Polaritonic Neuromorphic Computing Outperforms Linear Classifiers

Dario Ballarini,* Antonio Gianfrate, Riccardo Panico, Andrzej Opala, Sanjib Ghosh, Lorenzo Dominici, Vincenzo Ardizzone, Milena De Giorgi, Giovanni Lerario, Giuseppe Gigli, Timothy C. H. Liew, Michal Matuszewski, and Daniele Sanvitto

ABSTRACT: Machine learning software applications are ubiquitous in many fields of science and society for their outstanding capability to solve computationally vast problems like the recognition of patterns and regularities in big data sets. In spite of these impressive achievements, such processors are still based on the so-called von Neumann architecture, which is a bottleneck for faster and power-efficient neuromorphic computation. Therefore, one of the main goals of research is to conceive physical realizations of artificial neural networks capable of performing fully parallel and ultrafast operations. Here we show that lattices of exciton-polariton condensates accomplish neuromorphic computing with outstanding accuracy thanks to their high optical nonlinearity. We demonstrate that our neural network significantly increases the recognition efficiency compared with the linear classification algorithms on one of the most widely used benchmarks, the MNIST problem, showing a concrete advantage from the integration of optical systems in neural network architectures.

KEYWORDS: Exciton-polaritons, optical microcavities, neuromorphic computing, reservoir computing, semiconductors

INTRODUCTION

It is well-established that artificial atoms in the form of excitons in semiconductors can hybridize with photons to form exciton-polaritons when placed inside a cavity. Polaritons demonstrate many properties analogous to those of cold-atom systems, such as Bose–Einstein condensation, together with properties analogous to those of nonlinear optical systems, such as four-wave mixing and soliton formation. Thanks to their hybrid nature, polaritons are interesting both for fundamental studies and for their possible integration into optical devices. Indeed, photonic systems have been advocated for use in information processing for a long time because of their potentially superior speed and energy efficiency with respect to their electronic counterparts. Polaritons are suitable candidates for this purpose because of their high optical nonlinearity and femtosecond reaction time. Several experimental milestones using polaritons to reproduce binary logic elements have been completed, including the construction of individual logic gates, switches, routers, and transistors, although polariton systems have not been sufficiently developed to perform any practical information-processing task.

Rather than using polaritons to reproduce binary logic systems, we demonstrate here the use of polaritons in an alternative neuromorphic architecture and show how the nonlinearity and complexity in a polariton lattice are able to distinguish patterns very efficiently. Polariton lattices were previously shown to exhibit a range of intriguing highly nonlinear effects, including transitions between ground and excited states, gap solitons, and spin correlations. More recently, lattices of localized polariton condensates have been used to form topological protected states and to solve many-body Hamiltonians. Aside from providing an example of the application of nonlinear polariton lattices, we find that their complexity remarkably generates pattern recognition success rates higher than originally expected and highly competitive with previously considered hardware implementations.

Biologically inspired artificial neural networks (NNs) are typically defined by a set of neurons, which are the individual processing elements, and their interconnections. Each neuron receives a combined signal from many surrounding neurons, reacts to that signal nonlinearly, and passes its own signal to subsequent neurons in the network. The interconnections transmit signals with different weights, and it is the choice of these weights that is critical to determining the network performance.
function. If such weights can be engineered precisely enough, one obtains powerful information processing systems for the recognition and classification of patterns even with imperfect, noisy, or incomplete information. NN architectures are radically different from existing (von Neumann) computational architectures, as there is no physical separation of memory and processing units. Thus, NN architectures avoid a key bottleneck in the data transfer rate, the von Neumann bottleneck, which limits the speed of modern computers. NNs are also highly efficient in parallelizing tasks, as all of the neurons can operate simultaneously. Aside from implementations using conventional binary logic to simulate NNs, hardware implementations have been based on memristors and spintronic and optical systems. However, the individual control of the connection weights remains a key challenge in the development of scalable hardware implementations.

A promising solution appeared in the form of reservoir computing (RC), which throws out much of the paradigm that weights inside the network need precise control and allows all weights inside the bulk of the network to be random. Instead, only a single layer of weights that are applied to an output set of neurons is controlled. Remarkably, this scheme still results in a working NN, and because only a small part of the whole system needs to be controlled, this design is much more accessible, with photonic systems, microwaves, and optoelectronic implementations of RC.

The particular choice of the random matrix has a limited effect. If such weights can be engineered precisely enough, one obtains powerful information processing systems for the recognition and classification of patterns even with imperfect, noisy, or incomplete information. NN architectures are radically different from existing (von Neumann) computational architectures, as there is no physical separation of memory and processing units. Thus, NN architectures avoid a key bottleneck in the data transfer rate, the von Neumann bottleneck, which limits the speed of modern computers. NNs are also highly efficient in parallelizing tasks, as all of the neurons can operate simultaneously. Aside from implementations using conventional binary logic to simulate NNs, hardware implementations have been based on memristors and spintronic and optical systems. However, the individual control of the connection weights remains a key challenge in the development of scalable hardware implementations.

A promising solution appeared in the form of reservoir computing (RC), which throws out much of the paradigm that weights inside the network need precise control and allows all weights inside the bulk of the network to be random. Instead, only a single layer of weights that are applied to an output set of neurons is controlled. Remarkably, this scheme still results in a working NN, and because only a small part of the whole system needs to be controlled, this design is much more accessible, with photonic systems, microwaves, and optoelectronic implementations of RC.

Results and Discussion

The schematic of the experimental procedure is shown in Figure 1. To distribute the information over the whole reservoir, each digit is converted into a matrix with the same number of elements as the number of nodes in the polariton network. This is done by multiplying the input digit (Figure 1a) by a random matrix, which is taken the same for all digits. The particular choice of the random matrix has a limited effect on the final recognition efficiency (see the Supporting Information for details on the data set preparation).

Each node in the network is optically created by using a spatial light modulator (SLM) to shape the pump laser beam (wavelength $\lambda = 836$ nm) into an array of spots with controllable intensity and phase. The laser light is injected into a semiconductor microcavity, in which exciton-polaritons are created by tuning the laser frequency to the polariton resonance (see the SI). The information on the digit is encoded in the intensities of the spots (Figure 1b), which excite different realizations of the polariton network, while junctions between neurons are created by imposing a phase difference of $\Delta \theta = \pi$ between adjacent nodes. The nonlinearity is provided by the Kerr-type interactions between polaritons: for resonant excitation, the curve of the total nonlinearity is provided by the Kerr-type interactions between exciton-polaritons.

The resulting data set is sent to the SLM to pattern the laser beam. The nonlinear transmission (c) of the polariton RC produces an output (d), which is recorded by a CCD camera and multiplied by the weight matrix (e) to obtain the digit classification (f). (g) Scheme of the experimental setup corresponding to the steps described in (b–d).
weights in the output layer (Figure 1e) that minimize the error rate of predictions. We implement logistic regression, an efficient algorithm for linear classification, on a computer to adjust the output weights in the readout matrix. In the testing phase, the trained network is used to predict the recognition rate on examples that the system has not seen before. It should be noted that at this stage the computer could be replaced by an all-optical device performing a linear vector–matrix multiplication operation.46 We always compare our results against the pure logistic regression on input data to evaluate the gain of the recognition rate due to the nonlinear transformation performed by the network. To this scope, the transfer function of the SLM is calibrated to be strictly linear (see the SI) in order to measure solely the contribution of the polariton nonlinearity to the image recognition efficiency. This is important because although the final efficiency can be increased by adding a nonlinear transfer function to the SLM, the polariton array can work much faster than the electronically controlled SLM, allowing operation rates up to the terahertz range. Moreover, while the SLM is used here as a convenient way to generate an array of polariton nodes, polariton lattices with a few thousand nodes can be realized through lithographic patterning for on-chip integration.47

In Figure 2, we test the recognition efficiency of the polariton RC on the MNIST data set. To avoid any spurious effect, such as the possibility that a continuous drift in the temperature would artificially increment the recognition efficiency, the input frames in the data set are randomly shuffled, confirming the proper operation of the polariton RC in the classification task (see the SI). Given the limited reservoir size considered here, we reduce the resolution of the MNIST digits from 28 × 28 pixels to 7 × 7 pixels (top row “M” in Figure 2 left). Although this operation loses some information, the recognition rate of the polariton RC remains higher than that set by logistic regression on the full-resolution MNIST, which is illustrated by the shaded gray region in Figure 2 right for different training set sizes. The exact recognition rate depends on which examples in the MNIST data set are taken for training and testing. Thus, the different data points have error bars, calculated from statistical analysis of different random choices of the training and testing sets. More training samples result in better performance, but it is remarkable that the polariton RC performs well even with a limited number of training samples. One can also compare to the recognition rate obtained from performing logistic regression on the 7 × 7 input digit set (point M in Figure 2 right), performing logistic regression on the 8 × 8 SLM pattern (point S in Figure 2 right), and performing logistic regression on the laser beam image after the SLM (point L in Figure 2 right).

These numbers indicate that nonlinear polariton lattices are particularly efficient in performing image recognition. Even though a direct comparison between different physical architectures is difficult, our system is able to obtain higher accuracy than previously reported RC approaches while working with lower-resolution images. In Table 1, different strategies are compared, including the realization of terahertz deep NN with metallic masks50 (81% MNIST accuracy), a convolutional approach in which the hopping conduction in a silicon contact is controlled by artificial evolution58 (96% MNIST accuracy), and a cellular automata approach to RC (ReCA) implemented on an FPGA59 (98% MNIST accuracy). While state-of-the-art software realizations of NNs achieve accuracies on the MNIST data set higher than 99.8%, the potentiality of photonic NN can be most appreciated in reducing the energy consumption or accelerating specific operations in complex problems of image classification.51–53

The picosecond dynamics and high accuracy of polariton lattices therefore seem to be suited for image recognition when fast operational speed, noisy information, or offline devices are involved. Moreover, the polariton RC works extremely well also when the resolution of the input image is reduced to only 4 × 4 pixels. Logistic regression performed on the nominal 4 × 4 input data set gives a recognition rate of 81.6%, while the signal processed by the polariton array allows an accuracy of 86.3% to be reached, with an increment of 4.7% (Figure S3). The accuracy can be further increased by employing a reservoir
with a larger number of nodes (Figure S4) or, equivalently, by increasing the readout resolution (Figure S5).

In the following, the nonlinear output of the polariton node and the connectivity of the neural network are analyzed in more detail. In Figure 3a, the output of a single node of the polariton RC is shown (only the node indicated by the red dot in the inset of Figure 3a is activated). The input laser is slightly blue-detuned with respect to the polariton resonance, resulting in low transmission at low power. For increasing power, polariton–polariton interactions blue-shift the polariton resonance up to the laser frequency, resulting in a nonlinear increase of the transmission through the cavity.14 If the input power is further increased, the transmitted intensity saturates to an output value, which depends on the energy detuning ΔE between the laser and the polariton resonance. The laser detuning (with respect to the polariton resonance at low power) and the range of input powers are optimized by maximizing the recognition rate of the whole polariton RC, as shown in Table S1. In Figure 3b, the whole polariton RC is activated, and the output intensities of single neurons are compared. While all of the nodes show a nonlinear transmission curve in the same range of powers, a specific behavior is associated with each node. From a comparison of Figure 3a and Figure 3b, it is evident that the behavior of polariton nodes is modified by the activation of other neurons in the network, with a redistribution of intensities between nodes.

This shows the importance of using hybrid light–matter particles to simultaneously obtain high nonlinearity and efficient connectivity between nodes. Indeed, the nonlinear behavior of polariton nodes can be simulated by a nonlinear calibration of the SLM transfer function and redirection of its output to the camera. However, even though the success rate is slightly increased over the linear classifier, in this case the result is still lower than that obtained by processing the information in the polariton RC. This is due to the lack of efficient connections between nodes in the SLM array, which are instead present in the polariton network. In Figure 4, we show that hopping of polaritons between nodes provides the required connectivity. In Figure 4b, all of the nodes except the central one, marked as A, are pumped using an almost homogeneous SLM pattern. The ReCA system in ref 49 was trained with an extended MNIST data set including distortions, which improved the recognition rate.

Table 1. Comparison of Different Hardware NN Performances on MNIST

<table>
<thead>
<tr>
<th>System</th>
<th>MNIST</th>
<th>Training</th>
<th>Testing</th>
<th>Resolution</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>polariton RC</td>
<td>93%</td>
<td>4000</td>
<td>1000</td>
<td>7 × 7 THz</td>
<td></td>
</tr>
<tr>
<td>memristor RC</td>
<td>88.1%</td>
<td>14000</td>
<td>2000</td>
<td>28 × 28 kHz</td>
<td></td>
</tr>
<tr>
<td>THz deep NN</td>
<td>81%</td>
<td>1500</td>
<td>50</td>
<td>28 × 28 THz</td>
<td></td>
</tr>
<tr>
<td>hopping current</td>
<td>96%</td>
<td>60000</td>
<td>10000</td>
<td>28 × 28 MHz</td>
<td></td>
</tr>
<tr>
<td>FPGA ReCA</td>
<td>98%</td>
<td>60000</td>
<td>10000</td>
<td>22 × 20 MHz</td>
<td></td>
</tr>
</tbody>
</table>

“The sizes of the testing and training sets as well as the resolution of the input digits are reported together with the success rate. The speed is an estimated maximum operational frequency for digit classification based only on the fundamental physical limits of the system.13 The hopping current experiment in ref 48 is not a full hardware implementation but rather a software simulation of thousands of identical dopant systems. The ReCA system in ref 49 was trained with an extended MNIST data set including distortions, which improved the recognition rate.

Figure 3. Interconnected nonlinear behavior of polariton RC. (a) Output of a single node of the lattice vs input power (only the node indicated by the red dot in the inset is activated). (b) Outputs of several nodes in the polariton lattice as functions of the input power. When multiple nodes are activated, the behavior of each node is influenced by the connection between neighboring nodes.

Figure 4. Connectivity between nodes. (a) Comparison of the output intensities of node A activated by the contributions from the first and second neighbor nodes A’ and A” (blue line) and from second neighbor nodes A only (red line). (b) Region of the network with the pump power at node A intentionally set to P = 0. (c) Same as in (b) but with the power set to P = 0 at the first neighbor nodes A’ as well. Node A is activated in this case also, with a transmitted intensity of about 1/3, with respect to the configuration shown in (b).
limited to first neighbors but extend farther in the polariton network. In Figure 4c, the first neighbor nodes A' are also switched off, and node A can be activated only by the second neighbor nodes (indicated as A'' in Figure 4). The red line in Figure 4a shows that nodes A'' contribute to A with a connection strength of about one-half compared with nodes A'.

Moreover, the redistribution dynamics within the network takes place on a time scale comparable to the polariton lifetime. As shown by simulations (see the SI), the steady state is reached in this case after about 50 ps. Even though the present experiments use a continuous-wave pump, we anticipate that the recognition rate can be further improved by following the transient dynamics of the system (see the section on time multiplexing in the SI). Finally, we note that in order to take advantage of the ultrafast polariton dynamics, the SLM and CCD used in our proof-of-principle experiment can be replaced by faster technologies. However, these elements are not essential components of the polariton RC. Indeed, the polariton RC can work as a photonic accelerator that directly processes an optical signal without any optical-to-electronic conversion. In this case, the RC output is then multiplied all-optically by the classification matrix and instantaneously sent to ultrafast photodectors. Such a system does not make use of an SLM or CCD and can fully exploit the terahertz operation rate of the polariton neural network.

In conclusion, neuromorphic computing in a physical network of multiple connected nonlinear nodes is demonstrated. We show that exciton-polaritons are a suited system thanks to their mixed light–matter components: polariton–polariton interactions bring the desired nonlinearities, while the photonic component assures the connectivity between nodes and high operational speeds. We measured the recognition efficiency on the MNIST data set and found a significant increase in the accuracy with respect to linear classifiers and previous examples of photonic NNs. We note that a further increase in the recognition rate can be obtained in polariton networks by exploiting the spin degree of freedom and the phase difference between nodes. 4,5 With larger polariton arrays, success rates comparable to software implementations of reservoir models can be achieved with present technologies, moving toward a realistic integration of applied artificial intelligence based on optical systems.

ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.nanolett.0c00435.

Data set and input preparation; formation and excitation of the polariton RC; linear SLM calibration; logistic regression; combination of data from series of measurements; collecting data from emission; working conditions; numerical simulations; time multiplexing (PDF)

AUTHOR INFORMATION

Corresponding Author
Dario Ballarini — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy; orcid.org/0000-0002-2453-5849; Email: dario.ballarini@cnr.nanotec.it

Authors
Antonio Gianfrate — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Riccardo Panico — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Andrzej Opala — Institute of Physics, Polish Academy of Sciences, PL-02-668 Warsaw, Poland
Sanjib Ghosh — School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371; orcid.org/0000-0002-5014-9466
Lorenzo Dominici — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Vincenzo Ardizzone — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Milena De Gorgi — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Giovanni Lerario — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Giuseppe Gigi — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy
Timothy C. H. Liew — School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371
Michał Matuszewski — Institute of Physics, Polish Academy of Sciences, PL-02-668 Warsaw, Poland
Daniele Sanvitto — CNR NANOtec—Institute of Nanotechnology, 73100 Lecce, Italy

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.nanolett.0c00435

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partially supported by the FISR—C.N.R. Project “Tecnomed—Tecnopolo di nanotecnologia e fotonica per la medicina di precisione” and by ERC “ElecOpter” Grant 780757. S.G. and T.C.H.L. were supported by the Ministry of Education (Singapore), Grants MOE2017-T2-1-001 and MOE2018-T2-02-068. A.O. acknowledges support from the National Science Center, Poland, via Grant 2016/22/E/ST3/00045. M.M. acknowledges support from the National Science Center, Poland, via Grant 2017/25/Z/ST3/03032 under the QuantERA program. We thank R. Houdré for growing the sample.

REFERENCES


