Neuromorphic computing through photonic integrated circuits

G. Mourgias-Alexandris\textsuperscript{a,b}, A. Totovic\textsuperscript{c}, N. Passalis\textsuperscript{a}, G. Dabos\textsuperscript{a,b}, A. Tefas\textsuperscript{a}, N. Pleros\textsuperscript{a,b}

\textsuperscript{a}Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
\textsuperscript{b}Center for Interdisciplinary Research and Innovation, Aristotle University of Thessaloniki, Greece

ABSTRACT

The identification of neuromorphic computing as a highly promising alternative computing system has emerged from its potential to increase rapidly the computational efficiency that is currently restricted by Moore’s law end. First electronic neuromorphic chips like IBM’s TrueNorth and Intel’s Loihi revealed a tremendous performance improvement in terms of computational speed and density; however, they are still operating in MHz rates. To this end, neuromorphic photonic integrated circuits can further increase the computational speed and density, having a large portfolio of components with GHz-bandwidth and low-energy. Herein, we present an all-optical sigmoid activation function as well as a single-λ linear neuron. The all-optical sigmoid activation function comprises a Semiconductor Optical Amplifier-Mach-Zehnder Interferometer (SOA-MZI) configured in differentially-biased scheme followed by an SOA. Its thresholding capabilities have been experimentally demonstrated with 100psec optical pulses. Then, we introduce an all-optical phase-encoded weighting scheme and we experimentally demonstrate its linear algebra operational credentials by the means of a typical IQ modulator operated at 10Gbaud/s.

Keywords: Neuromorphic photonics, photonic integrated circuits, neural networks, neural network accelerators, optical computing, artificial intelligence

1. INTRODUCTION

With Moore’s law coming to an end [1], neuromorphic computing appears as an alternative for sustaining computational advances and overcoming the digital energy efficiency wall of von-Neumann architectures. Recent state-of-the-art neuromorphic accelerators confirm the potential of non-von-Neumann brain-inspired processors like IBM’s TrueNorth [2], SpiNNaker [3] and Intel’s Loihi [4] to 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 [5–9], recent research efforts are already attempting to transfer the neuromorphic computing principles over optics [10,11]. This has led to the introduction of neuromorphic photonics [10–12] as a new scientific area, indicating already its huge energy and footprint efficiency perspectives in case of successfully transferring the large bandwidth advantages of optics onto a neural network operational platform [12,13].

However, the optical neurons that have been demonstrated so far [12,14–16] rely on a rather limited set of photonic activation functions compared to the typical range of activations employed in Deep Neural Network (DNN) layouts, despite the variety of photonic non-linear thresholding elements that have been demonstrated as stand-alone modules [17–21]. DNN layouts typically require the ReLU, PReLU and variations of the sigmoid transfer function, including tanh and the logistic sigmoid but the vast majority of photonic activation functions has been based so far either on close to Heaviside step transfer functions offered usually by directly modulated laser-based neurons [15,22] or on sinusoidal [11,21] responses that are usually obtained via an electro-optic modulator driven by the output of a balanced photodiode circuitry [22] or by optical switching elements [12]. Spiking NNs have also been mainly implemented with sinusoidal photonic integrate-and-fire transfer functions typically provided by fiber- or SOA-based thresholding devices [23] or with close to step-wise unitary functions [24]. Although our recent work outlines the training procedure and confirms that optically-enabled sinusoidal activation can lead to converging Deep Convolutional Neural Networks (CNN), sigmoid-based activations continue to constitute the backbone of powerful Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM) architectures, being responsible for gating and confining the output between two well-defined levels. Furthermore, neural network attention mechanisms, which significantly improve the accuracy of neural models, can be directly implemented using sigmoid activations [25]. The efficient deployment of Artificial Intelligence (AI)
workloads based on the well-known RNN and LSTM architectures requires a complete portfolio of photonic modules including an all-optical sigmoid activation unit.

The linear neuron stage is another critical part of an all-optical neuron since its responsible for carrying out the linear algebra operations supporting at the same time both positive and negative weight representations. To cope with the need for both positive/negative weights, the weighting layouts that have been proposed so far have mainly relied on Wavelength Division Multiplexing (WDM) schemes [11,26,27] encoding every input signal onto a different wavelength and using a pair of balanced Photodetectors (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 photodetector followed by an optical modulator [11] 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 [28] or sigmoid [29–31] activation functions, requiring the deployment of new training frameworks to accommodate non-typical activation functions, like the modulator’s sinusoidal response [32]. 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 O-E-O conversion, but the scaling of fan-in still requires at least two different lasers for positive/negative weight representation [26]. 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 MZIs [10]. This design follows the Reck-proposal and requires a N^2 number of MZIs for an N-input configuration, scaling quadratically with the fan-in value.

Herein, we present a single-λ linear algebra module based on a conventional IQ modulator and an all-optical sigmoid activation function. First, we experimentally demonstrate a sigmoid activation function and present its successful operation performing thresholding over 4 different optical power levels of a multi-level WDM analog input signal, offering a 100% improvement over power level resolution in state-of-the-art photonic neuromorphic activation elements [23]. The sigmoid optical element comprises a deeply saturated differentially-biased SOA-MZI followed by an SOA. Then, we validate the linear algebra principles of a typical IQ modulator in weighted optical addition and subtraction operations of two optical signals, showing the potential of such layouts to scale into larger fan-ins by demonstrating experimentally by the means of an IQ modulator all-optical linear algebra operations at 10Gbaud/s.

2. EXPERIMENTAL DEMONSTRATION OF AN ALL-OPTICAL SIGMOID ACTIVATION FUNCTION

Artificial neural networks are usually arranged in feed-forward topologies, as shown in Fig. 1(a). One input layer, at least one hidden layer and one output layer, with each one being consisted by a tuple of artificial neurons. The layout of an artificial neuron is depicted in Fig. 1(b). Every single input X_k is weighted by a dedicated weight W_k before all of the weighted signals reach the summation stage with the resulted signal being forwarded into the activation function. Finally, the output is broadcasted to the neurons of the next layer, where the same process is followed for each neuron. The corresponding photonic layout of our proposed optical neuron is depicted in Fig. 1(c). Every input X_k is imprinted to a separate wavelength, while the weight W_k is realized by the corresponding Variable Optical Attenuator (VOA). The weighted signals enter then an Arrayed Waveguide Grating (AWG) for being multiplexed into a single waveguide, with the resulting stream being the sum of the four individual optical powers and then injected into the “Activation Unit”. The “Activation Unit” is realized by a SOA-MZI with its output connected to an SOA, with the first device operating in the deeply saturated regime and both of them operating as wavelength converters to yield a logistic Sigmoid activation function. The AWG-multiplexed summed signal is split into 2 identical streams and enters the SOA-MZI in co- and counter-propagation, respectively. Two laser beams (λ_4 and λ_5) are also fed into the SOA-MZI as input and holding beams, respectively, so that the SOA-MZI operates in its differentially-biased wavelength conversion scheme. In this way, the AWG-multiplexed signal is inversely imprinted on λ_4 that forms the switched output of the SOA-MZI, which is forwarded as the control signal into the successive SOA after being filtered in an Optical Bandpass Filter (OBPF). A CW laser beam at λ_6 is used as the SOA input signal, so that the output of the SOA and consequently of the optical neuron gets finally recorded on λ_6. More details about the device employed as the “Activation Unit” and its principle of operation can be found in [30], including a theoretical and experimental analysis transfer function.

The 4 optical inputs were provided by four laser beams emitting light at λ_1 = 1549.4nm, λ_2 = 1550.2nm, λ_2 = 1551.0nm, and λ_3 = 1551.8nm and modulated in 4 respective Lithium Niobate (LiNbO_3) modulators, each of them driven by a
Programmable Pattern Generator (PPG) to realize a periodic pulse signal with 100psec long pulses and 200psec spacing between successive pulses. In particular, within a duration of 1.6nsec, 4 pulses at $\lambda_1$ were generated by the first modulator, 3 pulses at $\lambda_2$ by the second modulator, 2 pulses at $\lambda_3$ by the third and 1 pulse at $\lambda_4$ by the fourth. Each one of these signals was weighted by a VOA and then properly time-synchronized via respective Optical Delay Lines (ODLs) prior being launched into the AWG multiplexer to realize the multi-level WDM signal. Afterwards, a VOA is responsible to adjust the optical power of the multi-level WDM signal defining the threshold level of the optical activation function unit. A 3dB coupler splits the optical signal into 2 identical streams, with the first one being injected into port A of the SOA-MZI. Because of the differentially-biased scheme, the second stream has to be attenuated by the corresponding VOA before is launched into the port H of the SOA-MZI. Two laser beams at $\lambda_4 = 1547.7$nm $\lambda_5 = 1548.5$nm are injected into the ports B and D, respectively. The wavelength converted signal is filtered by an Optical Bandpass Filter (OBPF) at $\lambda_4 = 1547.7$nm and along with a laser beam at $\lambda_6 = 1550.13$nm is forwarded into the SOA. The output signal is filtered by an OBPF at $\lambda_6 = 1550.13$nm before it reaches the oscilloscope for recording time traces and eye diagrams.

Figure 2 depicts the obtained experimental results during the evaluation of the proposed optical neuron. The experimental measurement of transfer function for the proposed neuron layout along with the transfer function obtained when using a simple SOA-MZI switch is depicted in Fig. 2(a), showing clearly that a simple SOA-based switch offers a close to sinusoidal transfer function, while the proposed activation function provides a sigmoid shape between two well-defined lower and upper-level power plateaus. The mathematical fitting of the sigmoid generalized formula provided by Eq. (1) to the experimental transfer function of the proposed activation function reveals an excellent agreement when using $A_1=0.060, A_2=1.005, x_0=0.145, \text{and } d=0.033$.

$$f(x) = A_2 + \frac{A_1 - A_2}{1 + e^{(x-x_0)/d}}$$  \hspace{1cm} (1)

The complete theoretical analysis of the SOA-based sigmoid activation function has been carried out in [30]. A time trace and the corresponding eye diagram from the input multi-level WDM signal is depicted in Fig. 2(b). The time synchronization between the four optical input streams and the weights have been adjusted in order to produce a 4-level signal with 0.54dB ER for the first pulse, 3.01dB for the second, 5.3 for the third and 8.66dB for the last and highest one. Injecting 11.5dBm optical power into the port A of the SOA-MZI and 10.7dBm into port H, sets the threshold of the Sigmoid activation function at the lower power level, denoted as “level 1” in the figure. The output of the neuron in this

Fig. 2. (a) Experimental transfer function and mathematical fitting of the employed activation element vs a typical SOA-MZI switch, (b) Time trace (200 psec/div) and eye diagram (50 psec/div) of the summed optical multi-level signal obtained at the AWG output, with the dashed lines denoting the different levels used then for thresholding, (c)-(f) Time traces at the output of the neuron (200 psec/div) along with the corresponding eye-diagrams (50 psec/div).
case is shown in Fig. 2(c), clearly showing that all 4 incoming pulses that are above the threshold “level 1” have resulted to respective power equalized output pulses with an amplitude modulation (AM) of 4.5dB and an extinction ratio (ER) of 4.7dB. Figure 2(d) illustrates the corresponding time trace and eye diagram when the threshold is set at “level 2” by decreasing the SOA-MZI control signal power by 3.4dB on both SOA-MZI control ports. As can be noticed, the shortest input optical pulse that was below the threshold has been eliminated and the three pulses above the threshold have turned into a power equalized pulse sequence, with an AM of 4dB and an ER of 6dB. Reducing the SOA-MZI control signal optical power by additional 3.3dB sets the threshold at “level 3”, so that the neuron operation forces both pulses being below the threshold to disappear, as shown by corresponding time trace and eye diagram of Fig. 2(e). AM and ER of the obtained 2-pulse signal were measured to be 2.5dB and 5.8dB, respectively. To move the threshold at “level 4” requires the reducing of optical power by 3.9dB in addition, resulting in a neuron output that comprises a single optical pulse with an ER of 6dB, shown in Fig. 2(f). The operational conditions during the demonstration of the proposed optical neuron were as follows: For the SOA-MZI, the SOA1 and SOA2 were driven at 240mA and 280mA, respectively. The input (port B) and the auxiliary (port C) laser beam was 5.5dBm and 6.5dBm respectively, while the optical power for the control signals was mentioned above. The output SOA was driven by 300mA and fed by a laser beam of -10.5dBm.

3. OPERATE AN IQ MODULATOR AS A 2-INPUT OPTICAL LINEAR ALGEBRA UNIT

Figure 3(a) 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 and enforcing via the phase shifter a phase shift of \( \psi_i = 0 \) or \( \psi_i = \pi \) 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 \( w_1x_1 = \sin\left(\frac{\pi V_I}{V_{sa}}\right) \) and \( w_2x_2 = \sin\left(\frac{\pi V_Q}{V_{sa}}\right) \) the IQ modulator output can be expressed as:

\[
E_{out} = \frac{1}{2}E_{in}[w_1x_1 + w_2x_2e^{j\psi_i}] = \frac{1}{2}E_{in}[w_1x_1 \pm w_2x_2]
\]

(2)

where \( w_i \), \( x_i \) denote the weight and the input value, respectively, of each input. Equation (2) 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 \( \psi_i \), 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 (2), we have used the experimental setup shown in Fig. 3(b). A laser beam at \( \lambda_0 = 1550nm \) is injected into a LiNbO\(_3\) IQ modulator where both the \( I \) and \( Q \) MZI structures are driven by two electrical Gaussian pulse-shaped sequences with a 104-ps pulse width, generated by an Arbitrary Waveform Generator (AWG) to emulate the \( w_1x_1 \) 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 \( \pi \) 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 photodiode (PD) and is monitored via an oscilloscope for signal evaluation purposes.

The pulse traces that were captured at the photodiode electrical output for different cases of \( I \) and \( Q \) modulated pulse sequences are illustrated in Fig. 4. 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. 4

Fig. 3. (a) Typical layout of an optical IQ modulator, (b) the experimental setup for the emulation of the 2-input COLN.
illustrates a different set of two input signals \( w_{1x1} \) and \( w_{2x2} \) corresponding to the \((w_{1x1})^2\) and \((w_{2x2})^2\) 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, Figs. 4(a) and 4(b) depict the traces of \((w_{1x1})^2\) and \((w_{2x2})^2\) in the case when both \((w_{1x1})^2\) and \((w_{2x2})^2\) are identical. Normalizing every pulse sequence to its highest peak power, each of the \( w_{1x1} \) and \( w_{2x2} \) pulse sequences reveals electric field peak amplitudes of 1, 0.9 and 0.7 for its three constituent pulses. Inducing a 0 or \( \pi \) phase shift by means of the phase shifter, coherent addition or subtraction is achieved resulting in \((w_{1x1} + w_{2x2})^2\) and \((w_{2x2} - w_{1x1})^2\) at the output of the PD, as can be seen in Figs. 4(c) and 4(d), respectively. Normalizing again with respect to the highest input pulse peak power, Fig. 4(c) confirms the successful coherent addition of \( w_{1x1} \) and \( w_{2x2} \) products, as the ratio between the intensities of the \((w_{1x1} + w_{2x2})^2\) output pulse peak powers is identical to the respective ratios of the constituent \((w_{1x1})^2\) and \((w_{2x2})^2\) pulse peak powers. Figure 4(d) shows clearly the successful result of \((w_{2x2} - w_{1x1})^2\) that equals zero in this case.

The second column depicts the case when using the same \((w_{1x1})^2\) and \((w_{2x2})^2\) signals with \((w_{2x2})^2\) being delayed by one pulse period with respect to the \((w_{1x1})^2\) pulse sequence, as shown in Figs. 4(e) and 4(f), respectively. Following the same procedure as explained for Figs. 4(a)-(d), successful coherent addition and subtraction can be again verified through the output pulse traces illustrated in Figs. 4(g) and 4(h), respectively. The third and the fourth column of Fig. 4 illustrate two additional scenarios where the \((w_{2x2})^2\) is delayed by two [Fig. 4(j)] and three [Fig. 4(n)] pulse periods, respectively, with respect to the \((w_{1x1})^2\) signal. In both cases, successful addition and subtraction has been obtained as can be verified by Figs. 4(k), (l) and 4(o), (p), respectively.

However, this operational arrangement can only provide the absolute value of the subtraction through the signal’s power, as can be clearly seen in Figs. 4(l) and 4(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 \( w_{1x1} \) 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 \( w_{1x1} \) + \( b \) instead of the \( w_{1x1} \) value considered in the theoretical analysis summarized in (2). In this way, (2) can be rewritten as:

\[
E_{out} = \frac{1}{2} E_{in} \cdot [(w_{2x2} + b) \pm w_{1x1}]
\]

(3)

![Fig. 4. Time traces of the initial \( w_{1x1} \) and \( w_{2x2} \) signals generated by \( I \) and \( Q \) MZIs, summed and subtracted when: (a)-(d) \( w_{2x2}(t) = w_{1x1}(t) \), (e)-(h) \( w_{2x2}(t) = w_{1x1}(t-1) \), (i)-(l) \( w_{2x2}(t) = w_{1x1}(t-2\tau) \), (m)-(p) \( w_{2x2}(t) = w_{1x1}(t-3\tau) \) and (r)-(u) \( w_{2x2}(t) = w_{1x1}(t-3\tau)+bias \). (x-axis scale: 100psec/div, y-axis scale: 2.5mV/div)
Where we have assumed that the DC bias level has been enforced at the $Q$-MZI branch that generates $w_{2x2}$. 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. 4. Figures 4(r) and 4(s) illustrate the optical power levels corresponding to the $(w_{1x1} + b)^2$ when $w_{2x2} = 0$ and $(w_{2x2} + b)^2$ when $w_{1x1} = 0$, 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. 4(t) where obviously all pulses are atop the biasing beam. The coherent subtraction when enforcing a $\pi$ phase shift at the $Q$-branch is shown in Fig. 4(u), where indeed the positive differences are imprinted as pulses atop the biasing beam but the negative $w_{2x2} - w_{1x1}$ differences emerge as inverted pulses below the power level of the biasing beam. The IQ modulator can be used as the basis for a generic layout of an n-input coherent linear neuron that has been described and mathematically analyzed in [33].

4. CONCLUSION

We have demonstrated experimentally an all-optical sigmoid activation function by a SOA-MZI interferometer followed by SOA as well as an IQ modulator acting as single-wavelength linear algebra module supporting both positive and negative value encoding through the electric field phase information. The sigmoid activation has been employed to perform experimentally thresholding in 4 different power levels for 100ps long pulses, offering a 100% improvement in the number of threshold levels resolved by state-of-the-art neuromorphic optical activation elements [23]. Finally, the linear algebra operational credentials of a typical IQ modulator have been demonstrated at 10Gbaud/s. Both layouts along with novel training methods adapted in neuromorphic photonic hardware [34] can form a complete portfolio including all the necessary building blocks towards the realization of neuromorphic photonic layouts.

5. ACKNOWLEDGEMENTS

This work has been supported by the EU H2020 project ICT-PLASMONIAC (871391) and ICT-MASSTART (825109).

REFERENCES


N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, "Variance Preserving Initialization
