Research Article | Vol. 8, No. 12 | 1 Dec 2018 | OPTICAL MATERIALS EXPRESS 3851
Optical Materials EXPRESS

All-optical nonlinear activation function for photonic neural networks [Invited]

MARIO MISCUGLIO,[1] ARMIN MEHRABIAN,[1] ZIBO HU,[1] SHAIMAA I. AZZAM,[2] JONATHAN GEORGE,[1] ALEXANDER V. KILDISHEV,[2] MATTHEW PELTON,[3] AND VOLKER J. SORGER[1,*]

[1]*Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA*
[2]*School of Electrical & Computer Engineering and Birck Nanotechnology Center, Purdue University, West Lafayette, Indiana, 47907, USA*
[3]*Department of Physics, UMBC (University of Maryland, Baltimore County), Baltimore, Maryland 21250, USA*

Abstract: With the recent successes of neural networks (NN) to perform machine-learning tasks, photonic-based NN designs may enable high throughput and low power neuromorphic compute paradigms since they bypass the parasitic charging of capacitive wires. Thus, engineering data-information processors capable of executing NN algorithms with high efficiency is of major importance for applications ranging from pattern recognition to classification. Our hypothesis is, therefore, that if the time-limiting electro-optic conversion of current photonic NN designs could be postponed until the very end of the network, then the execution time of the photonic algorithm is simple the delay of the time-of-flight of photons through the NN, which is on the order of picoseconds for integrated photonics. Exploring such all-optical NN, in this work we discuss two independent approaches for implementing the optical perceptron's nonlinear activation function based on nanophotonic structures exhibiting i) induced transparency and ii) reverse saturated absorption. Our results show that the all-optical nonlinearity provides about 3 and 7 dB extinction ratios for the two systems considered, respectively, and classification accuracies of an exemplary MNIST task of 97% and near 100% are found, which rivals that of software based trained NNs, yet with ignored noise in the network. Together with a developed concept for an all-optical perceptron, these findings point to the possibility of realizing pure photonic NNs with potentially unmatched throughput and even energy consumption for next generation information processing hardware.

1. Introduction

The past decades have been marked by an exponential increase in demand for high speed and energy efficient computer architectures. The established microelectronics technology faces its biggest limitations with respect to handling real-time processing, large data volumes and complex systems. The exponential growth in transistor count and integration, which marked the previous century, cannot be pushed further, highlighting the twilight of Moore's law for CMOS technology. In this changing landscape, multiple companies and research groups are pursuing novel solutions to provide high-performance computing to the next level, using better performing algorithms and novel technological approaches, where the work presented here falls under the latter. A widely investigated solution is to replace general-purpose processors (von Neumann architecture) with more specialized and task-specific processors. Graphic Process Units (GPUs) or Field Programmable Gate Array (FPGAs), architectures, which are optimized for data computation and parallelism provide efficiencies for their specialized tasks up to an order of magnitude higher than generic-task CPUs [1–3]. However,

#347314 https://doi.org/10.1364/OME.8.003851
Journal © 2018 Received 2 Oct 2018; revised 10 Nov 2018; accepted 10 Nov 2018; published 21 Nov 2018

Optical Materials EXPRESS

these approaches still rely on electronic transport and are bound by the speed and power limits of the interconnects inside the circuits due to RC parasitic effects.

A subclass of algorithms which receives benefits when implemented in non von Neumann architecture is represented by Neural Networks (NN). NNs are designed to simultaneously process large arrays of data in order to learn representations of them with multiple levels of abstraction. These architectures consist of neurons that perform the following basic functions: 1) Interpret multiple incoming signals in a form of multiple arrays (e.g. images) through weighted addition (Multiply and Accumulate, MAC); 2) Apply a nonlinear (NL) activation function for discriminating the data; 3) Transmit the result to multiple destination neurons (i.e., fan-out).

Recently emerging Photonic Neural Networks (PNN) demonstrated the potential to increase computing speed by 2-3 orders of magnitude [4]. In order to implement the functions of the NN into a PNN, two classes of devices and their respective functions need to be engineered, the weighted sum and the NL activation. The weighted sum, as previously investigated [5], can be obtained by a combination of balanced photodetection, Mach-Zehnder interferometer, or ring resonator, which constitutes the weights of the signal, and y-junctions or MMI (Multimode interference coupler) for summation. The weighted addition in analog photonics, which is the equivalent of the MAC, uses optical interference, while the coherent electromagnetic waves propagate through the photonic integrated circuit (PIC), and offers MAC energy consumption that does not trade-off with MAC speed [6]. Moreover, photonic interconnects can have large bandwidth and low power consumption.

On the other hand, the absence of a straightforward and efficient NL in optics has severely limited its usefulness in DeepLearning computing. Several attempts [7–9] showed that a NL modulation of the optical signal could be achieved through Electro-optic Absorption Modulators (EOMs). These rely on the on-chip implementation of an electronic platform that is able to vary the optical effective mode index by tuning the voltage. Despite their relatively straightforward implementation and controllability, EOM results in substantial trade-offs in terms of speed and power efficiency of the otherwise intrinsically instantaneous transmission of the signal through the PIC.

Current NL activation function unit mechanisms [10] are based on devices that would need first to convert an optical signal into an electrical signal by means of a detection mechanism and afterwards convert it back into an optical signal, a solution that would hamper the speed and cascadability of the network due to both the movement of charge carriers and noise (e.g. shot and thermal noise of a photodetector). One of the most commonly used optical nonlinearities unit, in reservoir optical computing [11] and more recently in feed-forward neural network [12] is represented by saturable absorbers, such as graphene layers [13] or based on 2-photon absorbtion [14]. Other mechanisms are instead based on the nonlinearities of bistable switches [15] and ring resonators [16] have also been investigated. Approaches based on single graphene excitable laser [17] have recently shown significant progress in the field of all optical spiking neural networks. Nevertheless, the integration of the NL optical modules in photonic circuit, the enhancement of the modulation strength and operational speed and consequently the effectiveness both at the device and system level, represent still an open challenge. For this reason, the implementation of fast and energy efficient all-optical nonlinearity becomes a key task for boosting the throughput of a neural network and consequently lowering the latency and power dissipation.

In this work, we aim to overcome these limitations by describing, through a combinatory approach of electromagnetic simulations and network emulation, novel devices and approaches based on light matter interaction (LMI) in assembly of nanoparticles in PICs. A first device prototype relies on a reversible transparency induced by a Fano resonance in a plasmon-exciton system. This system consists of a semiconductor nanocrystal, i.e. quantum dot (QD), within two gold nanorods. We also studied the NL response induced by reverse saturable absorption in films of buckyballs (C60). The optical response of the two hybrid

Optical Materials EXPRESS

systems shows a strong non-linearity suitable for being employed as activation functions in PNN circuits as investigated here. We find the optimal configurations that optimized the mode coupling between the signal transmitted in a waveguide and the hybrid structures using numerical tools. Finally, a NN architecture that exploits the proposed activation functions has been emulated on an open source machine-learning framework, i.e. Tensorflow. We then compared the performances in terms of speed and accuracy of the proposed system with others, which employ conventional nonlinear functions (e.g. Tanh, Sigmoid, ReLu), for a specific DeepLearning tasks. Ultimately, this work points towards further use of NL activation functions of optically triggered devices, integrated with silicon photonics, for the next generation of photonic processors.

2. Discussion

As a first step in our study we investigate the LMIs in hybrid systems consisting of a pair of gold nanorods separated by a 10-nm gap that contains a single CdSe Quantum Dot (QD) Fig. 1(a). Despite the apparent complexity, these quantum-assemblies could be either fabricated through a bottom-up approach, i.e. colloidal synthesis, as well as fabricated through a combination of top-down and bottom-up approach, by combining electron beam lithography and guided colloidal deposition [18]. As reported before [19–22], and illustrated in Fig. 1(a), the hybrid system could be modeled as two coupled oscillators, one representing the dipole of the QD transition, and the other representing the dipole of the plasmon. The surface plasmon is driven by the incident field, and the QD is driven due to its coupling with the plasmon. Destructive interference between the two oscillators leads to cancelation of the optical response of the coupled system, known as an induced transparency or a Fano resonance.

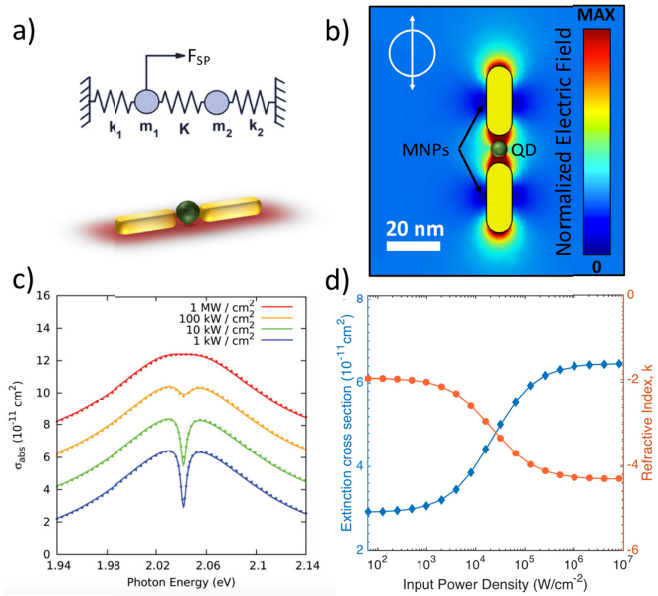

Fig. 1. Physical modeling and material characterization: a) Phenomenological description system. Two coupled oscillators that provides a classical analogue of the plasmon-exciton coupling induced transparency in the represented system and schematic representation of the MNPs/QD system consisting of a single quantum dot (QD) between a pair of gold nanoparticles (MNP). b) Normalized electric field distribution of the hybrid system computed for a 2.04 eV impinging plane wave. Scale bar 20 nm. c) Calculated absorption spectra for a system illustrated in the inset of (a) taken from [23]. The different curves show the response of the system for different amounts of energy in an incident laser pulse. d) Extinction cross section (left y-axis) and imaginary part of the refractive index (right y-axis) of the MNPs/QD system, as function of the input power density.

Optical Materials EXPRESS

Figure 1(b) displays the normalized electric near field distribution computed at 2.04 eV, in resonance with the gold rods of the proposed MNPs/QD system. The model parameters for the metal MNPs/QD system are obtained from FDTD simulations [20]. The complex refractive index of Gold and CdSe QD were obtained from refs [24] and [25]. The large electric field in the feed-gap of a gold optical antenna [26], generated by the plasmonic dipole resonances of the nano-rods, completely engulfs the CdSe QD, inducing the coupling between the plasmon and the QD that ultimately gives rise to the Fano resonance, which shows the obtained absorption spectra of the hybrid system for different incident optical powers, taken from ref [21] (Fig. 1(c)). It is worth mentioning that the data in Fig. 1(c) are based on the assumption of low (liquid-nitrogen) temperatures. This calculation, however, represents still a conservative result given the recent experimental results obtained at room temperature, where the Fano resonance, or transparency dip, is apparent for low incident optical power [19,27]. The assumption of low temperature can be in fact relaxed enhancing the plasmon to exciton coupling by either optimizing the geometry of the plasmonic nanostructure or controlling the interparticle spacing.

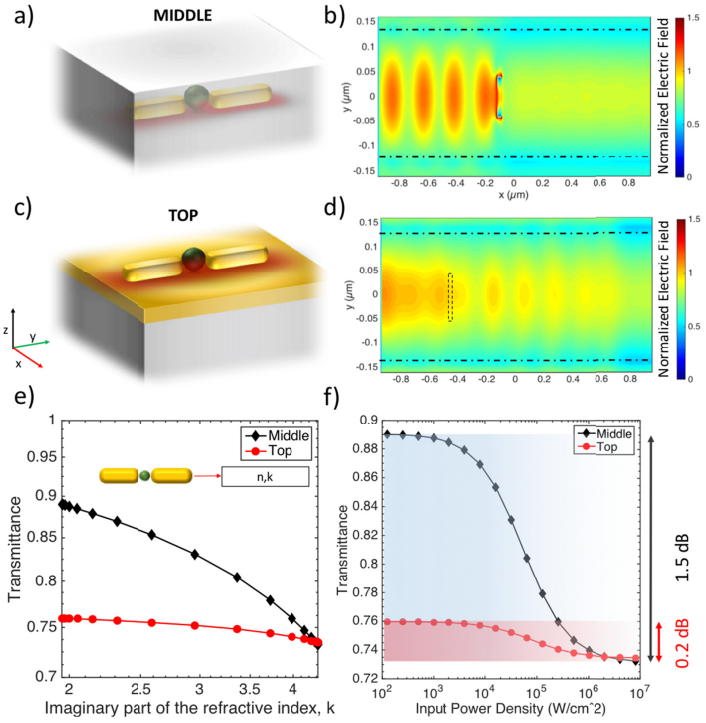

Fig. 2. Hybrid-Nanoparticle waveguide integration for an all-optical nonlinearity for photonic neural networks: a-c) Schematic representation of waveguide and MNPs/QD system coupling. The MNPs/QD system placed in the middle (a) and top (c). b-d) Normalized electric field distribution for a middle horizontal cut plane of the waveguide considering the assembly being placed in the middle (b) and on top (d), for its maximum absorption (highest *k*). c) Computed transmittance of the waveguide as function of the tunable absorption of a MNPs/QD system placed in the middle (Black solid line) and on top of the waveguide (Red solid line). d) Computed waveguide transmittance as function of the input power density and respective nonlinear modulation ranges in dB.

As the incident power increases, the Fano resonance dip disappears, due to saturation of the QD transition. Thus, the amount of energy dissipation in the coupled system is a NL function of the input power. This nonlinearity can be seen in the extinction cross section (left

Optical Materials EXPRESS

y-axis) spectra of the assembly computed at 2.04eV, as function of the input power density (Fig. 1d), suggesting a narrow spectral operation band. In order to later embed the system in a photonic waveguide, the imaginary part of the refractive index of the assembly as function of the input power density, was derived using a simplified geometry and Mie Scattering model [28] (right y-axis), which manifests a specular behavior.

We proceed further in our study by embedding the MNPs/QD system in a silicon photonic waveguide and studying their interaction through a finite-difference time-domain (FDTD) solver. The previously described physical model was investigated in two different device configurations, placing the assembly in the middle and on top of the waveguide, as schematized in Fig. 2(a) and (c), respectively. When placed on top, a 50 nm thin gold layer was used for enhancing the mode overlap between the assembly and the transmitted waveguide signal.

Figure 2(b) and (d), maps the normalized electric field travelling within the waveguide in the horizontal plane. Here, the absorbance of the assembly is set to its maximum; i.e., the imaginary part of the refractive index, *k*, is maximized. In this case, the overall transmitted power is similar for both configurations. As shown in Fig. 2(e)-(f), we evaluate these configurations by analyzing the modulation range of the transmitted power as a function of the imaginary part of the refractive index (e), which, for its part, depends on the input power transmitted in the waveguide (f). It may be observed that in both the configurations, the sole presence of an assembly produces a tangible NL modulation as function of the input optical power, while showing a sigmoidal shape. As expected, the configuration with the MNPs/QD system in the middle gives rise to a larger modulation range, reaching ~1.5 dB compared to the 0.2 dB obtained with a particle placed on top.

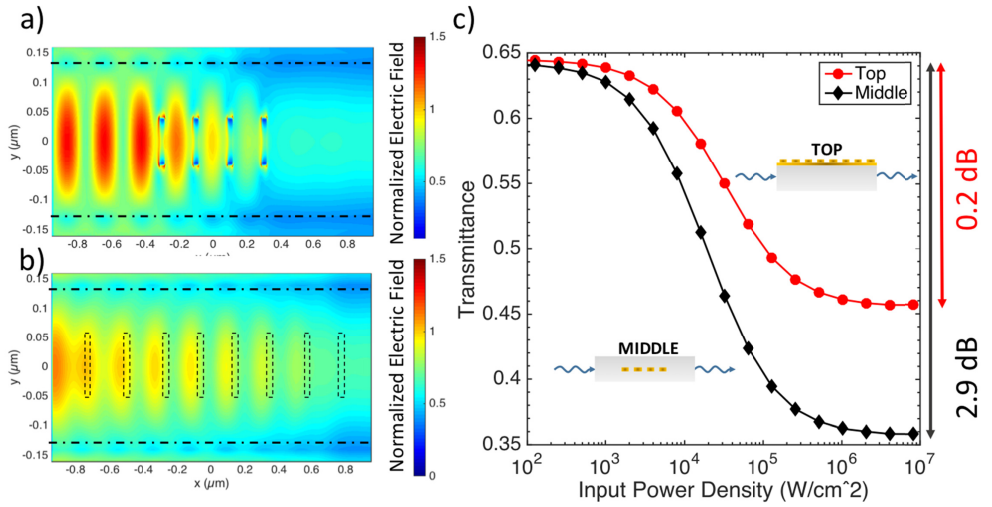

Fig. 3. Engineering the all-optical nonlinearity via array: a-b) Normalized electric field distribution for a middle horizontal cut plane of the waveguide, considering an array of closely spaced hybrid-structures, being in the middle (a) and on top (b) of the waveguide, for their maximum absorption (largest k). (c) Transmittance as function of the input power density.

Proceeding further with our analysis, in order to enhance the modulation range and consequently to introduce larger NLs in the transmitted signal, we consider the system composed by arrays of MNPs/QD assemblies that are distributed either on top or in the middle of the waveguide. Similarly to what is shown for single assemblies, Fig. 3(a,b) shows the electric field distribution in the horizontal cut-plane of the waveguide, when either 4 or 8 assemblies are placed in the middle and on top. The assemblies are spaced with a 200 nm pitch in both configurations. In both the configurations, the electric field transmitted at the

Optical Materials EXPRESS

end of the waveguide is significantly attenuated, when the assemblies retain their maximum *k*. More specifically, the modulation depth now approaches 50% for the top configuration and 65% for the middle configuration. Figure 3(c) shows the NL modulation range of the proposed devices as function of the input power density. The NL can affect the transmitted power by more than 2 dB in the top array configuration and almost 3 dB in the middle array configuration. The nonlinear modules' performance, hereby presented, are comparable in terms of energy threshold to those of commercially available bulkier resonant absorbers, regularly used for free space or fiber coupling. Nevertheless, the underlying physical mechanisms are substantially different as well as the strength of the nonlinear absorption change for comparable input power. Moreover, the relaxation time of commercially available saturable absorbers is of the order of tens or hundreds of ps, while the proposed NL unit, based on induced absorption, would have a femtosecond response, thus shortening the delay of this feed-forward NN, which is governed by the photon's time-of-flight through the system since no O-E-O conversion are needed.

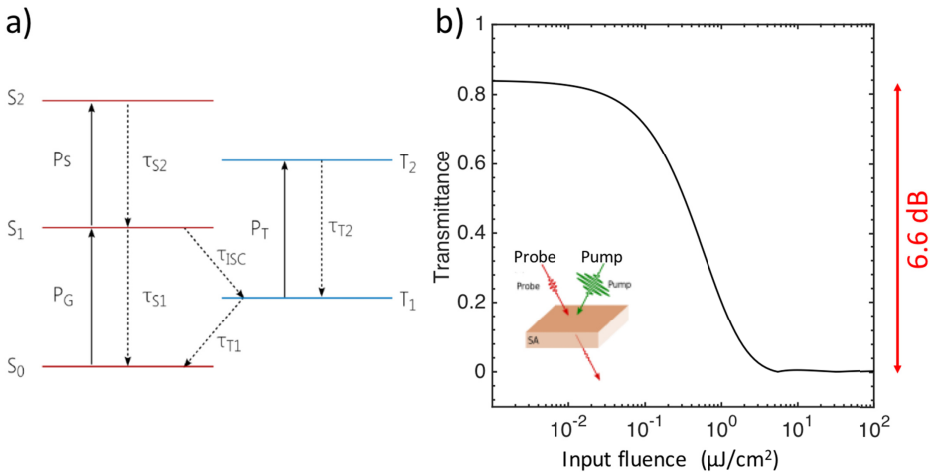

Fig. 4. Optical NL in a Reverse Saturable Absorber. Nonlinear transmission of a reverse saturable absorber made of high-concentration C_{60} in a PVA host thin film. **(a)** The simplified band-diagram of the reverse saturable absorber modeled by a five-level system. **(b)** Transmission vs. input fluence at 532 nm. Pump/Probe parameters: full-width at half maximum of the pump pulse = 1 ps and of the probe = 5 fs. Both pump and probe are centered at a wavelength of 532 nm. The lifetimes, τ_{S1} = 30 ns, τ_{T1} = 280 μs, τ_{ISC} = 1.2 ns, $\tau_{S2} = \tau_{T2}$ = 1ps.

Another possibility for inducing optical NL in a PIC is exploiting reverse saturable absorption (RSA). RSA is a property of a material for which the absorption increases as the light intensity increases. The response of the coupled MNPs/QD assemblies is an ultra strong, "artificial" RSA that arises from the Fano interference between the components. However, a broad range of materials show "natural," albeit weaker, RSA, such as organic compounds [29,30], heavy metals [31], and clusters of metallic particles [32]. In analogy to the discussion before, we next consider the use of C_{60} (Buckminsterfullerene or Buckyball) dispersed in polyvinyl alcohol (PVA) as a reverse saturable absorber to obtain an intensity-dependent transmission as a NL activation function. The band diagram of the majority of reverse saturable absorbing materials can be approximated by five energy levels representing two singlet and one triplet states, as depicted in Fig. 4(a). We use rate equations to model the transitions among the different energy levels and to capture the carrier kinetics in the system and model the nonlinear light-matter interaction. Full details of the model can be found in Ref [33]. The model is integrated with a finite-difference time-domain (FDTD) solver using an

Optical Materials EXPRESS

auxiliary differential equation approach to perform full-wave multiphysics analysis of the structures using the same structured grid and time-stepping.

A pump-probe analysis is used to extract the nonlinear transmission where a strong pump signal first illuminates the structure followed by a weak probe signal to probe the system response (Fig. 4(a)). The respective lifetimes are taken from optical characterization [30]. The concentration of the C_{60} molecules is chosen to be 10 mM to provide sufficient absorption from a thin film. The transmission of 1 μm-thin film of C_{60} in PVA (refractive index ~1.45) is calculated as a function of the input fluence and shown in Fig. 4(b). The linear transmission of the film is 0.84, and the saturation fluence is about 0.67 μJ/cm^2. We can also conclude that the NL modulation associated with the input power is approximately 7 dB. Given a feasible integration in photonic circuit, the smaller footprint and the completely passive strong NL modulation, the proposed devices could represent valuable candidates for optical nonlinear units in optical neural network.

After obtaining a small library consisting of 3 different NL optical responses, we introduce these as nonlinear activation functions on a standardized NN training set, MNIST classifiers of handwritten digits.

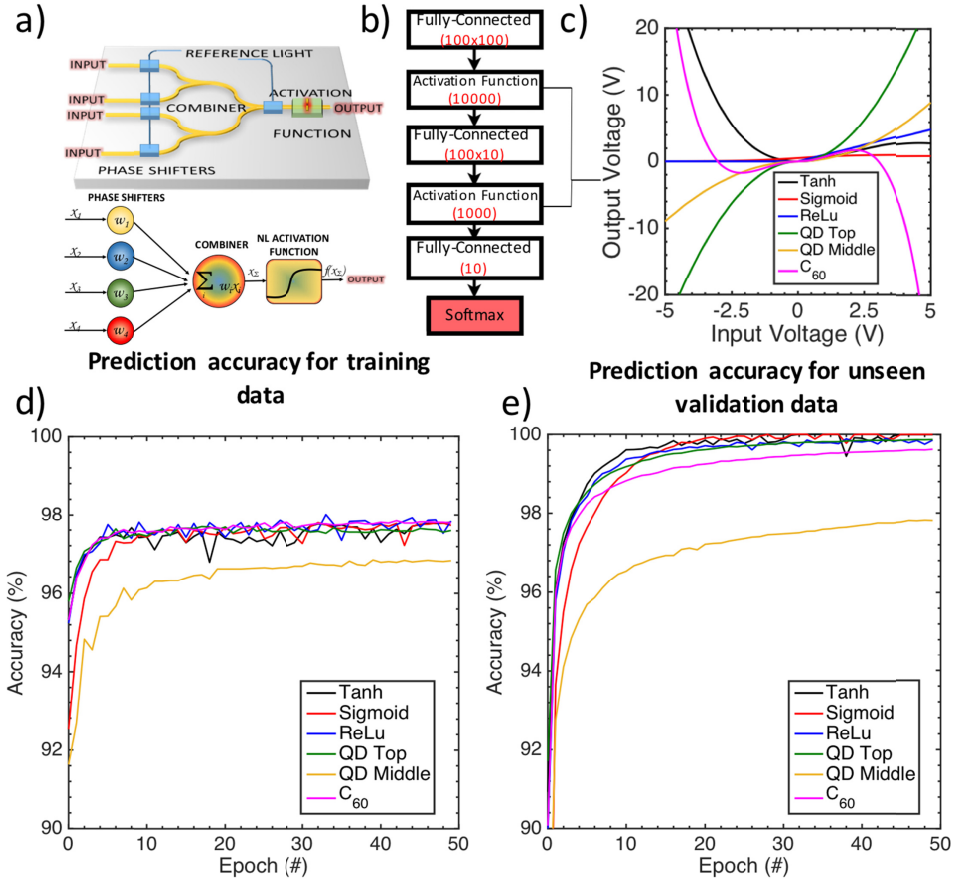

Fig. 5. Design of an AO neuron and evaluation of the Neural Network performance a) Schematic of an all-optical neuron. b) Representation of the emulated NN (b). c) Different activation functions, V_{IN} to V_{OUT}, ReLU (light blue dashed line), Sigmoid (blue dashed line), hyperbolic tangent (dark blue line) and the 2 proposed NL function, top (solid orange) and middle (solid red) d-e) Prediction accuracy as function of epoch for the training (d) and validation (e) phase of the network.

Optical Materials EXPRESS

Figure 5(a) shows the overall scheme of the AONN using photonic integrated circuits. As previously reported [5,34], the input optical signals are weighted by phase shifters and integrated by photonic combiners. Hereby, the NL activation functions (AF) can be achieved either by means of induced transparency through plasmon-exciton coupling or by reverse saturable absorption of C_{60}. The proposed all-optical NL modulation of the signal presents several advantages over an electro-optical counter part such as no charge fluctuation-induced noise or ps-short response times.

In the successive steps, we evaluate the functionality of the proposed NL optical response of the studied approaches as neural AF, by emulating their behavior in a 3-layer fully connected NN implemented in the Google Tensorflow tool and for the MNIST data set (details presented in the methods section). The software implementation MNIST is a known machine-learning data set comprised of 60,000 grayscale images of handwritten digits. The task is to identify the representing numbers for each image. In the MNIST data set each image has 28x28 pixel resolution. In order to feed each image to a fully-connected layer we flatten each 2D image into a vector of 784 pixels.

The first layer of the network is comprised of 100 neurons. As the name suggests, fully-connected networks have all-to-all connections among the neurons in each layer and the incoming inputs. Thus, the first layer requires 784x100 connections. These are identity connections and do not perform any type of weighting to the inputs. The second layer in our network also has 100 neurons, which receive inputs from the first layer. With all-to-all connections, that translates into 100x100 connections. It should be noted that the NL AFs are placed between two consecutive layers on each input connection. As a result, we will have 100x100 NL AFs operating between the first and the second layer. The third layer in our network has only 10 neurons, corresponding to 10 classes of output representing 10 digits from 0 to 9. This last layer can be regarded as a summarizing layer that reduces the dimension of the network output to the 10 required dimensions. For all the layers explained so far, an optical realization is envisioned. However, almost all modern NNs take advantage of a "Softmax" function implemented at the final layer. Softmax converts the incoming input into a probability distribution function with one incoming value receiving a high probability and the remainders receiving much lower probabilities. We have yet to envision an optical realization for this layer, but it is worth noting that this layer can be replaced with similar AFs at the cost of lower accuracy.

We train our network with two sets of AFs, namely, optical AF and commonly used AF in DeepLearning applications. The optical AF are those calculated for MNPs/QD assemblies on top of the waveguide (QD-Top), assemblies in the middle of the waveguide (QD-Middle), and C_{60} films. Respectively, the commonly used software-based AF we used to compare our photonic ones against the Rectified Linear Unit (ReLU), Sigmoid, and Tanh (Fig. 5(b)) [35]. It is worth mentioning that the actual input and output units of our optical AFs are μWatt. The shown optical AF are for devices of size of 80nm x 30nm. In order to obtain the mathematical model transfer function of the AFs to be used in Tensorflow, we fit quadratic curves to data points from the device simulations, with maximum Root Mean Squared Error (RMSE) due to fitting of 0.23 μWatt. The quadratic AFs are then reconstructed in Tensorflow. It should be noted that the NL AFs acts on the optical power of the incoming electromagnetic radiation associated with the photon flux, which quadratically dependents on the electric field. Moreover, in this study we aim to validate the NN performance by focusing on the shape of the AFs, without taking into account the noise on both system and device level, which will be the subject of an in-depth future investigation. Amplitude and phase noise induced by the NL device, as well as noise in the input signal and weights, could indeed seriously compromise the ability of the AF to discriminate amongst the data, hindering the accuracy of the solution.

We randomly split the MNIST data set into two subsets of 50,000 and 10,000 images for training and validation respectively. The network was initially set to train for 50 epochs; however, we can see that for QD-middle and C_{60}, the model started to over-fit and the training

Optical Materials EXPRESS

and the validation accuracy dropped. Figure 5(c) depicts the accuracy as the network is being trained for 50 epochs, showing that all of the optical AFs are comparable to those of the software-based networks in terms of the accuracy. QD-Middle is the only one performing relatively poorly, converging to a maximum accuracy of 96%. The validation data accuracy corroborates that of training data, as depicted in Fig. 5(d). For all the optical NL activation functions a ~100% accuracy is reached, apart from the QD-middle configuration, which reaches a plateau at 99%. It is worth mentioning that the accuracy curves for the validation data are smoother due to the fact that at each validation epoch, the network is evaluated over the whole set of validation images. In contrast, during training we used batch processing and a variation of gradient descent termed Stochastic Gradient Descent (SGD) for training. As a result, at each epoch during the training the network only receives a subset of training images. This results in slight variation of accuracy for different batches of data for different epochs.

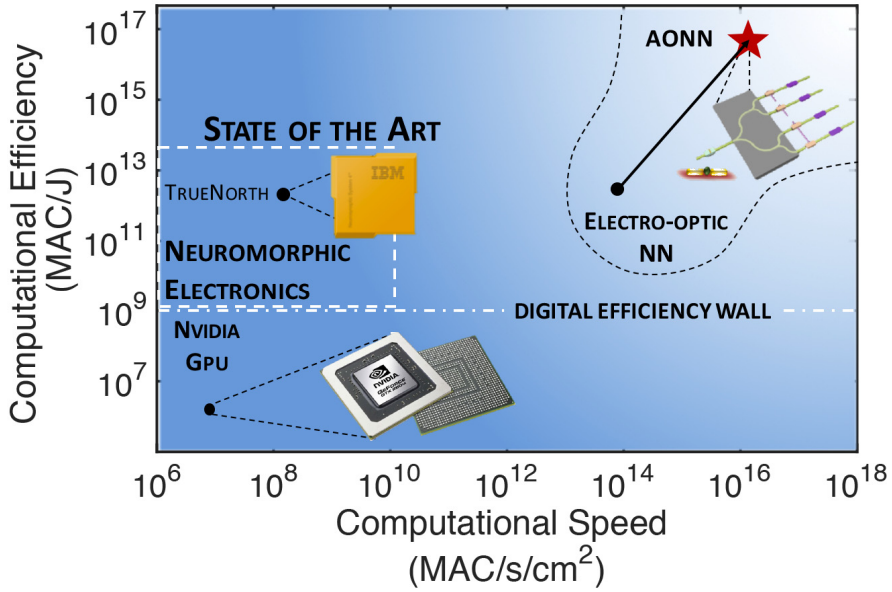

Fig. 6. Comparison of computational energy efficiency and processing speed between existing electronic neuromorphic demonstrations and our proposed programmable photonic platform. NN = neural network, AONN = all-optical NN, GPU = graphical processing unit.

The proposed activation mechanisms based on NL LMI, showed compatible performance, in terms of accuracy as function of epoch with respect to the well-established NL activation functions known from the software-based machine-learning community. The proposed architecture offers possible benefits of the absence of parasitic switching and short delays, since the run-time would simply be given by time-of-flight of a photon through the network. Therefore, an estimate on the processing time can be given by considering the physical length of the components of this photonic integrated circuit NN and its effective waveguide index; for instance the weighted addition obtained through combiners and phase shifters has a physical length of about 100 μm, a passive waveguide synaptic summation has an on-chip coverage of around 200 μm, and an estimated NL activation module is less than 10 μm long. Thus, photonics allows a single neuron to be integrated well within 100's of μm in length, leading to ~ps computation time-scales [36]. This AONN thence would have a delay of about few ps, or 10^{12} MAC/s, and $<10^{17}$ MAC/J efficiency given, for example, the power levels of the Fano resonance discussed above. Such performance of the proposed AONN would potentially be several orders of magnitude more efficient and faster than GPU and electro-optical neural networks, provided noise is negligible (Fig. 6) [37].

Optical Materials EXPRESS

Table 1. Comparison of different neuromorphic technologies in terms of computing speed and power efficiency per MAC, i.e. per neuron. This speed is the time delay to 'active' or 'use' one neuron, and does not equal the system delay, which depends on the number of neurons and would make the comparison arbitrary. In van Neumann architectures the neuron delay is set by the clock cycle speed (100's MHz to few GHz). For electro-optic integrated photonic-based NN, this delay is the sum of the electro-optic components plus the waveguide propagation delay of the weights such as MZIs or tunable ring filters (10's-100 GHz). In an all-optical version al electronic delays are not-present and the delay depends only on the photons time of flight (~1ps) given by the neuron dimensions of hundreds of micrometers and an optical waveguide index = 3.

Technology	*Efficiency (MAC/J)*	*Speed = (1/Delay) Per neuron (MAC/s)*
NVIDIA GPU[36]	3×10^{6}	10^{9}
Electro-optical NN [36]	4×10^{12}	10^{10}
All-optical NN [this work]	4×10^{16}	10^{12}

3. Conclusion

In summary, we have investigated a MNPs/QD system, based on two metal nanoparticles sandwiching a QD, which showed a coherent nonlinear optical response. This phenomenon was due to interference between the dipoles of the plasmon oscillation in the metal nanoparticles and the exciton transition in the QD. Furthermore, we modeled integration of the assembly in a waveguide platform and optimized the modulation range of the NLs associated with the transmitted signal, reaching a fully NL optical modulation of the transmitted signal up to 3 dB. Moreover, we also studied the reverse saturable absorption mechanism in a film of C_{60}. The film displayed a clear NL optical response as function of the impinging power density, with a modulation range of approximately 7 dB. Moreover, the studied platforms can provide insights into the speed of a complete NN architecture based on integrated photonics and all optical activation functions. The proposed NL optical responses were used as activation functions for fully-connected neural networks, emulated in Tensor Flow. We tested these nonlinear activation functions on a standardized NN training set, MNIST classifiers of handwritten digits. Our results show that the accuracy of the ONN can match others commonly employed for up to 50 number of reconfigurations in both training and validation phase. From an architecture point of view, our all-optical NN has the potential to significantly outperform in terms of computing speed and energy efficiency the established architectures based on either electronics or electro-optics. We estimated an efficiency of $<10^{17}$ MAC/J and a speed of 10^{12} MAC/s for a fully optical NN, which is several order of magnitude higher than the electro-optic and FPGA counterparts. These new insights could contribute to the design and fabrication of optical NL modulators, which could pave the way for all-optical high-speed and efficient NNs. Future experimental work is needed to validate this potential.

4. Methods

Electromagnetic simulations

We use a commercial solver (FDTD Solutions from Lumerical Inc.) built on the finite-difference time-domain method, which solves the Maxwell equations on a discrete spatio-temporal grid, for all the simulations related to the MNPs/QD system to waveguide coupling. The MNPs/QD system is modeled as a 3-D box with the absorptance being swept for modeling the effect of the induced transparency as function of the input power. The adaptive mesh algorithm is used to refine the grid in the QD domain. The RSA model of C_{60} response is realized with a proprietary FDTD-ADE multiphysics code that brings in the carrier kinetics.

Optical Materials EXPRESS

Neural network emulation

We simulate the functionality of this proposed photonic network with Google's Tensorflow platform. Tensorflow is an open-source machine learning software system. At heart, Tensorflow is a computation graph executor. A typical Tensorflow process includes defining a computation graph, by defining the flow of computations similar to their symbolic mathematical description. Tensorflow comes with a rich set of built in computational operations. These operations vary from low-level mathematical operations such as matrix operations to higher-level operations such as finding gradients or nonlinear activation functions (AF), which is a centerpiece of modern neural networks (NN). In order to simulate the effect of our all-optical nonlinear AF on the overall functionality of the NN, we emulated the behavior of the nonlinear AFs using their transfer functions. These transfer functions describe input/output relationships that reflect the nonlinearity. While a generic nonlinear transfer function is unitless, depending on the encoding scheme, they can be regarded as power in/ power out. Once the nonlinear transfer functions are derived as depicted in Fig. 5(c), we replace the built-in nonlinear AFs in Tensorflow with the calculated transfer functions we modeled for our all-optical AFs. This allows us to simulate a full NN but with our all-optical nonlinearities. The simulation processes in the two common steps, namely the training and the inference. During the training, our NN learns a collection of weights that aims to minimize a loss; here we used "categorical cross-entropy" as our loss function, which is a measure of how well predicted results of a neural network match the ground truth for a classification task. Since our demonstrative task of this all-optical NN is the classifying of handwritten digit images of MNIST into the right class, once the network is fully trained, we evaluate the performance of the network in term of prediction accuracy by feeding the input with unseen images during training. As the prediction accuracy of the network with our all-optical AFs is comparable with that of commonly used in deep learning, we conclude that all-optical AFs are able to provide the same function as those successfully adopted in modern neural networks.

List of Abbreviations

AF (nonlinear) Activation Function of the perceptron
ADE Auxiliary Differential Equation(s)
AONN All-optical Neural Network
CMOS Complementary Metal Oxide Semiconductor
GPU Graphic Process Unit
FDTD Finite Difference Time Domain
FPGA Field Programmable Gate Array
MAC Multiply Accumulate
MNP Metal Nanoparticle
PIC Photonic Integrated Circuit
QD Quantum Dot
NL Nonlinear
NN Neural Network
LMI Light matter Interaction
PNN Photonic Neural Network
RSA Reverse Saturable Absorbtion

Acknowledgments

The authors acknowledge fruitful discussion with the team of Prof. Prucnal.

References

1. F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha,

Optical Materials EXPRESS

"TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. **34**(10), 1537–1557 (2015).
2. J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," Front. Neurosci. **7**, 118 (2013).
3. B. Marr, B. Degnan, P. Hasler, and D. Anderson, "Scaling energy per operation via an asynchronous pipeline," IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **21**(1), 147–151 (2013).
4. B. J. Shastri, A. N. Tait, T. F. de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, "Principles of Neuromorphic Photonics," arXiv:1801.00016 [physics] 1–37 (2018).
5. A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 7430 (2017).
6. M. A. Nahmias, B. J. Shastri, A. N. Tait, T. F. de Lima, and P. R. Prucnal, "Neuromorphic Photonics," Optics & Photonics News, OPN **29**(1), 34–41 (2018).
7. R. Amin, J. George, J. Khurgin, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, "Attojoule Modulators for Photonic Neuromorphic Computing," in Conference on Lasers and Electro-Optics (2018), Paper ATh1Q.4 (Optical Society of America, 2018), p. ATh1Q.4.
8. R. Amin, S. Khan, C. J. Lee, H. Dalir, and V. J. Sorger, "110 Attojoule-per-bit Efficient Graphene-based Plasmon Modulator on Silicon," in Conference on Lasers and Electro-Optics (2018), Paper SM1I.5 (Optical Society of America, 2018), p. SM1I.5.
9. J. George, R. Amin, A. Mehrabian, J. Khurgin, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, "Electrooptic Nonlinear Activation Functions for Vector Matrix Multiplications in Optical Neural Networks," in Advanced Photonics 2018 (BGPP, IPR, NP, NOMA, Sensors, Networks, SPPCom, SOF) (OSA, 2018), p. SpW4G.3.
10. J. George, A. Mehrabian, R. Amin, J. Meng, T. F. de Lima, A. N. Tait, B. J. Shastri, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, "Neuromorphic photonics with electro-absorption modulators," arXiv:1809.03545 [physics] (2018).
11. A. Dejonckheere, F. Duport, A. Smerieri, L. Fang, J.-L. Oudar, M. Haelterman, and S. Massar, "All-optical reservoir computer based on saturation of absorption," Opt. Express **22**(9), 10868–10881 (2014).
12. 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**(7), 441–446 (2017).
13. Q. Bao, H. Zhang, Z. Ni, Y. Wang, L. Polavarapu, Z. Shen, Q.-H. Xu, D. Tang, and K. P. Loh, "Monolayer graphene as a saturable absorber in a mode-locked laser," Nano Res. **4**(3), 297–307 (2011).
14. R. W. Schirmer and A. L. Gaeta, "Nonlinear mirror based on two-photon absorption," J. Opt. Soc. Am. B, J. Opt. Soc. Am. B **14**(11), 2865–2868 (1997).
15. M. Soljacić, M. Ibanescu, S. G. Johnson, Y. Fink, and J. D. Joannopoulos, "Optimal bistable switching in nonlinear photonic crystals," Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **66**(5), 055601 (2002).
16. F. D. Coarer, M. Sciamanna, A. Katumba, M. Freiberger, J. Dambre, P. Bienstman, and D. Rontani, "All-Optical Reservoir Computing on a Photonic Chip Using Silicon-Based Ring Resonators," IEEE J. Sel. Top. Quantum Electron. **24**(6), 1–8 (2018).
17. B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," Sci. Rep. **6**(1), 19126 (2016).
18. H. Leng, B. Szychowski, M.-C. Daniel, and M. Pelton, "Dramatic Modification of Coupled-Plasmon Resonances Following Exposure to Electron Beams," J. Phys. Chem. Lett. **8**(15), 3607–3612 (2017).
19. K. Santhosh, O. Bitton, L. Chuntonov, and G. Haran, "Vacuum Rabi splitting in a plasmonic cavity at the single quantum emitter limit," Nat. Commun. **7**, 11823 (2016).
20. A. Hatef, S. M. Sadeghi, and M. R. Singh, "Plasmonic electromagnetically induced transparency in metallic nanoparticle-quantum dot hybrid systems," Nanotechnology **23**(6), 065701 (2012).
21. X. Wu, S. K. Gray, and M. Pelton, "Quantum-dot-induced transparency in a nanoscale plasmonic resonator," Opt. Express **18**(23), 23633–23645 (2010).
22. M. Pelton and G. W. Bryant, *Introduction to Metal-Nanoparticle Plasmonics* (John Wiley & Sons, 2013).
23. R. A. Shah, N. F. Scherer, M. Pelton, and S. K. Gray, "Ultrafast reversal of a Fano resonance in a plasmon-exciton system," Phys. Rev. B Condens. Matter Mater. Phys. **88**(7), 075411 (2013).
24. P. B. Johnson and R. W. Christy, "Optical Constants of the Noble Metals," Phys. Rev. B **6**(12), 4370–4379 (1972).
25. M. P. Lisitsa, L. F. Gudymenko, V. N. Malinko, and S. F. Terekhova, "Dispersion of the Refractive Indices and Birefringence of CdS_xSe_{1-x} Single Crystals," Phys. Status Solidi, B Basic Res. **31**(1), 389–399 (1969).
26. J. A. Schuller, E. S. Barnard, W. Cai, Y. C. Jun, J. S. White, and M. L. Brongersma, "Plasmonics for extreme light concentration and manipulation," Nat. Mater. **9**(3), 193–204 (2010).
27. H. Leng, B. Szychowski, M.-C. Daniel, and M. Pelton, "Strong coupling and induced transparency at room temperature with single quantum dots and gap plasmons," Nat. Commun. **9**(1), 4012 (2018).
28. V. Yannopapas, E. Paspalakis, and N. V. Vitanov, "Plasmon-Induced Enhancement of Quantum Interference Near Metallic Nanostructures," Phys. Rev. Lett. **103**(6), 063602 (2009).
29. C. Li, L. Zhang, R. Wang, Y. Song, and Y. Wang, "Dynamics of reverse saturable absorption and all-optical switching in C60," J. Opt. Soc. Am. B **11**(8), 1356–1360 (1994).
30. W. Su, T. M. Cooper, and M. C. Brant, "Investigation of reverse-saturable absorption in brominated porphyrins," Chem. Mater. **10**(5), 1212–1213 (1998).

Optical Materials EXPRESS

31. J. W. Perry, K. Mansour, S. R. Marder, K. J. Perry, D. Alvarez, Jr., and I. Choong, "Enhanced reverse saturable absorption and optical limiting in heavy-atom-substituted phthalocyanines," Opt. Lett. **19**(9), 625–627 (1994).
32. Y. Gao, X. Zhang, Y. Li, H. Liu, Y. Wang, Q. Chang, W. Jiao, and Y. Song, "Saturable absorption and reverse saturable absorption in platinum nanoparticles," Opt. Commun. **251**(4-6), 429–433 (2005).
33. S. I. Azzam and A. V. Kildishev, "Full-wave analysis of reverse saturable absorption in time-domain," arXiv:1808.02436 [physics] (2018).
34. A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, "PCNNA: A Photonic Convolutional Neural Network Accelerator," arXiv:1807.08792 [cs, eess] (2018).
35. P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," arXiv:1710.05941 [cs] (2017).
36. B. J. Shastri, P. R. Prucnal, "Principles of Neuromorphic Photonics" *Encyclopedia of Complexity and Systems Science* (Springer, 2017).
37. M. A. Nahmias, B. J. Shastri, A. N. Tait, T. Ferreira de Lima, and P. R. Prucnal, "Neuromorphic Photonics," Opt. Photonics News **29**(1), 34 (2018).