Research Article | Vol. 27, No. 4 | 18 Feb 2019 | OPTICS EXPRESS 5181

Optics EXPRESS

Neuromorphic photonics with electro-absorption modulators

JONATHAN K. GEORGE,[1] ARMIN MEHRABIAN,[1] RUBAB AMIN,[1] JIAWEI MENG,[1] THOMAS FERREIRA DE LIMA,[2] ALEXANDER N. TAIT,[2] BHAVIN J. SHASTRI,[2] TAREK EL-GHAZAWI,[1] PAUL R. PRUCNAL,[2] AND VOLKER J. SORGER[1,*]

[1]*Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA*
[2]*Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA*
[*]*sorger@gwu.edu*

Abstract: Photonic neural networks benefit from both the high-channel capacity and the wave nature of light acting as an effective weighting mechanism through linear optics. Incorporating a nonlinear activation function by using active integrated photonic components allows neural networks with multiple layers to be built monolithically, eliminating the need for energy and latency costs due to external conversion. Interferometer-based modulators, while popular in communications, have been shown to require more area than absorption-based modulators, resulting in a reduced neural network density. Here, we develop a model for absorption modulators in an electro-optic fully connected neural network, including noise, and compare the network's performance with the activation functions produced intrinsically by five types of absorption modulators. Our results show the quantum well absorption modulator–based electro-optic neuron has the best performance allowing for 96% prediction accuracy with 1.7×10^{-12} J/MAC excluding laser power when performing MNIST classification in a 2 hidden layer feed-forward photonic neural network.

1. Introduction

Photonic neural networks (NN) have the potential for both high channel capacity (i.e. data baud rate) and low operating power. The former is enabled by a) only having to charge-up the capacitors of the electro-optic active devices allowing the 'wire' (waveguide) to not limit RC-signal delays [1], and b) 'bosonification' where many photons are allowed to occupy the same quantum state, such as technologically utilized in wavelength division multiplexing (WDM) [2]. If the light-matter-interaction is enhanced, such as realized via sub-wavelength photonics or plasmonics [3–5], the energy-per-compute (e.g. bit, or multiply-accumulate, MAC for NNs) can surpass electronic efficiency (i.e. about 1-10 GMAC/J) [6]. An artificial neuron requires two functions; first, it must have a synaptic-like linear function, weighting the set of inputs from other neurons (Figs. 1(a) and 1(b)). Secondly, it must apply a nonlinear activation function to the sum of the weighted inputs (Fig. 1(e)). In photonic NNs, these functions can also be separated into a passive weighted interconnect and a set of active photonic neurons.

Interferometric [7] and integrated microring resonator (MRR)-based [8] weighting have both been previously realized in integrated photonic platforms. With interferometric weighting, the phase of coherent light is utilized in a mesh of Mach-Zehnder interferometers (MZI) to produce a vector dot product of the neuron's input vector and its weights. Similarly, in ring-based weighting, photonic rings are selectively tuned to apply a dot product, a potentially more compact method utilizing WDM to associate each neuron with a specific wavelength towards realizing waveguide-efficient all-to-all network topologies [8, 9]. In both cases linear optics achieves

#344511 https://doi.org/10.1364/OE.27.005181
Journal © 2019 Received 16 Nov 2018; revised 29 Dec 2018; accepted 3 Jan 2019; published 12 Feb 2019

Optics EXPRESS

an efficient weighting where the wave nature of light computes the inner product simply by propagating forward in time. After the weights are tuned, the only additional energy consumed in creating the inner product is due to the additional laser power required to counteract propagation losses.

An activation function is a nonlinear function that is applied to the weighted sum of the inputs of a neuron (Fig. 1(e)). The nonlinearity of the activation function allows the network to converge into definitive states by eliminating infinitely cascading noise, similar to the nonlinearity of the transistor forcing the digital computer into binary states. Apart from the requirement of nonlinearity and differentiability for training, there are no limits to the shape of the activation function itself, and many activation functions have been proposed each with strengths in different applications [10].

While neural network architectures have been modeled and fabricated with external nonlinear activation functions [7] and Mach-Zehnder modulators [9], absorption modulators have not been modeled in photonic neural networks. External nonlinear-activation functions are often implemented with digital conversion between layers which quickly becomes the most costly piece of the system [11]. External nonlinear analog components increase system complexity and interconnect latency. Mach-Zehnder modulators, while appealing from their popularity in communications systems, have been shown to require a larger footprint [5]. In neural networks larger footprint results in lower neural density. This points to monolithically integrated absorption modulators or micro-scale lasers [8] as the optimal modulators for photonic neural networks.

In this work, we develop a model to guide the optimal choice of design parameters in the development of a monolithically integrated photonic neural network using on-chip absorption modulators. We begin by introducing the electro-optic absorption modulators and examine their operating parameters. To this we add a circuit for coupling the photodiode to the modulator. We introduce an integrating circuit with capacitive coupling to allow higher voltages to be reached with minimal impact to RC delay. To the circuit and modulator models, we add a noise model and sweep the free parameters of the modulators, length and input optical power, to maximize SNR. We introduce a power consumption model to allow results to be compared on a basis of energy efficiency. Finally, we simulate our model by training it on the well-studied MNIST dataset and compare results with different modulators and optical powers.

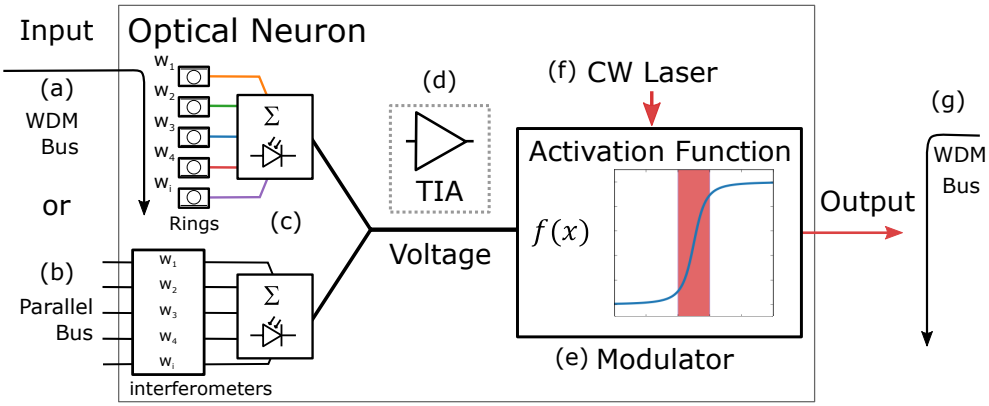

Fig. 1. An electro-optic neuron taking an input from (a) a WDM bus and weighting by wavelength with rings or (b) from a parallel bus weighting with an interferometer network. The neuron sums the optical signal with a photodiode converting the signal to a voltage (c), optionally amplified by a TIA (d), drives an electro-optic modulator (e) modulating a CW laser (f), which produces a nonlinear transfer function at the output (g) of an photonic neural network (NN).

Optics EXPRESS

2. Modulators for nonlinear activation

A variety of approaches are available for generating nonlinearity in photonic NNs. Applying the nonlinear activation function in an external CPU, such as demonstrated in [7], or digitally by chaining an analog to digital converter (ADC) to a digital multiplier to a digital to analog converter (DAC) increases latency and adds to power consumption [11]. With digital conversion a latency of several clock-cycles of a potentially slower clock rate are required between every layer of the network. Likewise, external electronic analog nonlinearity, such as external amplifiers, adds to latency by introducing interconnect delay. Finally, optical nonlinearity, while appealing in terms of latency, currently requires large optical powers for minimal effect. This points to monolithically integrated active photonic components as the optimal solution to multi-layer optical neural networks with today's material systems.

To set the context for the discussion of electro-optic modulators for NNs, in this work we are interested in photonic chip-based NNs, where integration density is a key value proposition. Modulators either alter the signals amplitude via direct electro-absorption (EAM), or shifting the phase inside an interferometer (electro-optic modulator, EOM) such as in linear MZI (or MRR) to modulate the amplitude. To provide the nonlinear activation function in photonic NNs either modulator can be utilized yet with different rationales; the advantage of EAMs is that they do not rely on interferometeric schemes to modulate a signals amplitude and, thus, can be designed more compact than EOMs. Reducing footprint allows increasing a) the neuron's areal density, and b) the photonic neurons' firing- or clock speed, since the MAC rate (i.e. MAC/s) scales with modulator 3dB-bandwidth (speed). However, if the vector matrix multiplication (VMM) is performed via shifting phase, then using phase for the activation function may be synergistic to the design layout. An advantage of field-driven EOM over carrier-based EAMs is that their intrinsic switching time is substantially shorter as compared to electronic clocking [12]. We note that both ref [7] and [8] used modulators only to perform the vector matrix multiplication and not the nonlinear activation.

In this work we consider carrier-based EAMs motivated by a) emerging material developments able of unity-strong optical index modulation which is 3-orders of magnitude stronger than the plasma effect in Silicon [5], and b) the chip density arguments made above. EAM devices absorb light as a function of their electrical bias. Either they absorb more light at zero bias and less light as the bias increases in magnitude or vice versa, depending on the type. This effect is used in optical communications to encode electrical signals on optical carriers. The shape of the voltage to absorption curve varies by EAM type [13] (Fig. 2). All EAM absorption curves are nonlinear due to the eventual saturation in the number of carriers (depletion or injection). The impact of EOM nonlinear activation functions on photonic NNs will be reported elsewhere.

3. Photodiode coupling

The electro-optic mediated nonlinearity is created first by converting an optical signal into an electrical one (O-E) and then performing the inverse electrical-to-optical (E-O) transduction. The performance of this nonlinear activation depends strongly on the choice of coupling between the photodiode and the electro-optic modulator performing the respective conversions. To this end, we model the photodiode electrically as a current source and the electro-optic modulator as a voltage-dependent capacitive load (Fig. 2). The current produced by the photodiode must then be converted into a voltage of sufficient magnitude to drive the modulator.

Generally, there are four options for coupling the photodiode to the modulator; the first adds a current-to-voltage converting resistor to the circuit (Fig. 3(a)). This is the simplest method and can be used to produce a voltage of any magnitude. However, the resistor in combination with the capacitance of the modulator and photodiode acts as a low pass RC filter [1]. The higher the required voltage, the greater the required resistance and the slower the device operates. The

Optics EXPRESS

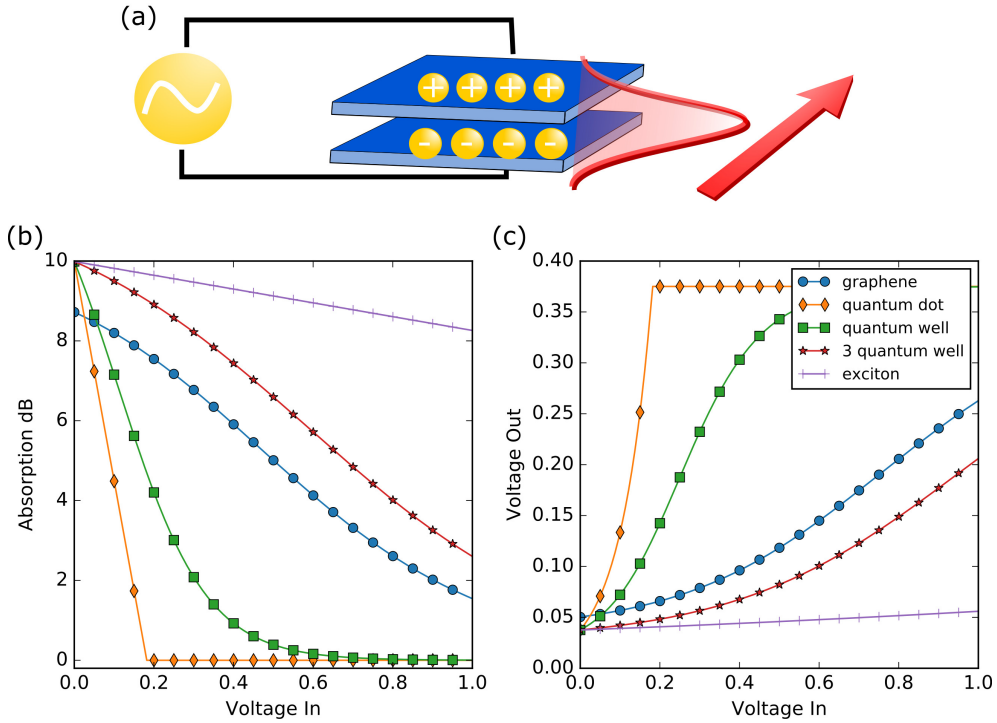

Fig. 2. (a) Schematic of a charge carrier-driven electro-absorption modulator. The optical mode and its propagation direction are shown in red. The modulator's nonlinearity use utilized in this work to provide the photonic perceptron's activation. (b) Modulator transfer function (absorption vs. their drive voltage (V_{in}) from the perceptron-summation photodiode. Color coded are five different electro-optic 'active' materials, whose response and equation-set was derived in [12]). (c) The photonic perceptron's nonlinearity varies by the type of tunable material used in the capacitively-biased modulator and translates into a voltage activation function (V_{out}) when a 100 μm modulator with gate oxide thickness $t_{ox} = 10\ nm$ is coupled through a 50 Ω current-to-voltage converting resistor to a photodiode (at the receiving neuron) with quantum efficiency of $\eta = 0.6$. These nonlinearities are used in the subsequent photonic neural network analysis below.

second method adds a transimpedance amplifier (TIA) to the circuit. This is the approach used by most optical receivers but is costly in terms of circuit complexity, power, and noise figure. The third method is capacitive coupling, connecting the photodiode directly to the modulator [1] (Fig. 3(c)). Here the photodiode acts as a constant current source to charge the capacitive modulator. The final method is to inductively load the photodiode to transiently convert the low voltage of the photodiode to a higher voltage to drive the modulator. Like resistive coupling, this method creates an LC filter and requires an inductor of sufficient magnitude to create the necessary transient voltage, limiting operating speed.

Of these three coupling methods, capacitive coupling is the most appealing because it only requires scaling capacitance and not resistance or inductance both of which can be minimized, increasing the maximum operating frequency of the circuit, and hence of the entire photonic NN.

To model the capacitively coupled electro-optic circuit, we begin with the optical power of the CW laser source. The electro-optic absorption modulator will then attenuate this optical signal with α (optical loss) as a function of voltage:

$$P_{out} = P_{cw} exp\left(-\alpha\left(V_{in}\right) L\right) \tag{1}$$

Optics EXPRESS

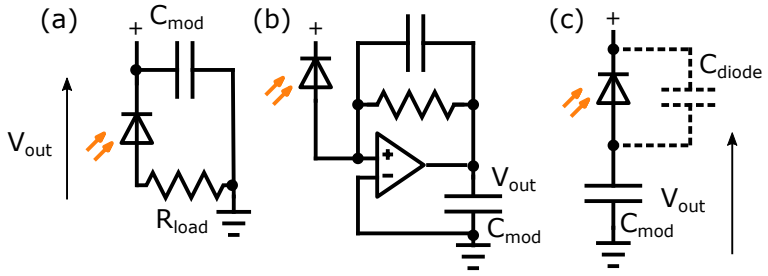

Fig. 3. Inside the photonic neuron (i.e. perceptron model Fig. 1), the photodiode is coupled to the modulator with a reverse bias either by a current to voltage converting resistive load, R_{load}, (a), a TIA (b), or directly coupling to the modulator as a capacitive load (c). When coupled with load resistance R_{load} must be large enough to to produce a voltage in the operating range of the electro-absorption modulator, limiting the RC frequency response of the circuit. For the NN results presented below, capacitive coupling of (c) was selected.

This optical power exiting the modulator creates an arrival rate of photons γ_p at the next NN layer's photodiode directly proportional to the optical power, P_{in}. The event rate in the interval is modeled with a Poisson distribution, and is proportional to optical wavelength λ_0 and inversely proportional to operating frequency f. The arriving photons act on the photodiode to move an elementary charge q with some quantum efficiency of η, to charge a circuit with total capacitance C, to ultimately reach a voltage change of ΔV_{out} at the time $1/f$, assuming the reverse bias is much greater than ΔV_{out}.

$$\gamma_p = \frac{\lambda_0}{fhc} P_{in} \tag{2}$$

$$\Delta V_{out} = P_{poisson}\left(\gamma_p\right) \frac{q\eta}{C} + V_{noise\_circuit} \tag{3}$$

From Eq. (3), we observe that adding an amplifier to the circuit, multiplying Eq. (3) by gain G, is equivalent to increasing the input laser power, ignoring noise. By adding optical gain G to the laser, the voltage root mean square (RMS) noise is increased by

$$\Delta N_{optical} = \sqrt{\gamma_p} \frac{q\eta}{C} \left(\sqrt{G} - 1\right) \tag{4}$$

To create equivalent gain with an electrical amplifier placed after the photodiode interfacing circuit, the increase in noise by the amplifier will be

$$\Delta N_{amp\_gain} = G_{amp}\left(N_0\right) - N_0 + N_{amp}, N_0 = \sqrt{\gamma_p} \frac{q\eta}{C} + N_{circuit} \tag{5}$$

Where N_0 is the original noise of the circuit, now amplified by the new electrical gain G. N_0 is composed of the optical noise and the noise of the circuit, $N_{circuit}$. It is interesting to note that adding a phase-insensitive electrical amplifier, such as the TIA used here, multiplies the original optical noise, while adding optical gain does not.

Looking just at noise and ignoring power we would like to identify the operation condition when it is beneficial to add an amplifier to the circuit vs. increasing the operating optical power. This happens when $\Delta N_{amp\_gain} < \Delta N_{optical}$. Then, solving for amplifier noise, N_{amp}, this is the region

$$N_{amp\_gain} < \sqrt{\gamma_p} \frac{q\eta}{C} \left(\sqrt{G} + N_{circuit}\left(1 - G\right)\right) \tag{6}$$

Optics EXPRESS

The electrical power consumption to produce optical gain is increased by $(G-1)\,P_0/\eta_{photodiode}\eta_{laser}$. These two limits create a noise and power budget, respectively, in which an amplifier is beneficial to the circuit.

While capacitive coupling offers an efficient means of interfacing charge-driven (capacitor) modulators to a photodiode, it will now act as an integrator. To bound the integration time, a clock cycle is introduced, dividing the alternating layers into a charging and a modulating cycle. In the charging cycle the integration is reset and the value of the lower modulating layer is captured in the capacitor. In the modulating cycle the capacitor charge is held to provide a steady optical power to the next higher layer. This effectively divides the throughput of the NN in half, with half of the layers held and half charging at any given time.

There are two options available for implementing a clock. The first is to add an electrical gate to isolate the capacitor from the photodiode after it has been charged. This method adds additional power and design complexity to drive the electronic gate at each modulator. The second method is to replace the CW laser source feeding each layer of the NN with a pulsed sources (e.g. source-gating). The pulses alternate between even and odd layers such that the lower layer is held while the upper layer is charged. This can be combined with an electronic gate to reset the layer at the beginning of the cycle or the modulator can be designed to leak charge at a set rate to reach a zero potential at the end of the held cycle.

4. Noise and cascadability

In neuromorphic photonics, the neuron must exhibit both nonlinearity and a sufficient SNR at each NN layer to produce a SNR greater than unity at the output of the final layer of the network. With a large number of layers, such as in deep-learning networks, the signal must cascade from layer to layer, and maintain a SNR greater than one at the output of the network within a reasonable power budget. This requirement bounds the type of modulator, operating power, and number of achievable layers, setting a multi-dimensional design space.

While the shape of the transfer function, including the nonlinearity, is primarily driven by the modulator type, the cascaded SNR of the electro-optic neuron is dependent upon several parameters. First, at the most rudimentary level, the optical power input generates a Poisson distribution in the quantized arrival time of the photons. Next, the interfacing circuit adds thermal noise, and potentially gain. Finally, the electro-optic modulator itself affects the cascaded SNR through modulation depth and the modulator's intrinsic nonlinearity.

Immediately apparent in the analysis is the SNR's dependence on each node's operating power. The output SNR of the lowest power optical nodes, those nodes with the smallest input, will be less than the highest optical power nodes, those nodes with the greatest input. The cascaded SNR of the system is then dependent on the input data and the trained weights of the network.

It follows that in optimizing device parameters this forces us to make assumptions about the statistics of the input data and the trained weights. In our analysis of cascaded SNR we assume that the input power to each neuron during operation is uniformly distributed in the nonlinear portion of the voltage transfer function. We define this portion of the transfer function as the region with slope greater than 0.1 between the minimum and maximum swept voltage (V_{in}). We then evaluate the SNR of the cascaded network in terms of the root mean square (RMS) power defined over the network as the square root of the mean variance of the signal, Eq. (7), for a uniformly distributed signal in this operation region. This definition of SNR power is similar to the SNR definition commonly used in image processing [14] except we define power as a deviation from the mean rather than a sum of intensity.

$$V_{rms} = \sqrt{\frac{\sum_{i=1}^{n}\left(V_i - \frac{\sum_{j=1}^{n} V_j}{n}\right)^2}{n}} \tag{7}$$

Optics EXPRESS

Using this definition of power of the entire network, the SNR was calculated assuming a uniform distribution in the range with and without noise. Next the signal was subtracted from the signal and noise to obtain an estimate of the signal and noise separately which were used to approximate the SNR of the network in dB terms with Eq. (8).

$$SNR_{dB} = 20\log_{10}\left(\frac{V_{rms\_signal}}{V_{rms\_noise}}\right) \tag{8}$$

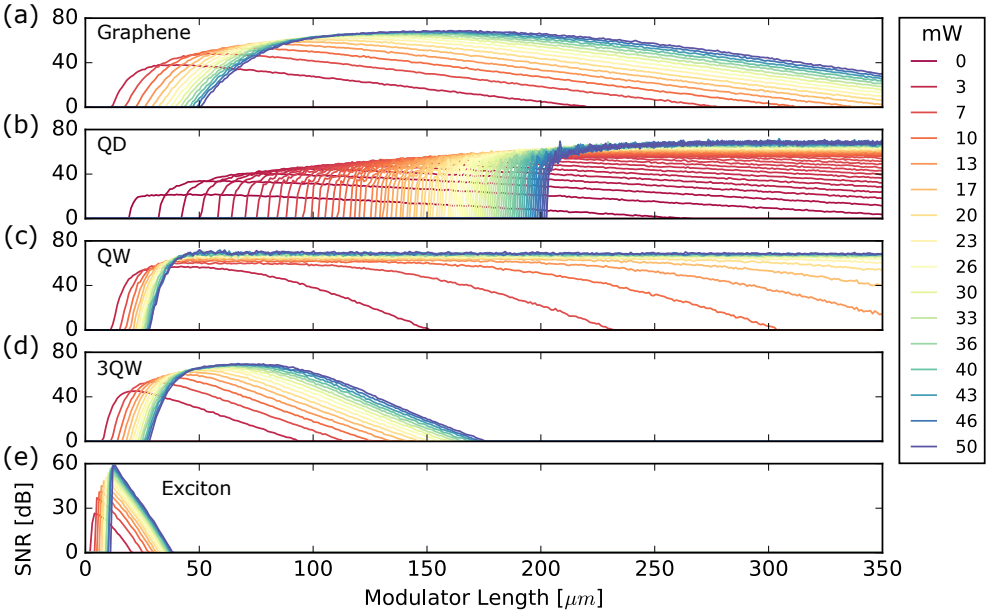

Fig. 4. The SNR of a neurons output after two NN layers for graphene (a), quantum dot (b), quantum well (c), three quantum well (d), and exciton (e) electro-absorption modulators against modulator device length over a range of optical powers from 0.01 mW to 50 mW shows low performance of graphene for low optical powers, almost no response for QD for optical power over 10 mW, a wide range of reasonable performance for QW, and a short region of peak performance for exciton, all for modulator lengths < 350 μm.

The SNR analysis for an exemplary 2-layer photonic NN shows distinct performance for a variety of modulator neurons (Fig. 4). While the results show a relatively similar obtainable SNR across the various modulators neurons, the modulator length and and power at which this SNR occurs is notably different. The exciton neuron, for instance, allows for most compact modulators (due to the high absorption), yet, also requires a high optical power to provide high SNRs. The QD case also requires higher optical power levels to achieve elevated SNRs, but since the absorption is relatively weak, also longer modulators. Notice, that increasing absorption-per-unit-modulator-length does not improve SNR for a same optical power and length (e.g. compare QW vs. multiple (3) QW layers), because the modulator became less sensitive to power changes from the higher absorption potential. This can be seen when comparing the QW, 3QW, and exciton cases, where each increases nominal absorption in that order.

5. Training

The electro-optic neuron produces an optical output from its weighted inputs. The power of this optical output must be distributed to the next layer's nodes. A naive layout would simply fan-out the output power to the next layer. This would create a $1/N$ power divider where N is the number

Optics EXPRESS

of nodes in the next layer. In this scheme, power is sent to every node in the next layer regardless of whether or not the weights in the next layer are set to accept that power. If the weights in the next layer are not set to accept the power they must absorb the power and the excess power is wasted. This severely limits the power available to the next layer. Alternatively, ring drop filters in a broadcast and weight network or an MZI network can be used to divide the power among the output layers proportionally to the trained weights [9]. In these configurable power dividing networks the total power ratio cannot exceed unity, i.e. no power is added ($G = 0$). This limitation requires enforcing a constraint on the weight matrix during training to keep the sum of the output weights of any node from exceeding unity. This is equivalent to enforcing a sub-unity L1 norm along the rows of a TensorFlow order (M Input x N Output) weighting matrix and is similar to the sub-stochastic matrices used in some Markov models [15].

6. Power analysis

The power dissipation of the capacitively coupled electro-optic modulator can be described by a charging and discharging a capacitor. The energy stored in each circuit is defined by the energy of the capacitance, $E = CV^2/2$. Each node must charge the modulator in each operating cycle. The total electric power dissipated (P_e) by the nodes of the network then is the charging and discharging of the capacitance of the modulator, photodiode, and gate at frequency f, $P_e = \sum_{k=1}^{N_{nodes}} \left(V_k^2 \left(C_{mod} + C_{photodiode} \right) + C_{gate} V_g^2 \right) \frac{f}{2}$ where V_k is the node voltage, C_{mod} is the capacitance of the modulator, $C_{photodiode}$is the capacitance of the photodiode, C_{gate} is the capacitance of the gate, and V_g is the gate voltage. In addition to the electrical power to operate the electro-optic modulator, the network requires a CW laser source of efficiency η_{laser} generating optical power to supply every node with enough optical power to sufficiently reach the necessary SNR found in the previous analysis (Fig. 4), $P_{laser} = N_{nodes} P_o / \eta_{laser}$. The total power consumption of the entire NN system then becomes (Fig. 5(b))

$$P_{total} = \sum_{k=1}^{N_{nodes}} \left(V_k^2 \left(C_{mod} + C_{photodiode} \right) + C_{gate} V_g^2 \right) \frac{f}{2} + \frac{N_{nodes} P_o}{\eta_{laser}} \tag{9}$$

With the goal to minimize the NN's power consumption to increase the MAC/J performance, aside from the trival engineering options to reduce the number of nodes in the network, or to use more efficient lasers, is to reduce the electrical capacitors of the electro-optic devices (modulator and photodiode). This may present a value proposition for light-matter-interaction enhanced devices such as plasmonics, or heterogeneous integration of emerging materials into photonics [3, 12, 13].

7. Methods

To gain further insights into performance of such photonic neuron-based NNs, we used Keras [16] to implement the custom activation function from each of the modulator types. These activation functions were composed of two parts: First, the photodiode coupling circuit, including noise variance over the time of the simulation step, is modeled with Keras's *random_normal* function, which adds a sampled random normal number at each simulated time period with mean and variance as parameters. Here the Gaussian noise mean is set to zero and the variance is modeled as a sum of shot noise from photon arrival and thermal noise of the circuit's capacitance (modulator plus photodiode). We assume that the photodiode capacitance is matched to the modulator's capacitance (Fig. 3(c)), given by the length modulator. For shot noise, the Gaussian model assumes that the Poisson distribution for photon can be approximated by the Gaussian for an arrival rate of $\gamma > 20$ photons per time period dependent on the NN's operating frequency f. The output of the circuit model is a voltage from the photodiode with noise, assuming integration over the cycle time.

Optics EXPRESS

The second half of the model defines how the modulator absorbs input light with the voltage derived from the first model. Here, we assume that the CW laser power is split directly from the laser to each of the EAMs in a network of branching power couplers. The EAM models were taken from our previous work [5] which define the absorption in dB per unit length of the modulator as a function of voltage, thus the amount of absorption is highly dependent on modulator length. Trivially, increasing the modulator length or the input laser power results in a higher modulated signal's SNR.

Finally, to simulate the WDM or interferometer optical network the standard implementation of the Keras NN must also be modified to require a sub-unity L1 norm weight matrix, requires that the sum of the output of each node must be equal to or less than one. We enforced this rule with a custom Keras *kernel_constraint* that clips the weight matrix to the range $[0, 1]$ and enforces a limit to the weight matrix such that the sum of each column (outputs of a node) is less than or equal to one.

8. Results and discussion

To test the performance of modulator neuron-based NNs, we turn to a simple image classification task; the MNIST dataset [17] is a set of images of handwritten digits in a grayscale 28x28 pixel format that is commonly used in comparing neural network performance. The NN is trained to classify the images into the 10 individual digits, and then evaluate prediction accuracy using a Python-developed code in Keras [16] and TensorFlow [18]. For this photonic NN we selected two hidden layers each with 150 nodes, a common configuration for minimal deep learning networks [19–22]. The weights are bound between zero and one to simulate input optical weighting by ring or interferometric modulators [7, 9]. The network is initialized with zero weights and trained with the Adagrad [23] method, the categorical cross entropy loss function, a learning rate of 0.005, 45 training epochs, and no decay. A row sum (L1 less than unity) constraint was placed on the TensorFlow ordered weight matrix during training to enforce power conservation in the optical weighting network. The simulated input laser power is swept (0.01-50mW), and for each optical power the modulator length is selected by maximizing the SNR.

The results (Fig. 5(a)) show that quantum well and quantum dot modulation outperforms graphene and exciton modulation in terms of accuracy in the low laser power limit. As optical power is increased, the graphene and exciton modulation approaches the accuracy of quantum well neurons. However, the latter performs well over a broad range of optical power inputs. In terms of capacitance and its effect on NN operating power, the quantum dot modulator outperforms all other modulators since it has the steepest transfer function at lowest drive voltage. (Fig. 5(b)).

Optics EXPRESS

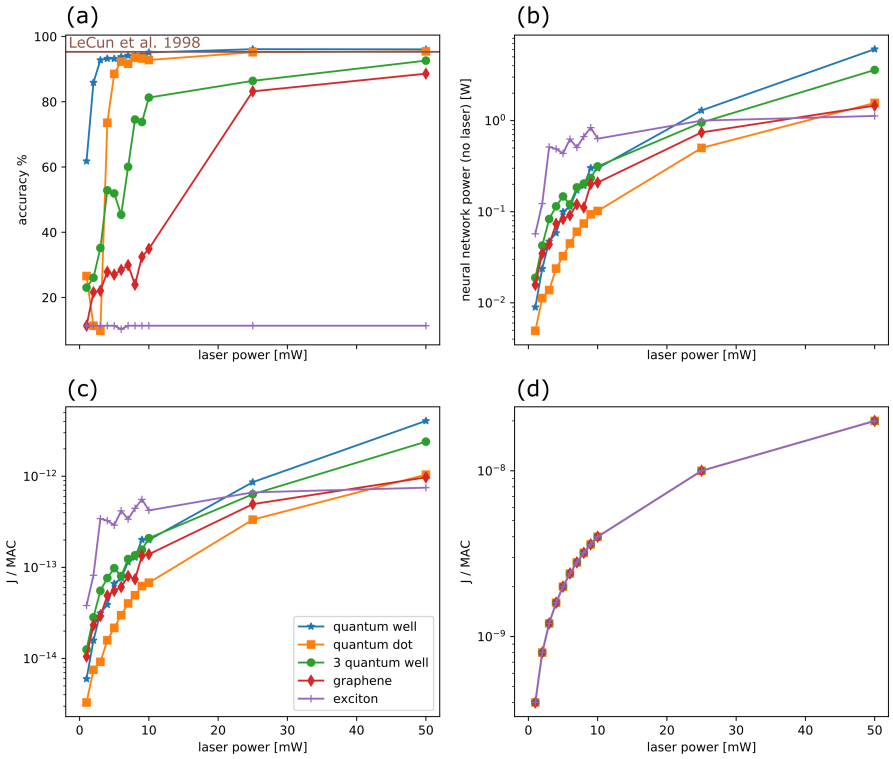

Fig. 5. Simulation of a 300 hidden node MNIST classification neural network with two hidden layers of 150 nodes each swept over range of laser optical powers from 0.01 mW to 50 mW show accuracy results (a) converging across modulator types, except exciton, to approach digital computer accuracy for a similar 300 hidden node network [19] as the optical power exceeds 30 mW. At lower power levels modulator types vary in performance with the quantum well modulator outperforms the others in terms of accuracy. Power dissipation, excluding electrical power to drive the CW laser and clocking overhead, (b) shows the quantum dot modulator consuming the least power. In energy per operation terms (c) QW performed at 93% accuracy using 6.3×10^{-14} J/MAC and QD at 92% accuracy using 6×10^{-14} J/MAC, both excluding laser power. Including laser power with wall plug efficiency of 50% (d) brings the J/MAC up significantly demonstrating that the power cost is dominated by laser efficiency. All results are with NN operating speed = 10 GHz (gated to 5 GHz), and training with 45 epochs of the Adagrad [23] method.

In conclusion, we find that the nonlinear transfer functions of electro-optical modulators, specifically electro-absorption modulators, are sufficient to generate accurate (>90%) feed-forward inference results in neural networks with a throughput of tens of GHz, low latency, and a power budget in the order of several watts. Training the photonic modulator-neurons for an exemplary MNIST image classification prediction task using our developed noise-material-capacitor-circuit model, shows a multi-parameter optimization space including (from front to back-end) optical laser power, modulator capacitor, signal-to-noise-ratio. Our results show that increasing the optical absorption-per-unit length improves SNR sensitivity, but not necessarily image classification accuracy when given a relatively low optical power budget (e.g. few mW). These attributes make the electro-optic neural network particularly appealing for small node networks in latency demanding applications such as communications and LIDAR.

Optics EXPRESS

Funding

National Science Foundation (1740262, 1740235); Semiconductor Research Corporation (SRC) (nCORE, E2CDA)

References

1. D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," J. Light. Technol. **35**, 346–396 (2017).
2. M. de Cea, A. H. Atabaki, L. Alloatti, M. Wade, M. Popovic, and R. J. Ram, "A thin silicon photonic platform for telecommunication wavelengths," in *2017 European Conference on Optical Communication (ECOC),* (2017), pp. 1–3.
3. V. J. Sorger, R. Amin, J. B. Khurgin, Z. Ma, H. Dalir, and S. Khan, "Scaling vectors for attojoule per bit modulators," J. Opt. **20**, 014012 (2018).
4. C. Ye, K. Liu, R. Soref, and V. J. Sorger, "A compact plasmonic mos-based 2x2 electro-optic switch," Nanophotonics **4**, 261–268 (2015).
5. R. Amin, C. Suer, Z. Ma, I. Sarpkaya, J. B. Khurgin, R. Agarwal, and V. J. Sorger, "Active material, optical mode and cavity impact on nanoscale electro-optic modulation performance," Nanophotonics **7**, 455–472 (2018).
6. H.-T. Peng, M. A. Nahmias, T. Ferreira de Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic integrated circuits," IEEE J. Sel. Top. Quantum Electron. **24**, 1–15 (2018).
7. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441 (2017).
8. A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," J. Light. Technol. **32**, 3427–3439 (2014).
9. A. N. Tait, T. F. Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Reports **7**, 7430 (2017).
10. K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems,* (2016), pp. 586–594.
11. J. K. George, H. Nejadriahi, and V. J. Sorger, "Towards on-chip optical ffts for convolutional neural networks," in *2017 IEEE International Conference on Rebooting Computing (ICRC),* (IEEE, 2017), pp. 1–4.
12. C. Haffner, W. Heni, Y. Fedoryshyn, J. Niegemann, A. Melikyan, D. L. Elder, B. Baeuerle, Y. Salamin, A. Josten, U. Koch, C. Hoessbacher, F. Ducry, L. Juchli, A. Emboras, D. Hillerkuss, M. Kohl, L. R. Dalton, C. Hafner, and J. Leuthold, "All-plasmonic mach–zehnder modulator enabling optical high-speed communication at the microscale," Nat. Photonics **9**, 525–528 (2015).
13. R. Amin, J. B. Khurgin, and V. J. Sorger, "Waveguide based electro-absorption modulator performance: comparative analysis," Opt. Express **26**, 15445–15470 (2018).
14. R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*, vol. 624 (Pearson-Prentice-Hall, 2004).
15. W. E. Pruitt, "Eigenvalues of non-negative matrices," Ann. Math. Stat. **35**, 1797–1800 (1964).
16. Keras, "Keras: the python deep learning library," https://keras.io (2015).
17. Y. LeCun and C. Cortes, "Mnist handwritten digit database," http://yann.lecun.com/exdb/mnist/ (2010).
18. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.
19. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2324 (1998).
20. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv **1312**, 6199 (2013).
21. K. Chellapilla, S. Puri, and P. Simard, "High performance convolutional neural networks for document processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition,* (Suvisoft, 2006).
22. R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems,* (2013), pp. 315–323.
23. J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," Tech. Rep. UCB/EECS-2010-24, EECS Department, University of California, Berkeley (2010).