the squared addition and subtraction outcomes, respectively. More specifically, Fig. 2(a) and (b) depict the traces of $(w_1x_1)^2$ and $(w_2x_2)^2$ in the case when both $(w_1x_1)^2$ and $(w_2x_2)^2$ are identical. Normalizing every pulse sequence to its highest peak power, each of the w_1x_1 and w_2x_2 pulse sequences reveals electric field peak amplitudes of 1, 0.9 and 0.7 for its three constituent pulses. Inducing a 0 or π phase shift by means of the phase shifter, coherent addition or subtraction is achieved resulting in $(w_1x_1 + w_2x_2)^2$ and $(w_2x_2 - w_1x_1)^2$ at the output of the PD, as can be seen in Fig. 2(c) and (d), respectively. Normalizing again with respect to the highest input pulse peak power, Fig. 2(c) confirms the successful coherent addition of w_1x_1 and w_2x_2 products, as the ratio between the intensities of the $(w_1x_1 + w_2x_2)^2$ output pulse peak powers is identical to the respective ratios of the constituent $(w_1x_1)^2$ and $(w_2x_2)^2$ pulse peak powers. Fig. 2(d) shows clearly the successful result of $(w_2x_2 - w_1x_1)^2$ that equals to zero in this case. The second column depicts the case when using the same $(w_1x_1)^2$ and $(w_2x_2)^2$ signals with $(w_2x_2)^2$ being delayed by one pulse period with respect to the $(w_1x_1)^2$ pulse sequence, as shown in Fig. 2(e) and (f), respectively. Following the same procedure as explained for Fig. 2(a)–(d), successful coherent addition and subtraction can be again verified through the output pulse traces illustrated in Fig. 2(g) and (h), respectively. The third and the fourth column of Fig. 2 illustrate two additional scenarios where the $(w_2x_2)^2$ is delayed by two [Fig. 2(j)] and three [Fig. 2(n)] pulse periods, respectively, with respect to the $(w_1x_1)^2$ signal. In both cases, successful addition and subtraction has been obtained as can be verified by Fig. 2(k) and (l) and 2(o) and (p), respectively. It should be noted that the results depicted in Fig. 2(c), (g), (k), (o) and (t) have been obtained by using a different normalization factor, so that, for example, the pulse amplitude of 2^2 in Fig. 2(c) is not four times the 1^2 in Fig. 2(b). This has been the result of applying different attenuation for the different traces prior monitoring them on the oscilloscope, so as to ensure that their optical power levels will be within the operational range of the PD. Concerning the distorted pulses of Fig. 2(l), (s) and (u), the sinusoidal transfer function of MZMs results in flat-top pulses with spikes on their rise and fall edges that stem from the subtraction of optical pulses with slight differences in their pulse shape, which in turn depend on whether the MZM is

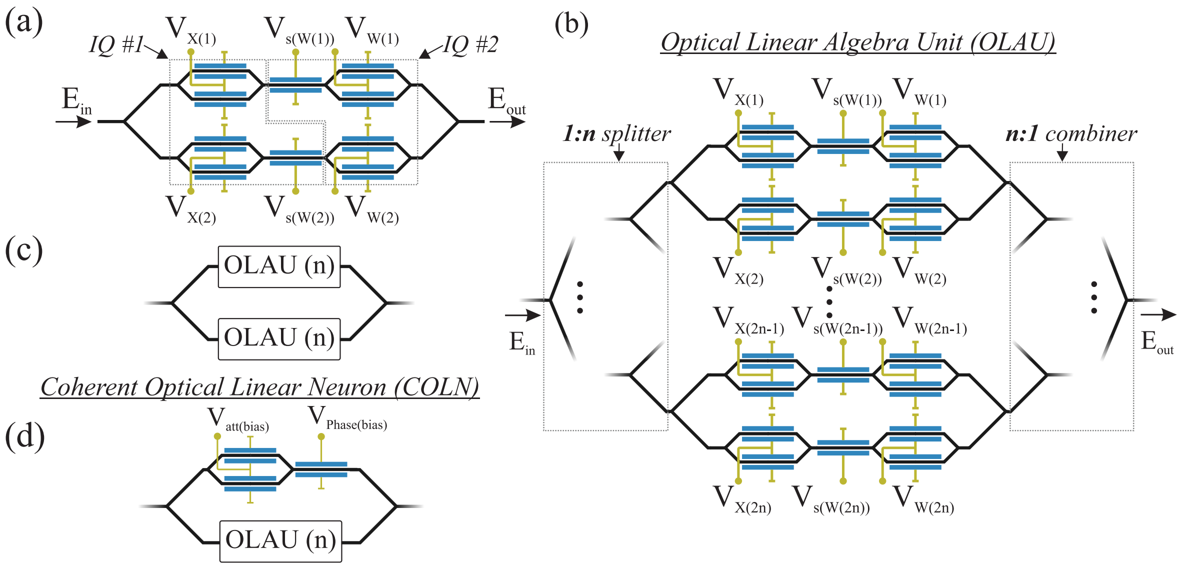

Fig. 3. Layout of the: (a) dual-IQ modulator, (b) $n = 2^m$ OLAU circuit, (c) $n = 2^{(m+1)}$ OLAU circuit, and (d) $n = 2^m$ OLAU along with the *bias* branch, forming a COLN.

operated at its linear or non-linear regime. Regarding the weight accuracy, the employment of IQ modulators towards implementing photonic linear neurons enables also the transfer of well-established techniques in optical communications for time alignment [24] and amplitude-accurate signal generation [25]. Additional improvements in weight accuracy can be eventually accomplished by exploiting also calibration methods utilized for fine-tuning the MRRs of a WDM weight bank [26].

However, this operational arrangement can only provide the absolute value of the subtraction through the signal's power, as can be clearly seen in Fig. 2(l) and (p), with the sign of the resulting difference staying concealed in the phase of the optical carrier. To reveal the sign of the difference in the optical power domain, one of the weighted input signals $w_i x_i$ needs to be superimposed onto a DC biasing power level, denoted as b, suggesting that at least one of the IQ module branches has to produce an electrical field proportional to $w_i x_i + b$ instead of the $w_i x_i$ value considered in the theoretical analysis summarized in (1). In this way, (1) can be rewritten as:

$$E_{out} = \frac{1}{2} E_{in} \left[(w_2 x_2 + b) \pm w_1 x_1 \right] \tag{2}$$

Where we have assumed that the DC bias level has been enforced at the Q-MZI branch that generates $w_2 x_2$. This can be easily obtained by biasing the respective Q modulation stage close to its quadrature point. The experimental results for this case are depicted in the fifth column of Fig. 2. Fig. 2(r) and (s) illustrate the optical power levels corresponding to the $(w_1 x_1 + b)^2$ when $w_2 x_2 = 0$ and $(w_2 x_2 + b)^2$ when $w_1 x_1 = 0$, respectively. Normalizing the pulse sequence to the power level of the DC biasing signal, the coherent addition between these signals can be successfully confirmed via Fig. 2(t) where obviously all pulses are atop the biasing beam. The coherent subtraction when enforcing a π phase shift at the Q-branch is shown in Fig. 2(u), where indeed the positive differences are imprinted as pulses atop the biasing beam but the negative $w_2 x_2 - w_1 x_1$ differences emerge as inverted pulses below the power level of the biasing beam.

III. The Coherent Optical Linear Neuron

Having validated the linear algebraic credentials of a simple IQ modulator, its layout can be expanded to support the individual control of the w_i and x_i values within the $w_i x_i$ product while controlling also the sign of both $w_i x_i$ product values by simply duplicating the I and Q modulation segments within the MZI structure. This yields a dual-IQ design as shown in Fig. 3(a), where two additional MZI modulators and an additional phase shifter have been incorporated within the initial IQ modulator cell, so that every MZI branch includes both an I and a Q modulation segment. The x_i and w_i values are defined by the two respective MZI modulators at each branch so that $x_i = \sin(\frac{\pi V_{x(i)}}{2V_\pi})$ and $w_i = \sin(\frac{\pi V_{w(i)}}{2V_\pi})$ with $i = 1, 2$, whereas $V_{x(i)}$ and $V_{w(i)}$ are the voltages applied to the respective MZI modulators and V_π the characteristic modulator voltage required for obtaining a π phase shift. The sign of the weight is determined by the phase modulator in the corresponding MZI branch through enforcing a phase shift $\phi_{s(w_{(i)})} = \pi \frac{V_{s(w(i))}}{V_\pi}$, with $V_{s(w(i))}$ being the driving voltage of the phase modulator. Using these expressions and representing $\Delta\phi_{x(i)} = \pi \frac{V_{x(i)}}{V_\pi}$, and $\Delta\phi_{w(i)} = \pi \frac{V_{w(i)}}{V_\pi}$, the output of this dual-IQ modulator structure can be then calculated to be:

$$\begin{aligned} E_{out} &= \frac{1}{2} E_{in} \left[\sin\left(\frac{\Delta\phi_{x_1}}{2}\right) \sin\left(\frac{\Delta\phi_{w_1}}{2}\right) e^{j\phi_{s(w_1)}} \right. \\ &\quad \left. + \sin\left(\frac{\Delta\phi_{x_2}}{2}\right) \sin\left(\frac{\Delta\phi_{w_2}}{2}\right) e^{j\phi_{s(w_2)}} \right] \\ &= \frac{1}{2} E_{in} \sum_{i=1}^{2} w_i x_i e^{j\phi_{s(w_i)}} \end{aligned} \tag{3}$$

This reveals that the dual-IQ module acts as a 2-input Optical Linear Algebra Unit (OLAU) where the optical x_i signals are optically multiplied by respective w_i weight values with the weight sign enforced via the optical phase shift $\phi_{s(w_i)}$ that can be either 0 or π. It should be noted that although the above analysis describes the dual-IQ cell as exploiting electro-optic modulation for all the x_i, w_i and ϕ_s modulation segments, the typically slow time scales of changes required at the weight values can in principle allow for the x_i and ϕ_s modulation stages to rely on thermo-optic modulation mechanism [27].

Scaling this design into an OLAU that supports a higher fan-in can be accomplished via the layout shown in Fig. 3(b), where $n = 2^m$ parallel dual-IQ cells have been incorporated into an interferometric arrangement formed between a passive $1 : n$ split and a passive $n : 1$ recombine stage, accomplished through cascading Y-junction splitters/combiners, with m being a positive integer number. This layout supports a total fan-in equal to $2n = 2^{(m+1)}$ and in the following will prove that the response of OLAU follows the relationship:

$$E_{out}(n = 2^m) = \frac{1}{2n} E_{in} \sum_{i=1}^{2n} w_i x_i e^{j\phi_{s(w_i)}}$$
$$= \frac{1}{2^{(m+1)}} E_{in} \sum_{i=1}^{2^{(m+1)}} w_i x_i e^{j\phi_{s(w_i)}} \quad (4)$$

by following the induction method, where w_i, x_i and $\phi_{s(w_i)}$ denote the weight amplitude, the input signal and the enforced phase shift at each branch within the dual-IQ structures, respectively. For $n = 1$, i.e., $m = 0$, (4) obviously describes the response of an elementary 2-input OLAU, i.e., one dual-IQ module, as it reduces to the already validated (3). Assuming that (4) is valid for $n = 2^m$, then for $n = 2^{(m+1)}$ the response of the respective configuration shown in Fig. 3(c) is given by:

$$E_{out}(n = 2^{(m+1)}) = \frac{1}{2}\left[\frac{1}{2^{(m+1)}} E_{in} \sum_{i=1}^{2^{(m+1)}} w_i x_i e^{j\phi_{s(w_i)}} + \frac{1}{2^{(m+1)}} E_{in} \sum_{i=2^{(m+1)}+1}^{2^{(m+2)}} w_i x_i e^{j\phi_{s(w_i)}}\right]$$
$$= \frac{1}{2} E_{in} \frac{1}{2^{(m+1)}} \sum_{i=1}^{2^{(m+2)}} w_i x_i e^{j\phi_{s(w_i)}}$$
$$= \frac{1}{2n} E_{in} \sum_{i=1}^{2n} w_i x_i e^{j\phi_{s(w_i)}} \quad (5)$$

This confirms the validity of (4) and proves that the layout of Fig. 3(c) acts as an OLAU for a number of $2n$ dual-IQ modulators, resulting in $4n$ inputs. In order to equip this structure with a discrete bias signal b avoiding the need for operating one of the modulation branches at its quadrature point, as was demonstrated in Section II, the whole OLAU design is incorporated into an additional MZI unit as one of its branches, with its second branch comprising a simple MZM followed by a phase modulator for defining the bias amplitude $w_b = |b|$ and its phase ϕ_b, used for compensating the optical path lengths deviation between bias and OLAU branches, resulting in a COLN as shown in Fig. 3(d). Analyzing COLN mathematically, it can be shown that its response is given by:

$$E_{out}(n = 2^m) = \frac{1}{2} E_{in} \left[w_b e^{j\phi_b} + \frac{1}{2n} \sum_{i=1}^{2n} w_i x_i e^{j\phi_{s(w_i)}} \right] \quad (6)$$

Equation (6) designates that the structure of Fig. 3(d) operates as a COLN offering both the weighted summation of the $2n$ input signals as well as its summation with an individually controlled biasing signal that allows to imprint the positive and negative summations as power signals formed beyond or below the DC biasing beam, respectively. To this end, this unit can be fully compatible with all-optical nonlinear activation modules without requiring sophisticated electro-optical schemes like balanced photodetection modules while supporting operation with a single optical wavelength at its input. It is also important to note that this COLN design scales linearly with the neuromorphic layout fan-in, requiring for N inputs a total number of $2N$ amplitude modulating MZIs and N phase modulating MZI within its OLAU unit plus one MZI and one phase modulator at its biasing branch. As such, it requires in total a number of $3N + 2$ phase shifting elements, implying significant benefits compared to the N^2-scaling architectures [9] suggested so far for lower losses and lower power consumption requirements.

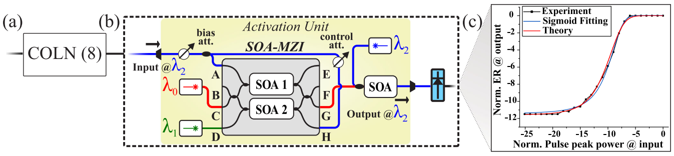

Fig. 4. (a) An 8-input COLN followed by a (b) Sigmoid activation unit [21] with its transfer function being illustrated in (c).

IV. Physical Layer Analysis of a Trained All-Optical Coherent Sigmoid Optical Neuron at 10 Gbaud/s

The linear photonic neuron of Fig. 3(d) has been validated via physical layer simulations in the VPIphotonics Design Suite using four parallel dual-IQ modulation cells that correspond to a total number of 8 electrical $V_{x(i)}$ input signals. The 8-input COLN was arranged to perform as a trained all-optical sigmoid neuron for the MNIST benchmarking dataset, having its output connected to an all-optical sigmoid activation module recently demonstrated experimentally by us in [21] and having its weights determined by an off-line training process targeting the recognition of handwritten digits $(0 - 9)$. The VPI implementation, schematically illustrated in Fig. 4(a) and (b), has considered optical Mach-Zehnder and phase modulators parametrized to comply with the performance specifications of typical commercial LiNbO$_3$ IQ modulators [28]. The all-optical sigmoid activation unit has been realized by implementing a deeply saturated differentially-biased Semiconductor Optical Amplifier (SOA)-MZI followed by a Cross-Gain Modulated (XGM)-SOA module operated in its small-signal gain regime, following exactly the layout demonstrated experimentally and described in [21] and using the SOA parameters for closely matching the experimental SOA response, as already analyzed in [29]. Fig. 4(b) depicts the experimentally obtained response that almost perfectly fits a sigmoid curve, as has been shown also mathematically in [21]. The simulated activation response obtained by the VPI model is also shown in Fig. 4(b) with the red curve, designating almost a perfect matching with both the experimental curve and its sigmoid fitting. The 8 electrical input signals were obtained from the MNIST dataset for every different evaluation case considered, driving the respective optical modulators at 10 Gbaud/s. The phase modulators were constantly operated either under a zero or a V_π DC driving voltage to enforce a positive or a negative weight sign, respectively, depending on the specifications received by the offline training process. The DC driving voltages of the weight value optical modulators were similarly enforced by the training process. The output of the all-optical sigmoid neuron was then captured by a PD and was monitored on an oscilloscope.

A. Training With MNIST Dataset

The training procedure utilized the experimentally obtained photonic sigmoid activation function reported in [21], following the sigmoid fitting:

$$f(x) = A_2 + \frac{A_1 - A_2}{\left(1 + e^{\left((x_0 - x)/d\right)}\right)} \tag{7}$$

where $A_1 = 0.060$, $A_2 = 1.005$, $x_0 = 0.145$ and $d = 0.033$. It should be noted that the proposed coherent linear neuron encodes the numerical values in the field of optical beam, while the employed activation function has been measured in optical power. This has been incorporated during the training procedure by squaring the linear neuron output prior incorporating this into the activation function. The MNIST dataset has been utilized in all experiments [30], containing 70,000 images of handwritten digits $(0 - 9)$ that were divided into 60,000 training samples and 10,000 testing samples. For evaluating the training and simulation performance of the proposed algorithm we used 4 different class pairs of varying difficulty: $0 - 1, 2 - 3, 4 - 5$ and $6 - 8$. The dimensionality of the MNIST dataset was reduced from 784 $(28 \times 28$ grayscale images) to 8 dimensions using PCA [31] to match the number of inputs in the employed photonic implementation. The proposed neuron was trained and evaluated separately on each digit pair. The weights of the neurons were optimized using a variant of stochastic gradient descent, Adam [32], to minimize the binary cross-entropy loss:

$$L = \sum_{i=1}^{N} (y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{8}$$

where y_i is the target class for the $i\text{-}th$ training sample, $\hat{y}_i$ is the output of the neuron and N is the number of training samples. However, note that $\hat{y}_i$ should never exceed 1 during the training, while the output range of this function ranges from $A_1 = 0.060$ to $A_2 = 1.005$, rendering it incompatible with the employed cross-entropy loss function which expects a probability value bounded in the [0, 1] interval. To ensure that meaningful gradients will be back-propagated, we appropriately shifted the neuron's output during the training by subtracting the value of $c = 0.005$ from the fitted function. This allowed for successfully training the neuron, without altering its function during the inference process (since this was only required during the training process). The initial weights of the neurons were drawn uniformly from the interval $[-\frac{1}{\sqrt{n_{in}}}, \frac{1}{\sqrt{n_{in}}}]$, where $n_{in} = 8$. The neurons were trained using a state-of-the-art DL framework, PyTorch [33], and then the weights were exported and realized in the VPI simulation to evaluate the performance of the all-optical coherent sigmoid neuron. During the training procedure, the learning rates were set to the largest possible values that could ensure the smooth convergence of the training process, i.e., 0.01 for the ideal sigmoid and to 0.0001 for the photonic sigmoid. All the neurons were trained for 50 epochs using a batch size of 128 samples.

B. Performance Evaluation

Applying the calculated weight values from training process onto the VPI-modelled photonic neuron, its performance could be validated via physical layer simulations at 10 Gbaud/s. Fig. 5(a)–(h) depict the weighted time traces of a 60-bit time-window from the MNIST $0 - 1$ dataset that show the weighted field amplitude $w_i x_i$ at every branch (axon) of the 8-input COLN, with red and blue lines indicating whether a positive or a negative weight sign has been applied, respectively. The coherent algebraic addition is illustrated in Fig. 5(i), while the summation of this signal with the appropriate level of optical bias towards converting the phase information into an optical power-related quantity is shown in Fig. 5(j). This signal is then forwarded into the activation unit with its optical output being illustrated in Fig. 5(k), while the corresponding electrical signal after being detected by the subsequent 10 GHz PD is shown in Fig. 5(l), together with the decision threshold, depicted by the dashed line. Indicative traces for the corresponding MNIST class $2 - 3, 4 - 5$ and $6 - 8$ datasets at the output of the coherent optical linear neuron, the activation unit and the PD, respectively, are shown in Fig. 6(a)–(c), (d)–(f) and (g)–(i). It should be noted that the readout of the neuron is carried out directly at the PD output by utilizing the detection threshold that was defined during the training phase. More specifically, the physical neuron training is performed using the 60,000 training samples from the MNIST dataset aiming at a two-fold optimization process for the read-out stage: (i) to define the position of the sigmoid's quiescent point around its linear region through the optical signal attenuation and (ii) to optimize the decision threshold. The combined influence of these two quantities maps to the bias at neurons output. Following the training phase, 10,000 MNIST testing samples are employed to validate the neurons performance by using the operational point and detection threshold that were defined as the training outcomes. After detection, the signal was synchronized with the global clock through retiming, sampled, and compared to the decision threshold, yielding an output binary stream. This stream is compared to the 100% accurate digit classification of the test samples to determine the corresponding accuracies.

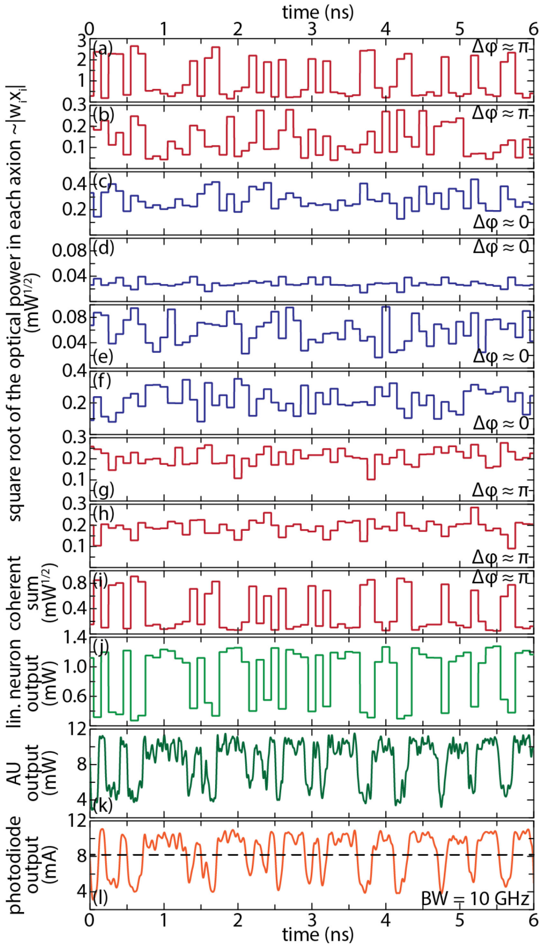

Fig. 5. Numerical analysis of the 8-input all-optical sigmoid neuron with its physical model implemented in VPI Design Suite for MNIST class 0–1. (a)–(h) Color-coded waveforms showing square roots of the optical power in each axon (in $\sqrt{\mathrm{mW}}$), proportional to the optical field amplitude carrying the optical equivalent of $|w_i x_i|$ product, with blue denoting positive, and red negative optical signals. (i) Result of the coherent algebraic summation of the signals from all axons given in $\sqrt{\mathrm{mW}}$. (j) Waveform of the signal power at the output of the coherent optical linear neuron (COLN), after interfering with the attenuated bias CW signal, having the power of 2.72 mW. (k) Waveform of the optical signal at the output of the all-optical sigmoid neuron, after the all-optical sigmoid activation unit (AU). (l) Waveform of the filtered photocurrent (with the bandwidth of 10 GHz), together with the decision threshold (dashed line).

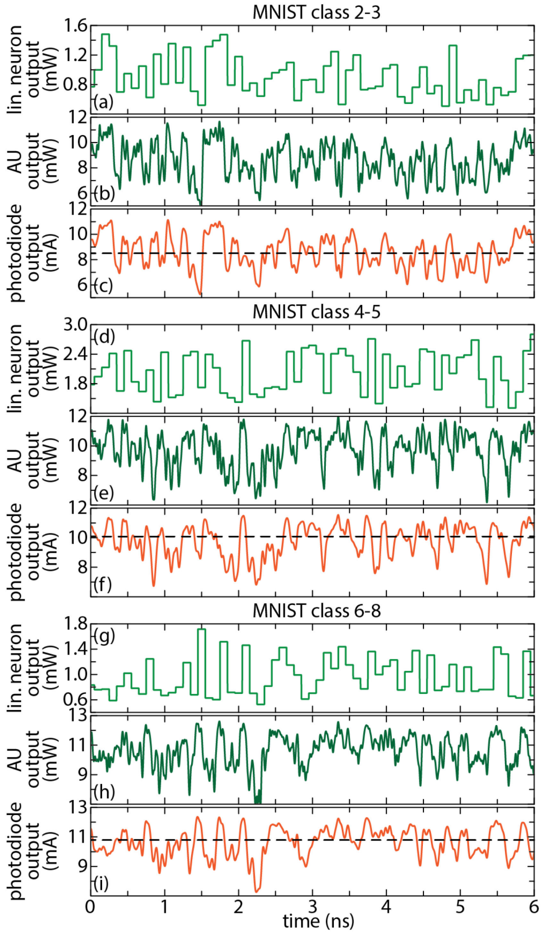

Fig. 6. Numerical analysis of the 8-input all-optical sigmoid neuron with its physical model implemented in VPI Design Suite for MNIST classes (a)–(c) 2–3, (d)–(f) 4–5, and (g)–(i) 6–8. (a), (d), (g) Waveform of the optical signal power at the output of the coherent optical linear neuron (COLN), after interfering with the attenuated bias CW signal. (b), (e), (h) Waveform of the optical signal power at the output of the all-optical sigmoid neuron, after the all-optical sigmoid activation unit (AU). (c), (f), (i) Waveform of the filtered photocurrent (with the bandwidth of 10 GHz), together with the decision threshold (dashed line).

TABLE I
Test Accuracy for MNIST Database

Class	Ideal sigmoid	Photonic sigmoid	VPI simulation
$0-1$	99.91%	99.81%	99.61%
$2-3$	94.37%	94.47%	90.86%
$4-5$	96.96%	96.96%	91.31%
$6-8$	97.46%	97.72%	95.38%

The test accuracies for the four evaluated MNIST class pairs obtained via the VPI physical layer validation process are shown in Table I and are compared to the respective accuracies calculated during the off-line training procedure for an ideal sigmoid and the photonic sigmoid activation function. As can be easily observed, the physical layer validation of the MNIST digit pair recognition process yields accuracy values that are only 0.3%–5.65% lower compared to the ideal case where a typical ideal sigmoid function was employed and only 0.1%–5.65% lower compared to the software-obtained accuracies when being trained with the photonic sigmoid function. As can be seen from Table I, the training with the photonic sigmoid performs comparable to the corresponding training with the ideal sigmoid mathematical expression. This result confirms that the experimentally validated optical sigmoid activation allows indeed for their successful employment in neuromorphic circuitry when properly adapting the training process [13], [14], despite the slight differences in its mathematical expression compared to the ideal sigmoid activation typically used in DL networks. The learning curves (on the training set) for the two different activation functions are depicted in Fig. 7. Note that even though a smaller learning rate was used for the photonic sigmoid, it is still leading to a steeper learning curve, possibly due to the largest magnitudes of its gradients during the training. This also leads

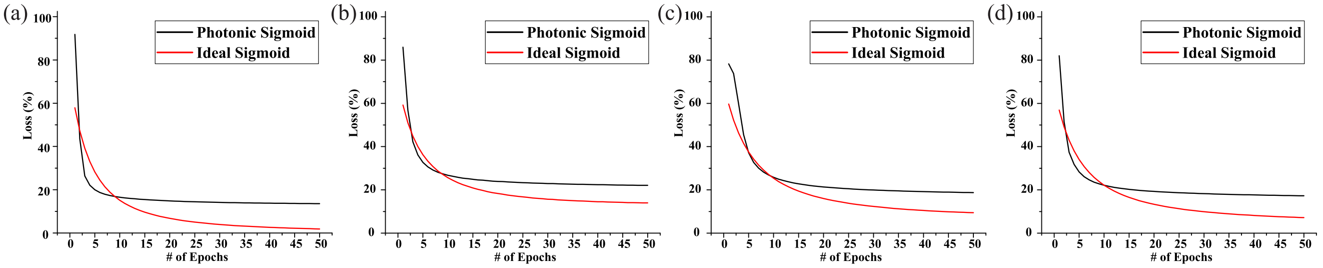

Fig. 7. Learning curves for a single neuron employing the photonic sigmoid and the ideal sigmoid, respectively, using data from MNIST class (a) 0–1, (b) 2–3, (c) 4–5 and (d) 6–7.

to slightly higher training loss in all the cases. At this point, it is worth noting that this behavior can be possibly attributed to the employed initialization scheme that was not optimal for the photonic sigmoid. Adapting the initialization scheme to the characteristics of the activation function has been shown to alleviate these issues [14], offering a first indication that an optimized hardware-software co-design approach where the training framework is optimally adapted to the new underlying neuromorphic photonic hardware can yield performance accuracies close to the well-established and mature software-based platforms. Nonetheless, even using generic (and potentially suboptimal) initialization schemes, allows for successfully training the neuron and achieving results comparable to the ideal sigmoid. This suggests that an optimized hardware-software co-design approach where the training framework is optimally adapted to the new underlying neuromorphic photonic hardware, can yield performance accuracies close to the well-established and mature software-based platforms. As a result, the combination of the well-known energy efficient, high clock-rate and computational density advantages of neuromorphic photonic platforms [11] will be allowed with the performance standards of current digital neural network implementation. Towards this goal, the proposed COLN layout expects to expedite the hardware-software co-design by introducing the capability of i) having both positive and negative numbers also in neuromorphic photonic deployments, overcoming the constraint of using mainly sinusoidal activation functions that are so far enforced by optoelectronic solutions adopted for dealing with negative values [10], ii) relying on well-established IQ modulator schemes so as to draw from their technological maturity and their implementations in several fabrication platforms [34]–[36] towards constructing more advanced and higher fan-in neuromorphic architectures.

The energy efficiency of the proposed architecture, which requires a dual-IQ modulator for a 2-axon layout, equals the energy efficiency of a single IQ modulator structure, which will obviously depend on the specific integration technology platform that will be adopted. Recent demonstrations of IQ modulators as Silicon-Organic hybrid [35], Plasmonic [36] and InP-based [34] devices suggest that energy efficiency can span from tens of pJ/baud offered by InP deployments down to tens of fJ/baud reported by plasmonics, with InP comprising the more mature IQ modulator technology but with silicon and plasmonic IQ structures offering pronounced benefits when energy efficiency and footprint form the main performance metrics. In an energy efficient deployment path, the rather power-hungry SOA-based sigmoid activation module should also be replaced by lower footprint and lower energy nonlinear technologies like the III/V-on-Si photonic crystals [37] that have been demonstrated to offer nonlinear functions similar to SOAs but with energy efficiencies of just a few fJ/baud.As far as processing time requirements are concerned, the proposed photonic neuron retains the low-latency characteristics of all-optical neuromorphic layouts that stem from their time-of-flight and GHz-scale bandwidth of photonic technologies [11].

V. Conclusion

We have demonstrated a single-wavelength COLN that relies on a nested interferometric arrangement of dual-IQ modulation cell structures and supports both positive and negative value encoding through the electric field phase information, validating experimentally the linear algebra operational credentials of a typical IQ modulator at 10 Gbaud/s and confirming mathematically and via simulations its extension towards high fan-in linear photonic neurons. This linear photonic neuron setup was then extended into an 8-input linear neuron that was followed by an experimentally validated photonic sigmoid activation function to form an all-optical sigmoid neuron at 10 Gbaud/s, with its physical layer performance as a fully trained neuron for handwritten digit recognition being validated via VPI simulations. The training of this neuron with MNIST digit pair datasets revealed that the photonic sigmoid activation function can yield accuracy values almost identical to the ideal sigmoid case, while the transfer of the training parameters into the VPI modelled photonic neuron resulted in physical-layer validated accuracy values that were always >90% and deviated only between 0.3%–5.65% from the accuracy values expected during the training.

References

[1] T. N. Theis and H.-S. P. Wong, "The end of Moore's law: A new beginning for information technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, Mar. 2017.

[2] F. Akopyan *et al.*, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[3] S. B. Furber *et al.*, "Overview of the SpiNNaker system architecture," *IEEE Trans. Comput.*, vol. 62, no. 12, pp. 2454–2467, Dec. 2013.

[4] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.

[5] A. H. Atabaki *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, p. 349, 2018.

[6] M. Moralis-Pegios *et al.*, "Chip-to-chip interconnect for 8-socket direct connectivity using 25 gb/s o-band integrated transceiver and routing circuits," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2018, pp. 1–3.

[7] S. Pitris *et al.*, "A 40 Gb/s chip-to-chip interconnect for 8-socket direct connectivity using integrated photonics," *IEEE Photon. J.*, vol. 10, no. 5, pp. 1–8, Oct. 2018.

[8] S. Pitris *et al.*, "A 4×40 Gb/s o-band WDM silicon photonic transmitter based on micro-ring modulators," in *Proc. Opt. Fiber Commun. Conf.*, Optical Society of America, 2019, p. W3E.2.

[9] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jun. 2017.

[10] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Rep.*, vol. 7, no. 1, p. 7430, 2017.

[11] H.-T. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, pp. 1–15, Nov.–Dec. 2018.

[12] 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, 2019, Art. no. 14009.

[13] 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.*, pp. 1–10, 2019.

[14] N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, "Variance preserving initialization for training deep neuromorphic photonic networks with sinusoidal activations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, IEEE, 2019, pp. 1483–1487.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14 Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[16] 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. Vision*, IEEE, Dec. 2015, pp. 1026–1034.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Demonstration of WDM weighted addition for principal component analysis," *Opt. Express*, vol. 23, no. 10, pp. 12 758–12 765, May 2015.

[19] A. N. Tait *et al.*, "Microring weight banks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, Nov.–Dec. 2016, Art. no. 5900214.

[20] M. Miscuglio *et al.*, "All-optical nonlinear activation function for photonic neural networks [invited]," *Opt. Mater. Express*, vol. 8, no. 12, pp. 3851–3863, Dec. 2018.

[21] 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, p. 9620, 2019.

[22] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, and N. Pleros, "Experimental demonstration of an optical neuron with a logistic sigmoid activation function," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Mar. 2019, pp. 1–3.

[23] T. Sakamoto, A. Chiba, and T. Kawanishi, "Electro-optic synthesis of multi-level coherent signals," in *Proc. 14th OptoElectronics Commun. Conf.*, IEEE, Jul. 2009, pp. 1–2.

[24] T.-H. Nguyen *et al.*, "Blind transmitter IQ imbalance compensation in M-QaM optical coherent systems," *J. Opt. Commun. Netw.*, vol. 9, no. 9, pp. D42–D50, 2017.

[25] P. W. Berenguer *et al.*, "Nonlinear digital pre-distortion of transmitter components," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1739–1745, 2015.

[26] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Continuous calibration of microring weights for analog optical networks," *IEEE Photon. Technol. Lett.*, vol. 28, no. 8, pp. 887–890, Apr. 2016.

[27] M. Harjanne, M. Kapulainen, T. Aalto, and P. Heimala, "Sub-μs switching time in silicon-on-insulator Mach–Zehnder thermooptic switch," *IEEE Photon. Technol. Lett.*, vol. 16, no. 9, pp. 2039–2041, Sep. 2004.

[28] 40 GHz IQ $LiNbO_3$ modulator. [Online]. Available: https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=3948

[29] E. Kehayas, K. Vyrsokinos, L. Stampoulidis, K. Christodoulopoulos, K. Vlachos, and H. Avramopoulos, "ARTEMIS: 40-Gb/s all-optical self-routing node and network architecture employing asynchronous bit and packet-level optical signal processing," *J. Lightw. Technol.*, vol. 24, no. 8, pp. 2967–2977, Aug. 2006.

[30] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[31] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Rev. Comput. Statist.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv preprint arXiv:1412.6980*.

[33] F. W. Pfeiffer, "Automatic differentiation in pyTorch," *ACM SIGNUM Newslett.*, vol. 22, no. 1, pp. 2–8, 2007.

[34] A. Aimone *et al.*, "DAC-free ultra-low-power dual-polarization 64-QAM transmission with InP IQ segmented MZM module," in *Proc. Opt. Fiber Commun. Conf.*, Optical Society of America, 2016, pp. Th5C–6.

[35] S. Wolf *et al.*, "Silicon-organic hybrid (SOH) IQ modulator for 100 GBd 16 QAM operation," in *Proc. Opt. Fiber Commun. Conf.*, Optical Society of America, 2017, pp. Th5C–1.

[36] W. Heni *et al.*, "Plasmonic IQ modulators with attojoule per bit electrical energy consumption," *Nature Commun.*, vol. 10, no. 1, p. 1694, 2019.

[37] T. Alexoudi *et al.*, "III–V-on-Si photonic crystal nanocavity laser technology for optical static random access memories," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, pp. 295–304, Nov.–Dec. 2016.