

Energy-Efficient Neural Network Inference with Microcavity Exciton Polaritons

M. Matuszewski^{1,*}, A. Opala,¹ R. Mirek², M. Furman,² M. Król², K. Tyszka², T.C.H. Liew,³ D. Ballarini,⁴ D. Sanvitto,^{4,5} J. Szczytko,² and B. Piętka²


¹*Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, Warsaw PL-02-668, Poland*

²*Institute of Experimental Physics, Faculty of Physics, University of Warsaw, ul. Pasteura 5, Warsaw PL-02-093, Poland*

³*Division of Physics and Applied Physics, Nanyang Technological University 637371, Singapore*

⁴*CNR NANOTEC-Institute of Nanotechnology, Via Monteroni, Lecce 73100, Italy*

⁵*INFN, Sez. Lecce, Lecce 73100, Italy*

 (Received 31 May 2021; revised 21 July 2021; accepted 29 July 2021; published 25 August 2021)

We propose all-optical neural networks characterized by very high energy efficiency and performance density of inference. We argue that the use of microcavity exciton polaritons allows one to take advantage of the properties of both photons and electrons in a seamless manner. This results in strong optical nonlinearity without the use of optoelectronic conversion. We propose a design of a realistic neural network and estimate energy cost to be at the level of attojoules per bit, also when including the optoelectronic conversion at the input and output of the network, several orders of magnitude below state-of-the-art hardware implementations. We propose two kinds of nonlinear binarized nodes based either on optical phase shifts and interferometry or on polariton spin rotations.

DOI: [10.1103/PhysRevApplied.16.024045](https://doi.org/10.1103/PhysRevApplied.16.024045)

I. INTRODUCTION

The progress in communications, information processing, and mobile technologies resulted in the advent of the era of big data. This received an incredible burst with the developments in parallel computing hardware, and in particular in artificial intelligence and neural networks (NNs) [1,2]. Practical applications of machine learning quickly became an important part of the economy, and currently include, among others: natural language processing, image and sound recognition, autonomous vehicles, finance, marketing, and research. At the same time, these developments put a strain on computing systems, which have to process data faster and more efficiently than ever before. This becomes a serious issue as the energy consumption of information processing and communication systems is set to surge. It is already a significant part of the global energy consumption, and is expected to reach over 20% of global electricity use by 2030 [3], becoming one of the main bottlenecks of further progress. This problem has been recognized by the machine-learning community [4–6].

Meanwhile, the increase of performance of CMOS saturates, significantly deviating from Moore's law [7]. In recent years, the shift to architectures with many parallel computing units, such as graphical processing units (GPUs) or tensor processing units (TPUs), has been the

dominant trend, but this avenue is limited by Amdahl's law [8]. In result, much effort has been dedicated to research on possible alternatives to the CMOS technology for information processing [9]. In particular, there have been great advancements in machine learning with photons, in both all-optical and optoelectronic systems [10–12]. Photonic information processing benefits from high speed, parallelization, low communication losses, and high bandwidth. Fully functional photonic neurons including nonlinear activation functions were demonstrated [11–14], including spiking neurons [15,16], as well as neural networks [17–28]. Certain networks achieved high performance in challenging machine-learning tasks, such as image and video recognition [29–35]. Scalable vector-matrix multiplication operations and convolutions, which are at the core of neural network implementations, were demonstrated using photons [36–39]. All these remarkable developments indicate that photonic information processing is maturing, and may become a serious competitor for electronics in the near future, as recognized both in the academia and the industry [40,41].

The most serious issue taming progress in electronic systems is related to energy consumption and heat generation, which results in the phenomenon of “dark silicon” [7,42]. The main source of losses at high data rates is not the cost of the actual logic operations, but communication using electronic channels [43]. On the other hand, communication using photons can be almost lossless. For this

*mmatu@ifpan.edu.pl

reason, electric wiring has been successively replaced by photonic waveguides at long and medium distances, and are now being commercially implemented even at the chip scale [44]. However, electronics is still necessary for signal amplification and processing. The drawback of photonics is the weakness of nonlinearity, or effective interactions between photons. Nonlinearity is crucial for the implementation of nontrivial information processing, either in the form of a transistor or a neuron activation function. All-optical information processing has been impractical due to the high intensity of the optical beams required, which effectively results in high energy cost per operation.

Alternatively, since nonlinearity occurs at very low energy levels in electronics, optoelectronic conversion can be used, so that nonlinear transformations are implemented electronically, and photons are used for communication or linear operations only. However, there are still serious issues with the implementation of scalable optoelectronic information processing. Integration of light sources is difficult, and the spatial scales for efficient electronics (nanometers) and photonics (micrometers) are incompatible. The energy cost of optoelectronic conversion in logical elements is typically such that it overcomes any gain from the use of optics for communication. It is estimated that net energy-efficiency benefits can only be expected if the spatial extent of photonic modes is reduced to the nanometer scale [43]. This may be difficult in practice without incurring additional losses, either due to absorption or imperfect confinement of photons at the subwavelength scale. Since practical solutions to these issues are yet to be demonstrated, it is desirable to find an all-optical alternative for energy-efficient information processing. Such an approach requires strong optical nonlinearity without optoelectronic conversion. This would allow for fully exploiting the intrinsic ultrashort time scales and high energy efficiency of photonics in complex information-processing tasks that require nonlinearity.

We recently demonstrated hardware neuromorphic systems where strong nonlinearity resulted solely from interactions of exciton polaritons, quantum superpositions of light and matter [30,31,45]. Such superpositions, in the form of mixed quasiparticles of photons and excitons [46,47], are characterized by excellent photon-mediated transport properties and strong exciton-mediated interactions [48,49]. Their unique properties led to the discovery of remarkable phenomena, including nonequilibrium Bose-Einstein condensation [47,50], superfluidlike states [51,52], interactions of quantum defects [53–55], lasing of topological edge states [56,57], as well as demonstration of polariton transistors and switches [49,58–62], gates [31,63], neurons [64,65], simulators [66], and nonlinear phenomena at the femtojoule level [48,49]. In our proof-of-principle neural network implementations, both photonic and electronic layers are used, but the latter performed linear operations only. These experiments, based

on reservoir computing [30] and binarized neural networks [31], achieved accuracy of 93% and 96% in the Modified National Institute of Standards and Technology (MNIST) handwritten digit benchmark, respectively. Moreover, the strong polariton nonlinearity, resulting from Bose-Einstein condensation, allowed the achievement of energy efficiency of 16 picojoule per synaptic operation (SOP) in an all-optical binarized neuron [31].

Here, we argue that semiconductor microcavity systems can be used to construct fully functional, all-optical neural networks characterized by extremely high energy efficiency of inference. We show why using polaritonics in place of standard nonlinear optical phenomena, is the key to achieving such a performance. Exciton polaritons can perform information processing that takes advantage of the properties of both photons and electrons in a seamless manner. We propose a design of a polariton-based neural network in which fast, parallel, and low-loss communication using photons is accompanied by the strong nonlinearity induced by exciton interactions, without the need for optoelectronic conversion. We estimate the performance of such a network in the case of resonant excitation, and predict that nonlinear inference at attojoules per bit energy cost can be achieved, outperforming current technologies by orders of magnitude. We propose how a simple network, based on nonlinear binarized nodes, could be implemented, taking advantage of either the interference of photons or polariton spin rotations. We estimate two key measures of the proposed design—the energy efficiency and performance density. Our estimations indicate that using currently available optical elements, the network could reach an energy efficiency of 4×10^{16} SOP s⁻¹ W⁻¹ (synaptic operations per second per watt) and performance density of 10^{16} SOP s⁻¹ mm⁻² (synaptic operations per second per millimeter squared), which are higher than the limits of the current semiconductor technology [67], as well as other technologies under investigation [68], by 4 and 3 orders of magnitude, respectively. Finally, we discuss the challenges and limitations of the proposed platform, including those related to optical losses, and indicate the possible applications where polaritonics may compete with CMOS electronics.

II. GENERAL ESTIMATES

A. Exciton polaritons

To estimate the theoretical minimum of energy required to perform a single nonlinear operation using exciton polaritons, we consider the simple device depicted in Fig. 1. The optical phase of the output pulse depends on the input pulse intensity due to nonlinear dynamics of polaritons inside the microcavity. We assume that the nonlinear process is third order, i.e., the phase shift depends linearly on the intensity, and results from the repulsive interactions between polaritons mediated by their exciton component

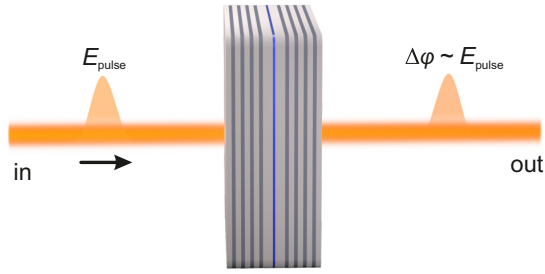


FIG. 1. Simple setup for the consideration of energy efficiency. The microcavity contains exciton-polariton quantum wells or a nonlinear optical medium. The optical phase shift of the output pulse depends on the energy of the input pulse. This nonlinear (i.e., optical-intensity-dependent) transformation is useful for information processing when phase shifts of the order of π can be achieved.

[47]. While other processes, including interactions with the reservoir of weakly coupled excitons, or the reduction of light-matter coupling strength, can also lead to a similar nonlinear response, we consider direct polariton-polariton interactions as the process that is likely to play the most relevant role at the shortest timescales.

We assume that to implement a useful nonlinear transformation, the phase shift induced by the nonlinearity should be of the order of π . Such a phase shift can be, for example, subsequently converted into full-scale amplitude modulation using a Mach-Zender interferometer. Therefore, it is enough to construct optical components that turn on or off transmission through the device depending on the nonlinear phase shift. However, we do not exclude the possibility that smaller phase shifts can also be used for computation. We compare the natural energy scales to estimate the number of photons necessary to realize such a phase shift. This leads to the formula $ng > \hbar\gamma$ where n is the quantum-well exciton density, g is the exciton-exciton interaction coefficient, and γ is the polariton decay rate, which is the inverse of the lifetime of polaritons in the cavity. Here $n \approx \chi N_{\text{photons}} / (N_{\text{QW}} S)$, where N_{photons} is the number of photons that are converted into polaritons, χ is the exciton Hopfield coefficient, S is the area of the illuminated surface, and N_{QW} is the number of quantum wells accommodating excitons. To estimate a realistic lower limit of energy consumption, we consider resonant excitation configuration, in which polaritons are created directly by a picosecond laser pulse. Such a configuration is much more energy efficient than the nonresonant excitation used in the previous experiment [31], since most of the photons can be converted directly into polaritons. As we are concerned with order-of-magnitude estimations, for the time being we neglect the photons reflected or absorbed by the microcavity or other optical components. More precise estimates will be presented in Sec. III C.

In the theoretical estimation of energy efficiency we use realistic microcavity parameters. Following Refs. [69,70] we use the values of parameters $g = 2 \mu\text{eV } \mu\text{m}^2$, $\gamma = (270 \text{ ps})^{-1}$. These parameters correspond to state-of-the-art GaAs microcavities, but a range of inorganic materials including CdTe, ZnO, GaN, and perovskites are expected [71] to be characterized by exciton interaction coefficients g in the range $2 - 5 \mu\text{eV } \mu\text{m}^2$. Many of them were used to observe exciton polaritons at room temperature [72–76]. In this work we assume a room-temperature operation to avoid the energy cost of cooling. We assume that a single quantum well is present in the microcavity, and use a conservative estimation of the surface area of a single polariton gate $S = 10 \mu\text{m}^2$. Note that polariton modes as small as $1 \mu\text{m}^2$ have been realized experimentally [77], but structuring light on such a small length scale could be challenging due to interference effects.

This gives the estimate of the minimum energy of a single input pulse $E_{\text{pulse}} = N_{\text{photons}} E_{\text{photon}}$ at the level of 3 attojoules ($3 \times 10^{-18} \text{ J}$), which corresponds roughly to 12 photons per pulse, for the photon energy $E_{\text{photon}} = 1.6 \text{ eV}$. One may worry that too small a photon number introduces errors due to quantum fluctuations; these are discussed in Sec. VII. Since a simple binary neuron (see Sec. IV) requires four pulses and performs two synaptic operations, the energy-efficiency bound is estimated as $1.7 \times 10^{17} \text{ SOP s}^{-1} \text{ W}^{-1}$. This can be compared, for example, with the energy efficiency of $4 \times 10^{10} \text{ SOP s}^{-1} \text{ W}^{-1}$ obtained with the IBM TrueNorth neuromorphic chip [78].

B. Nonlinear optical media

Microcavities can be used to construct nonlinear elements without the use of exciton polaritons. Instead, any medium exhibiting third-order nonlinearity can be embedded in a microcavity to realize the same concept. In this case, the energy required for a single operation can be estimated from the strength of optical phase modulation, knowing the parameters n_2 (nonlinear Kerr index), n_0 (refractive index), and n_g (group refractive index). The third-order nonlinearity results in the intensity-dependent refractive index change $\delta n = n_2 I = n_2 P / S$, where P is the power and S is the surface area. The resulting change of resonator frequency is $\delta\omega = -(\delta n / n)\omega$ as the wavelength is fixed by the cavity size and the number of antinodes. The corresponding shift of resonator energy is $\delta E = \hbar\delta\omega$. The minimum energy shift required for the operation of a gate can be estimated as in the case of polaritons, by comparing it to the natural energy scale of the cavity $|\delta E| = \hbar\gamma$. If this condition is fulfilled, photons can acquire a π phase when interacting in the microcavity.

Finally, the energy of the input pulse can be estimated from $E_{\text{pulse}} = P t_{\text{cav}}$, where t_{cav} is the time necessary for the light to perform a round trip inside the cavity, $t_{\text{cav}} = 2L n_g / c$, where L is the length of the cavity. This leads

to the formula $E_{\text{pulse}} = 2n_0n_g\gamma SL/(n_2\omega c)$ for the energy of the input pulse. We estimate that materials exhibiting strong and fast nonlinearity, such as silicon or gallium arsenide semiconductors in the weak coupling regime, with n_2 of the order of $10^{-17} \text{ m}^2 \text{ W}^{-1}$, would require at least 10^{-14} J (10 fJ) per synaptic operation, 4 orders of magnitude higher than in the case of exciton polaritons. At the same time, materials that exhibit slow nonlinear processes, such as photorefractive media, would not be energy efficient due to long illumination times required to build up the nonlinear index change.

C. Optoelectronic approach

Electronic devices are excellent for nonlinear operations, with energy per bit in a logical gate reaching possibly the level of 50–100 aJ [43,79]. However, communication on chip requires charging electrical wires to the level of 1 V, because lower voltage leads to strong leakage current in transistors. The energy cost of charging communication lines results in large energy dissipation, of the order of pJ per bit for floating point operations, memory access, and even larger cost for off-chip communication.

On the other hand, optoelectronic approach, where nonlinear operations are performed electronically, but communication is realized with light, is a possible solution that takes advantages of both methods. However, this requires optoelectronic conversion for each information bit, which itself generates pJ energy cost in typical devices [44]. Since the cost of optoelectronic conversion scales with size, it was suggested that net benefits can be expected if light modes are confined to volumes at the nanometer scale [43]. This may be difficult in practice, as the confinement of light modes to below-wavelength spatial scales is usually associated with strong losses either due to absorption by metallic mirrors, or imperfect confinement of dielectric mirrors.

III. ALL-OPTICAL NEURAL NETWORK

In this section we discuss a possible implementation of polariton nonlinearity in an all-optical neural network. A single hidden-layer network of binarized nodes is considered for its simplicity. Here and in the following, we use the term “binarized” to describe a neural network with binary inputs and activations in the hidden layer. While numerous optical neural networks or machine-learning systems have been described in the literature, only Refs. [30,31] considered the use of exciton polaritons.

A. Design

We propose a design of a simple all-optical neural network taking advantage of the polariton nonlinearity. There are no electrical or optoelectronic elements in the system and both inputs and outputs are assumed to take the

form of optical pulses. This makes the proposal distinct from previous experimental implementations [30,31], and is crucial for the energy efficiency of the network. Moreover, all elements of the network such as beam splitters, mirrors, lenses, light filters are passive and do not require an external power supply. In contrast to experiments realized in exciton-polariton systems [30,31], our design does not involve optoelectronic conversion or electronic information processing at any element within the network. Although we consider inputs and outputs as optical, we appreciate that they may have to be converted from or to electronic signals for compatibility with other systems. We account for the corresponding additional energy consumption in Sec. III C. The proposed network is one of the simplest possible applications of polaritons for optical information processing, but not the only one possible, and many other configurations can be envisaged. We discuss some of the possible extensions in Sec. III F. The aim of this work is to propose a simple design to indicate the potential of polaritonics for energy-efficient computing.

We consider a large network of all-optical binarized neurons where inputs and activations in the hidden layer are two level. Such a network can be used for complex classification tasks [31,80,81]. The logical structure of the network is analogous to the one introduced in Ref. [31] and shown schematically in Fig. 2. In Ref. [31], we experimentally realized a single binarized neuron using a polariton microcavity excited with a nonresonant laser pulse, achieving 16 pJ per bit energy efficiency. Here, we propose to use resonant excitation to achieve the same functionality, but at much higher energy efficiency, since resonant pumping selectively excites polaritons and allows most of the input photons to be converted to polaritons. The activation function of neurons is such that they effectively perform XOR logical gate operations. In the case of information encoding with light intensity, the intensity levels corresponding to “0” and “1” bits can be defined arbitrarily and differently at the input and output of the gates, as long as they are consistent between the gates. We consider a single hidden-layer network, composed of N_{nodes} polaritonic gates. The linear classification at the output layer is realized by an all-optical vector-matrix multiplier. Optical vector-matrix multiplication has been realized in many experiments [17,19,20,27,29,38,82], and recently multiplication of 3000 variables has been demonstrated [83].

The input information is encoded in space, and all input pulses are assumed to arrive at the same time. In the particular example considered, pixels from the MNIST dataset are binarized, and assigned to elements of two equally sized arrays. We assign random pairs of pixels to the elements of the two arrays, as depicted in Fig. 2. The same pairs of pixel positions denoted by $p_1 \dots p_n$ are assigned to the same array elements $1 \dots n$ for all digits. This allows the network to detect nontrivial correlations between

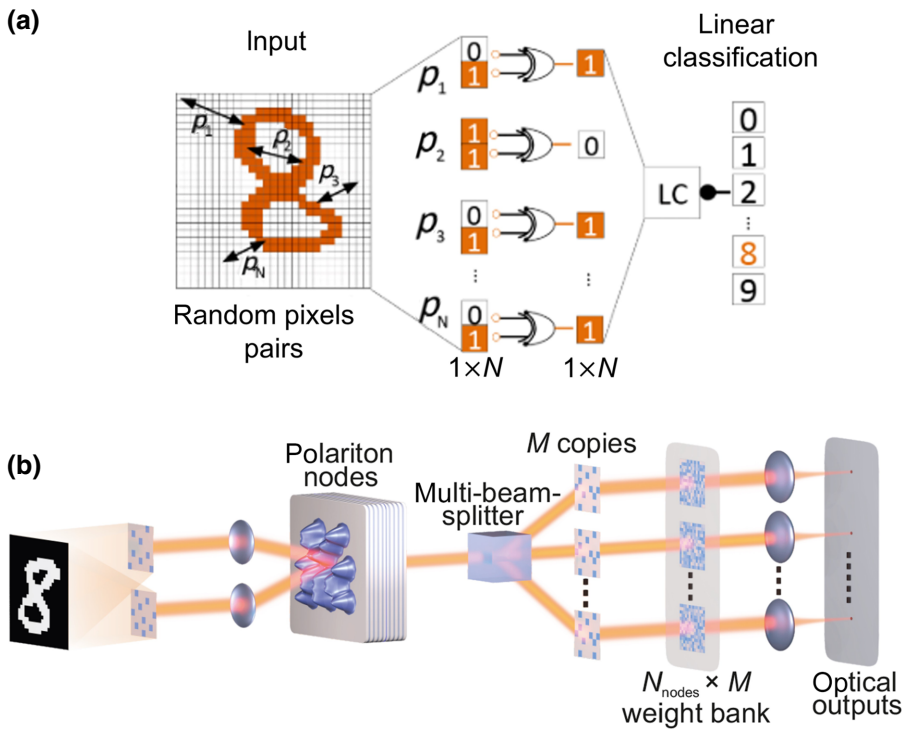


FIG. 2. (a) Conceptual scheme of a single hidden-layer network and (b) a possible all-optical implementation. The raw input (schematically depicted as a handwritten digit) is spatially encoded with light pulses, which are assumed to arrive simultaneously. It is split into two parts visualized with the dotted arrays, which encode the two inputs of multiple XOR gates. Each of the gates corresponds to a single pixel of the array. The transmitted intensity contains the result of the XOR gates if one of the methods of Sec. IV is used. This is subsequently split into M copies, which correspond to the number of classes. The weight banks implement linear classification of the result in the N_{nodes} -dimensional feature space. Light intensity at the output measured by the detectors contains information about the probability that the sample belongs to a certain class.

distant pixels even in the single-layer XOR network. The above stage does not require any nonlinear operation and can be implemented all optically, for example, using a diffractive optical element (DOE) or a three-dimensional laser-written waveguide array. The arrays of pulses, forming the two binary inputs of the gates, are directed at the same area of the sample, creating an array of N_{nodes} nonlinear nodes. The optical output (light transmission) from each of the nodes corresponds to the result of an XOR operation. To obtain an XOR activation function of the nodes, we propose to use one of the methods described in Sec. IV. Further on, a vector-matrix multiplication is realized with linear optical elements. Precisely, each of the N_{nodes} pulses, which constitute output of the nodes, is split into M copies (for example, using a diffractive optical element), where M is the number of classes. These copies are directed at an $N_{\text{nodes}} \times M$ optical filter array, which applies synaptic weights by selectively attenuating intensities of the pulses. The resulting weighted pulses are combined into M outputs. The output of the device consists of M pulses, which are the result of the multiplication. The optical vector-matrix multiplication has been realized both in the case of coherent and incoherent light [17,19,20,27,29,38,82,83]. The intensity at the output is related to the probability that the input sample belongs to the particular class. This intensity can be measured by a detector such as a photodiode to convert it to an electrical signal.

The optical elements necessary to build the above system are readily available. The input can be created using a laser generating picosecond optical pulses, and input information can be encoded with arrays of

micrometer-sized optoelectronic modulators [84–86]. The multi-beam-splitter used for the vector-matrix multiplication can be either implemented with an appropriately shaped diffractive optical element or a system of standard beam splitters. Finally, arrays of sensitive photodetectors can be used to detect the output signals [84]. The optical weight banks can be implemented with any form of a tunable semitransparent filter, for example, using programmable nanophotonics [10,36], a liquid crystal array, or a phase-change material array [25].

B. Numerical results

We simulate the proposed system, assuming that $N_{\text{nodes}} = 8000$ gates are characterized by the same level of inaccuracy in terms of shot-to-shot variance of intensity as in the experiment [31]. Despite the relatively large amount of experimental noise in that experiment, we find that such a network is able to recognize MNIST handwritten digits with 95.9% accuracy, only 0.1% lower than the network of ideal XOR gates. This is close to the 96% accuracy on the MNIST task of the state of the art for hardware neuromorphic implementations [31,87,88], although sophisticated software simulations of neural networks can reach much higher accuracy of recognition, at the cost of much longer operation time and higher energy consumption [89].

C. Total energy consumption

The estimations of energy efficiency presented in Sec. II have to be treated as theoretical lower bounds for the energy efficiency of an ideal system. Here, we attempt to

take into account all sources of energy loss and estimate more realistic energy efficiency including all necessary optical and optoelectronic components. Until now we have only taken into account the optical energy of input pulses. However, most of the information in the modern world is carried electronically, with the exception of optical fiber connections. In the case when information is encoded electronically, we expect that the conversion of electronic to optical signal can lead to a significant overhead in terms of energy efficiency and speed. In particular, commercial ultrashort pulse lasers reach wall-plug efficiencies, or conversion ratio of electric to optical energy, above 10%.

We present a calculation of the total energy consumption of a complete network, taking into account losses occurring at all elements of the system, i.e., laser source, modulators, microcavity, weight banks, and photodetectors. We consider two cases: (A) an “idealized” large scale system, with parameters corresponding to state-of-the-art optical elements, and (B) a proof-of-principle system with a relatively small number of nodes and accessible optical elements.

The average energy per synaptic operation in a binarized network can be calculated as $E_{\text{op}} = E_{\text{total}}/(2N_{\text{nodes}})$, where E_{total} is the total energy consumed by the system and N_{nodes} is the number of binary nodes (gates) in the hidden layer, each gate with two synaptic inputs. The total energy consumption for a complete neural network is

$$E_{\text{total}} = E_{\text{source}} + E_{\text{modulators}} + E_{\text{detectors}}, \quad (1)$$

where $E_{\text{modulators}}$ is the net electrical energy required to supply optoelectronic modulators that encode input data, and $E_{\text{detectors}}$ is the energy required to supply photodetectors at the output layer. In the proposed architecture, all data is injected, processed, and read out in parallel. The number of modulators is equal to the number of input bits, and the number of photodetectors is equal to the number of classes. The microcavity and the weight bank are passive elements, and do not appear directly in the formula above, but the optical energy loss occurring in these elements will influence the required laser power.

The energy of the source laser will be bounded from below by two main factors. It has to be high enough for the operation of all the polariton nodes in the hidden layer, and on the other hand, provide sufficiently high optical energy for the photodetectors in the output layer. Of the two constraints, the first one will be typically the more strict one, due to the large number of nodes and much lower number of classes measured by photodetectors in a standard network. We estimate the electrical energy required to supply the laser as

$$E_{\text{source}} \geq \frac{E_{\text{nodes}}}{\eta_L} = \frac{2N_{\text{nodes}}E_{\text{pulse}}}{\eta_C\eta_L}, \quad (2)$$

where E_{nodes} is the optical energy required to supply all the polariton nodes, $\eta_L < 1$ is the wall-plug efficiency of the laser, E_{pulse} is the minimum energy per pulse as estimated in Sec. II, $\eta_C < 1$ is the microcavity transmission coefficient (ratio of output to input energy transmitted for “1” output logic value). We overestimate the energy by assuming that the energy of polaritons interacting in the microcavity is equal to the energy of pulses transmitted through the microcavity in the “1” output state. We neglect losses at optical elements such as beam splitters and mirrors, as they will contribute marginally to the final efficiency. We can estimate the total energy per operation as

$$E_{\text{op}} \geq \frac{E_{\text{pulse}}}{\eta_C\eta_L} + \frac{N_{\text{mod}}}{N_{\text{nodes}}}E_{\text{mod}} + \frac{N_{\text{det}}}{N_{\text{nodes}}}E_{\text{det}}, \quad (3)$$

where N_{mod} is the number of modulators, E_{mod} is the energy cost per bit for a single modulator, N_{det} is the number of detectors, and E_{det} the energy cost per bit for a single detector. The above formula describes the fact that if a single bit of information is shared as input by many nodes in the hidden layer, which is a typical situation in neural networks, the cost of the input optoelectronic conversion will be divided between all the nodes that use this bit of information. This can reduce the energy cost of generating an input pulse for a particular hidden node by orders of magnitude with respect to the energy cost of the optoelectronic switch. Similarly, the energy of detectors per operation will be proportionally reduced.

The conversion of information from electronic to optical signal can be realized with ultraefficient modulators, which reach energy efficiency of several femtojoules per bit [84,85]. Polariton spin switches were also recently demonstrated to achieve energy efficiency at the femtojoule level [49]. The efficiency of photodetectors has also reached very high (femtojoule) levels [84]. The number of photodetectors N_{det} may be very low. For example, if the task of the network is to distinguish between two classes of objects, only two photodetectors will be necessary, and consequently their share in the total energy budget will be negligible.

A single laser pulse can be efficiently redistributed among inputs. It can be split with DOEs or beam splitters into multiple copies, and used to generate multiple bits of information. A single pulse can even be split and redirected into many devices if its energy is high enough to power many of them, so the energy of a single laser pulse should not be treated as the bound for efficiency of a network. Likewise, if the number of classes is large, sensor arrays incorporating many elements may be used to detect thousands of optical outputs simultaneously, which can make the optoelectronic conversion more efficient.

In Table I we show detailed estimations for two cases, corresponding to an “idealized” case A with parameters

TABLE I. Parameters of the two devices considered in the text and the total energy cost per synaptic operation.

	Case A	Case B
N_{inputs}	10 000	10
N_{outputs}	10 000	2
N_{nodes}	10^6	100
Surface area of a single node	$1 \mu\text{m}^2$	$10 \mu\text{m}^2$
Polariton lifetime γ^{-1}	100 ps	1 ps
Cavity transmission coefficient η_C	0.9	0.1
Laser wall-plug efficiency η_L	0.2	0.1
$E_{\text{modulator}}, E_{\text{detector}}$	1 fJ	1 pJ
Energy cost per operation E_{op}	24 aJ	200 fJ
Energy efficiency (SOP $\text{s}^{-1} \text{W}^{-1}$)	4×10^{16}	5×10^{12}

of state-of-the-art components and microcavities, used to solve a complicated machine-learning task (such as ImageNet image recognition) and a “proof-of-principle” case B with parameters closer to off-the-shelf optical components and not yet very optimized microcavities in which room-temperature polaritons have been already observed [72,73,76]. We consider polaritons with realistic interaction constant [71] $g = 3 \mu\text{eV}\mu\text{m}^2$. In the idealized case A, we assume the energy cost of modulators and photodetectors to be of the order of 1 fJ per pulse as in Ref. [84], cavity polariton lifetime 100 ps and transmission coefficient of 90% (see Sec. V) and 20% wall-plug efficiency of the laser. A large-scale neural network with a million nodes, 10 000 inputs and 10 000 output classes is estimated to reach energy efficiency of $E_{\text{op}} = 24$ aJ per synaptic operation. In the “proof-of-principle” case B, we assume the energy cost of modulators and detectors at 1 pJ, 10% laser wall-plug efficiency, 10% cavity transmission coefficient, and 1-ps polariton lifetime. A simple network with ten inputs, 100 nodes, and two output classes achieves efficiency of $E_{\text{op}} = 200$ fJ per operation, which is 2 orders of magnitude below electronic neuromorphic realizations [78] and 1 order of magnitude below the state-of-the-art specialized electronic neural network accelerator [90].

For completeness, we check if the output contains enough optical energy for photodetectors. A commercial photodetector characterized by noise equivalent power of $6 \text{ pW}/\sqrt{\text{Hz}}$, is capable of detecting a signal of a 130-nW optical power within a 200-MHz bandwidth at a signal to noise ratio equal to 1. This corresponds to the required optical energy of 650 aJ per sample in each photodetector. Considering the high ratio of the number of optical nodes to the number of detectors, we conclude that the output contains enough optical energy in both case A and case B. However, in the case of tasks that require a high bit depth, such as regression, the photodetector sensitivity may become the most relevant limiting factor of energy efficiency.

D. Comparison with other systems

We first compare the inference efficiency of our network and non-neuromorphic systems that implement software simulations of neural networks. We consider the most energy-efficient supercomputer from the Green500 list, which performs a single floating-point operation (FLOP) operation at the cost of approximately 60 pJ. We assume that in a computer simulation of an artificial neural network a single operation roughly corresponds to a single synaptic event, such as accumulation of a single weighted input, 1 FLOP \approx 1 SOP. This gives the energy efficiency of $1.7 \times 10^{10} \text{ OP s}^{-1} \text{ W}^{-1}$ (operations per second per watt). This estimation does not take into account the energy cost of external memory access, which is crucial in neural network simulations and can be much higher than the cost of the computation itself. Memory access is a bottleneck in CMOS neural network implementations [67] with the minimum cost per bit at the level of 10 pJ. The efficiency of $10^{13} \text{ OP s}^{-1} \text{ W}^{-1}$ was reported for the state-of-the-art energy-optimized neural network accelerator [91], but such a high efficiency was reported for convolutional layers of the network only, which are characterized by a relatively small number of external memory access operations. In contrast, in our neuromorphic design the cost of memory access is zero, as there is no need to store the state of the system in external memory during computation. Finally, we mention that off-the-shelf GPU systems [5] can reach the energy efficiency at the level of $10^9 \text{ OP s}^{-1} \text{ W}^{-1}$.

In some other works [35,88], efficiency of hardware neural networks was determined by estimating the number of FLOPs required to simulate the network rather than comparing the number of SOPs. Such an approach can give estimates orders of magnitude higher but here we choose to take a more conservative approach by focusing on the number of achievable SOPs. In the case of a disordered dopant atom network [88], it was suggested that a theoretical efficiency limit of $10^{13} \text{ SOP s}^{-1} \text{ W}^{-1}$ could be achieved [93], but this estimation does not account for the cost of the output signal thresholding and additional processing of the output, which have to be performed electronically in this case. The comparison with other neuromorphic and non-neuromorphic systems is presented in Fig. 3, which also indicates the efficiency of our recent experimental implementation of a network based on a binarized polariton node with nonresonant pumping [31]. According to this comparison, we estimate that an exciton polariton network has the potential to outperform other solutions by orders of magnitude. In the above comparison we do not include estimations made for optical systems that perform linear operations (in the function of inputs), such as vector-matrix multiplication [32,35–39,94], since such systems alone are not able to efficiently perform complex machine-learning tasks that require nonlinearity.

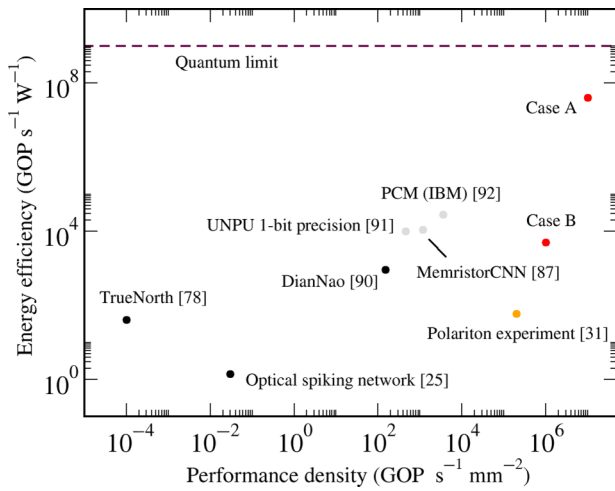


FIG. 3. Comparison of energy-efficiency and performance-density estimates of polariton systems (red dots) and other systems [25,31,78,87,90–92]. Gray points correspond to systems for which estimates of only a certain part of the network is provided [87,91], or the implementation is partly realized by software [87,92]. The orange dot corresponds to the experiment with nonresonantly pumped polaritons [31]. The dashed line is the single-photon quantum limit for optics.

E. Footprint and performance density

To estimate footprint, we consider the minimal physical size of optical weight banks that encode synaptic weights. The size of microcavity neurons is comparable to the size of weight banks, but their number is smaller. The size is limited by the optical wavelength and typically cannot be decreased below micrometer level. We cautiously assume the size of $10 \mu\text{m}^2$ for both the optical weight and optical node. Each weight can be realized with a semi-transparent filter, either embedded within a spatial light modulator array, or even with a nonadjustable filter such as a patterned glass surface, if the weights can be set at the fabrication stage.

While this size is much larger than the size of an electronic transistor, it is unlikely to be the main limiting factor for the footprint. One million polariton neurons or optical weights can be implemented on a 10mm^2 area, which is comparable to the density in a neuromorphic CMOS architecture [78]. The number of parameters in the leading artificial neural networks participating in the ImageNet competition [67] is of the order of 10^8 . Such a number of parameters would require an optical weight bank surface of a few cm^2 .

The most relevant performance measure is the performance density, which is the number of operations that can be performed per surface area per time [67]. The practical maximum rate of incident pulses in a polariton system can be estimated to be of the order $1/(100 \text{ps})$. After injecting the input pulses, a typical polariton system needs the “cooling-off period” of the order of 100ps to completely

recover to the original state, so that the response of the cavity is not affected by the previous pulses [31]. This rate is given by the lifetime of reservoir excitons, or unwanted excitations in the quantum well, and may strongly depend on the material used. Such excitations occur even in the case of resonant excitation. Here, we assume that the cavity polariton lifetime is shorter than 100ps . The resulting performance density is orders of magnitude higher than in the leading CMOS systems [67]. Figure 3 presents comparison of performance density with respect to other systems.

However, the performance density does not necessarily tell the full story. It is also necessary to consider the aspect of heat generation in the calculation of footprint. In electronic devices, due to heat generation, it is difficult to implement multilayer structure in three-dimensional chips. One of the dimensions is sacrificed for a heat drain, and it is the reason why performance density is measured in operations per surface and not per volume. Hence, calculation of footprint should not be considered without the consideration of energy dissipation, which is also here the ultimate limiting factor. In the case of a system with much lower energy loss but larger surface area, it is possible to consider multilayer wiring, exploiting the third dimension, which is typically occupied by a heat sink. For example, a physical multilayer network can be considered, where each layer is placed on top of the previous one. In the following section we discuss the possible integration of the proposed system inside an optical chip.

F. Extensions and generalizations

We consider several possible extensions of the proposed design. The system depicted in Fig. 2 is based on free-space optics, but an integrated version is required for a portable device. We note that the idea behind free-space optics does not necessarily imply propagation in free space. The same principles can be applied to design an integrated system, where light propagates in an optical medium, such as glass or a transparent semiconductor. In fact, the most common integrated devices that rely on free-space optics are mobile phone cameras. Millimeter sized camera objectives are used together with integrated semiconductor arrays, where each optical sensor is micrometer-sized. Analogously, it is possible to design portable neural networks based on micrometer-sized nodes and focal lengths of the order of millimeters. Figure 4 illustrates the possible geometry of an integrated polariton neural network. A device relying on a similar concept has been demonstrated in Ref. [82]. Propagation is in this case along the cavity plane and lenses are replaced by diffractive optical elements, which can be designed to precisely control light paths.

An alternative to free-space propagation is the use of integrated optical waveguides [25,36]. Waveguide systems

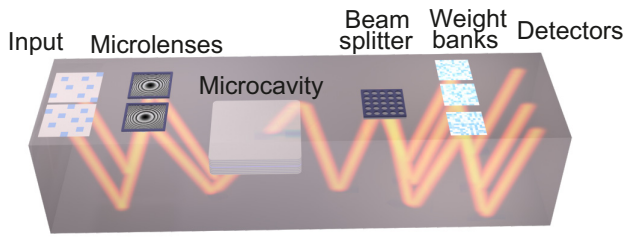


FIG. 4. Scheme of a possible integrated version of the free-space system. The subsequent layers of the system are arranged side by side in the plane of a planar waveguide. Elements such as lenses and beam splitters can be implemented with diffractive optical elements.

can be implemented in a compatible semiconductor platform, and can be used not only to transport light, but also to build efficient devices for linear operations, such as convolutions or vector-matrix multiplication [36,37]. It is also possible to consider a hybrid architecture where light propagation is partly based on free-space optics and partly on light confinement in waveguides. Waveguide systems have two potential drawbacks: they suffer from photon losses, in particular at waveguide bends, and require a separate waveguide for each optical path. On the other hand, free-space connectivity can be almost loss free, and allows arbitrary intersection of optical paths in a linear medium, which can greatly improve data bandwidth [43,95].

Another interesting possible extension is the use of holography for encoding synaptic weights [18]. In this case, data can be efficiently stored in a three-dimensional, rather than two-dimensional, weight bank. This can improve memory capacity significantly, but on the other hand, writing and retrieving information requires complicated optical systems.

A straightforward generalization of our design is to use neurons with analog, rather than binary, inputs and outputs. In this work we choose the simple XOR gate as an example of a neuron that already allows complex machine-learning tasks to be performed. From the point of view of machine learning, the XOR gate, in contrast to OR, AND gates, is itself a classification task that requires nonlinearity to be solved [31]. Moreover, because it is a sufficiently nonlinear function, it can be used as a basic building block for the construction of more complex networks [31,80,81]. In the case of analog neurons, a similar condition exists. The activation function has to be strongly nonlinear, with a negative differential response [31], in order to be useful for the efficient solution of complex machine-learning problems. Optical weights may also modify phases of optical pulses rather than intensities, which can reduce optical losses and increase the available parameter space.

An extension to a multilayer system is possible, thanks to the resonant character of the input, which means that input and output pulses have the same optical frequencies.

In this case, multiple weight banks apply the synaptic connections between neurons in subsequent layers, and each neuron layer can be implemented with a separate microcavity. On the other hand, in this case losses become a more serious issue and may accumulate exponentially as light is transmitted through a number of layers. In Sec. V we show how an appropriate design of a microcavity can circumvent losses to a large extent. Nevertheless, in the case of a large number of layers, some kind of pulse regeneration method will likely be required. Note that difficult tasks such as ImageNet can be solved very efficiently with networks containing only a few layers, for example by AlexNet (eight layers).

Finally, the vector-matrix multiplication can be replaced by a convolution operation [37,38]. This is particularly relevant for image recognition tasks, where convolutional layers play a very useful role [89], or in general in tasks that are characterized by translational invariance, either in space or in time.

IV. IMPLEMENTATION OF BINARIZED NEURONS

We propose two methods of implementation of XOR gates in a microcavity system in the case when input pulses are quaresonant with the frequency of the microcavity. Resonant excitation is better suited to energy-efficient processing than the nonresonant excitation method used in the recent experiment [31], as in the latter case input pulses do not create exciton polaritons directly. Instead, high-energy excitations in the so-called reservoir are created, and only a fraction of them are converted into polaritons. Moreover, with the resonant excitation method, the frequency of the input and the output are equal, which facilitates realization of multilayer networks. However, spectral filtering used in Ref. [31] cannot be directly used to obtain a negative response in the case of resonant excitation, as the frequency of photons is fixed.

The first method is based on cross-phase modulation and the use of interferometers. We assume that “0” is encoded with the absence of a pulse or a low-intensity pulse, and “1” with a high-intensity pulse. Before entering the microcavity, each of the input pulses is split into two equal copies, see Fig. 5. One of the two copies is sent around the microcavity, while the other copy is incident on the surface of the microcavity. The second input pulse is split and directed in the same way. The copies of the two input pulses meet at the microcavity, where they interact through polariton-polariton scattering. This induces both self-phase modulation, which is independent of the other pulse, and cross-phase modulation, which is dependent on the intensity of the other pulse. The phase acquired by cross-phase modulation is proportional to the interaction energy between polaritons in the two pulses, i.e., each pulse $i = 1, 2$ acquires the phase $\delta\phi_i \approx g_{\text{XPM}}(n_{3-i})\tau$,

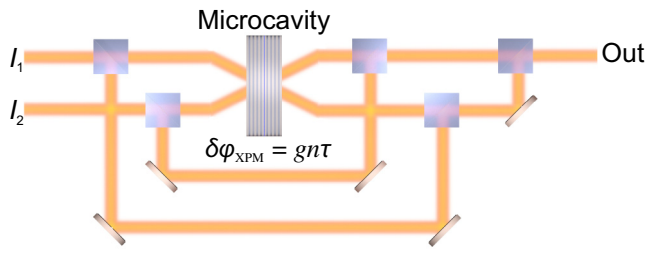


FIG. 5. Scheme of a possible XOR gate implementation. The principle of operation is based on cross-phase modulation (XPM) of two pulses inside the microcavity and interferometry. With appropriate tuning of optical path lengths, interference of cavity emission with copies of input pulses results in intensity minimum for the “11” input configuration.

where $\langle n_{3-i} \rangle$ is the average density of polaritons in the other pulse during the time of interaction, which is given by τ . As the two pulses are incident at different angles, they separate at the exit from the microcavity. Each of the transmitted pulses is then combined with its copy that did not go through the microcavity, which results in interference. Interference is destructive when the cross-phase modulation phase $\delta\phi$ together with the phase difference accumulated along optical paths is equal to π . By adjusting the delay lines between the interfering pulses accordingly, it is possible to obtain destructive interference for input pulses that are in the “11” state.

As we are only interested in order-of-magnitude estimations, we neglect the temporal dynamics of n and assume that it is an average of density calculated over the duration of the interaction τ . For optimal interaction, the temporal length of the pulse and the lifetime of polaritons in the microcavity should be similar. This ensures that polaritons can be effectively created in the microcavity and most of them can be present and interact there at the same time. It also ensures that the spectral density of the pulse matches the width of the spectral resonance of the microcavity. This requirement leads to the condition $gn = \hbar\gamma$ mentioned in Sec. II.

The drawback of the above method is that it uses an interferometric technique, which requires coherence of light and precise adjustment of optical path lengths for all nodes, which may be impractical. These shortcomings are not present in the second method, which utilizes the anisotropy of interactions of exciton polaritons [46]. In this case, we assume that the two input pulses are mutually incoherent. Interactions between polaritons with opposite spin (or circular polarization) are very weak, while polaritons with the same spin interact strongly [96]. This allows construction of a simple XOR gate using two 1/8-wave plates, one placed in front and one behind the cavity, and a polarizing filter. The principle of the method is shown in Fig. 6. In this case, the two input pulses are assumed to be linearly polarized in a well-defined direction (X). At

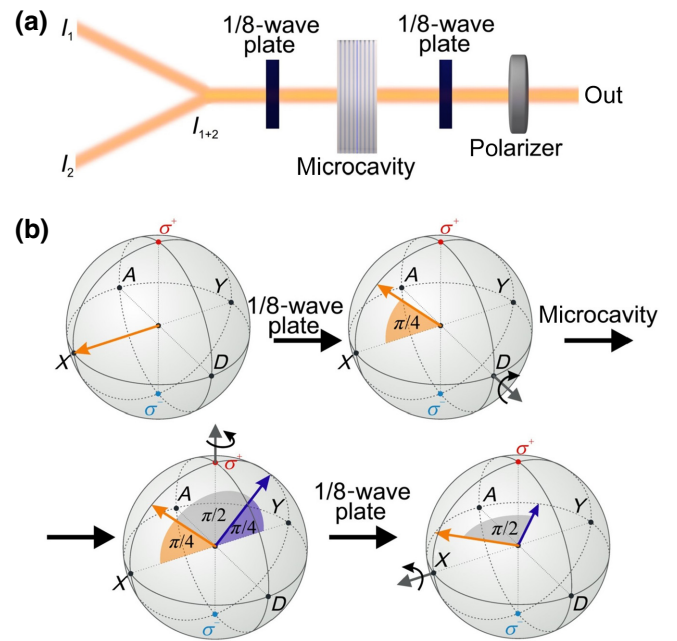


FIG. 6. (a) Scheme of the XOR gate setup based on polarization rotations. (b) Details of rotations visualized on the Poincaré sphere. The input pulses are assumed to be linearly polarized. The first 1/8-wave plate rotates the polarization by the angle $\pi/4$ around the diagonal-antidiagonal polarization axis. The spin-anisotropic polariton interaction results in the rotation of polarization around the circular polarizations axis, which is intensity dependent. This results in $\pi/2$ angle between polarizations of the outputs from the microcavity in the input configurations “10/01” (orange) and “00/11” (purple). The second 1/8-wave plate rotates the polarization back to the equator (linear polarization) plane. The polarizer at the output blocks the polarized emission corresponding to the “00/11” configuration.

the entrance, they are combined into one. Before entering the microcavity, the combined pulse goes through a 1/8-wave plate, which rotates the polarization by 45° on the Poincaré sphere. Inside the microcavity, the polaritons interact, which leads to nonlinear spin rotation around the z axis on the Poincaré sphere, which is dependent on the intensity of the input pulse. We assume that the intensity of the pulse in the “10” or “01” input combination leads to a rotation that is different by π with respect to the “00/11” combination. As in Sec. II, this requires that the density and time of interaction fulfill the condition $gn\tau \approx 1$. After exiting the cavity, the polarization is rotated again by 45° using a 1/8-wave plate back to the equator plane, which results in linear polarization of pulses, orthogonal in the case of “01/10” and “00/11” polarizations. The pulse now passes through the polarizer, which blocks the light corresponding to the “00/11” polarization, which results in an XOR gate output. A similar concept was proposed to realize a spin transistor [97], where spin rotation due to

the polariton energy splitting followed by interference was used.

V. MICROCAVITY TRANSPARENCY

Efficient operation of the proposed neural network requires low energy losses at optical elements. On the other hand, to enhance polariton-polariton interactions, microcavities with very high quality factor are required. The use of high-reflectivity mirrors in such cavities may seem to contradict the assumed high transmission coefficients as in Table I. However, this is not the case, and in this section we show that high- Q cavities can be characterized by high transmission rates at resonance.

Consider transmission of a light pulse through a dielectric microcavity structure depicted in Fig. 7, where mirrors consist of alternating layers of materials with refractive indices n_1 and n_2 , and the cavity medium has the index n_c . The coefficients of transmission and reflection are conveniently calculated using the transfer matrix formalism [46]. At the center of the stop band we assume $n_j d_j = \lambda/4$, where λ is the light wavelength in vacuum, d_j is the thickness of the mirror layer j , and the transfer matrix is

$$\mathbf{T}_j = \begin{pmatrix} 0 & i/n_j \\ in_j & 0 \end{pmatrix}. \quad (4)$$

Two layers of material with refractive indices n_1 and n_2 stacked next to each other correspond to the transfer matrix, which is the product of \mathbf{T}_2 and \mathbf{T}_1

$$\mathbf{T}_{12} = \begin{pmatrix} -n_1/n_2 & 0 \\ 0 & -n_2/n_1 \end{pmatrix}. \quad (5)$$

Perfect transmission can be achieved when the complete transfer matrix of the microcavity is proportional to the identity matrix, cf. Fig. 7

$$\mathbf{T}_{\text{mc}} = \mathbf{T}_{12}^N \mathbf{T}_c \mathbf{T}_{21}^N \sim \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (6)$$

where N is the number of dielectric layer pairs on each side. The above condition is fulfilled when the width of the cavity layer is such that $n_c d_c = \lambda/2$.

The bottom panel of Fig. 7 shows the transmission spectrum of a microcavity composed of 25 pairs of layers on each side of the microcavity, and fulfills the above condition. Despite the high- Q factor of the cavity and high reflectivity of the mirrors, almost perfect transmission occurs at the resonant frequency. Similarly, a polariton cavity containing quantum-well excitons can be designed to exhibit such a high transmission at resonance.

In practice, effects such as cavity imperfections, absorption, and nonradiative exciton decay will decrease the peak transmission rate. Currently, polariton cavities can reach transmission coefficients at the level of a few tens of percent.

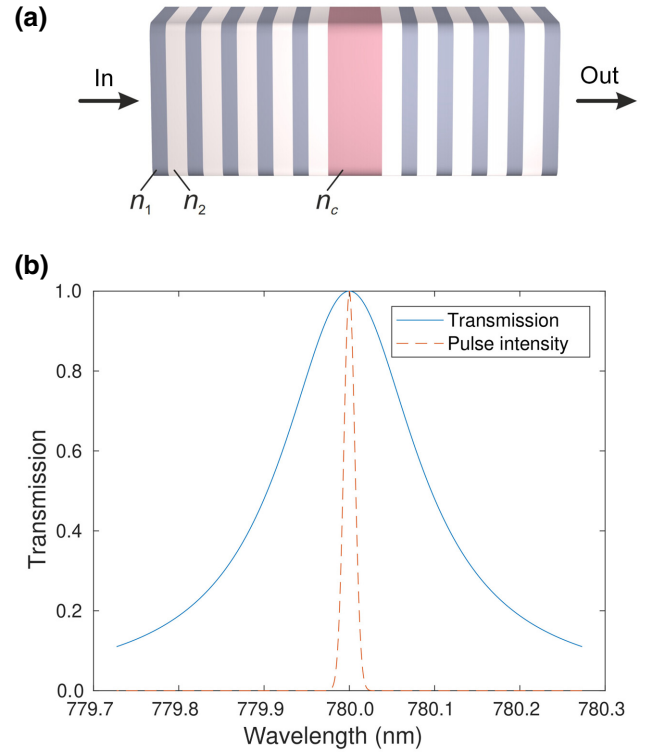


FIG. 7. (a) Example of a microcavity considered in the text. (b) Solid line shows numerically calculated transmission spectrum of a microcavity with mirrors composed of 25 pairs of layers each with refractive indices $n_1 = 3$ and $n_2 = 3.5$ and the cavity with refractive index $n_c = 3$. Widths of the mirror layers are $\lambda/4n_i$, where λ is the central wavelength, and the width of the cavity is $\lambda/2n_c$. The dashed line shows the normalized spectrum of a laser pulse with 50-ps temporal FWHM.

VI. NEURAL NETWORKS VERSUS UNIVERSAL COMPUTING

We would like to point out several advantages of our neural network design with respect to the more conventional universal digital logic. Miller [98] summarized the difficulties in using optical components for information processing and pointed out conditions that have to be fulfilled by a reasonable candidate for digital logic. However, some of these conditions, including logic level restoration condition, or “cleaning up” the signal, and the fan-out condition are only necessary in the implementations that follow the architecture of conventional computers. In particular, they are not necessary in the case of single hidden-layer networks implemented in hardware. They are also not absolutely necessary for deep feed-forward networks, if optical signals are not redirected recurrently.

Nevertheless, in our system some of the conditions of Ref. [98] are fulfilled: (a) critical biasing is not an issue since the input-output operating point is not thresholdlike; (b) input-output isolation can be achieved by a simple method of directing the input pulses in free space

at an angle with respect to optical surfaces, which results in reflected pulses propagating along different trajectories than the input pulses; (c) the condition of cascadability can be fulfilled in the case of resonant excitation, when the frequency of the output is the same as the input, provided that some method of optical signal restoration or amplification is implemented [99]; (d) while the logic level is dependent on loss in our system, losses can be minimized as described in Sec. V, and are proportional in each layer-to-layer connection, which enables definition of signal amplitudes for logic levels accordingly in each layer. At the same time, it is possible to correct for the difference in loss in different optical paths by adjusting the corresponding optical weights.

The use of a feed-forward neural network architecture instead of universal digital computing results in another useful aspect of the system, which is the absence of separate memory units. Currently, there is no reliable implementation of optical memory that could be used to implement a traditional computing architecture with polaritons. Even if such technology existed, it would be required to store and read information with ultrashort, ultralow energy pulses to comply with the energy efficiency of the entire system, which may be difficult to realize in practice.

VII. QUANTUM LIMIT OF ENERGY EFFICIENCY

Since our device is based on classical physics, it should operate well above the quantum limit to avoid quantum noise. It is interesting to compare the potential energy efficiency of our system to the single-photon limit, since the fundamental energy efficiency bound is imposed, as in the case of electronics, by quantum effects. The quantum limit corresponds to a single photon per bit of information or two photons per XOR gate. This leads to the estimation of 10^{18} SOP s^{-1} W^{-1} , as depicted in Fig. 3. The onset of the quantum limit has been already observed in tightly confined exciton-polariton microcavities [77,100]. We would like to emphasize that even if XOR nodes operated close to the quantum limit, few-photon emitters or detectors would not be needed. The “quantum limit” in the sense considered here applies to the operation of network nodes in the hidden layer. Since there are no separate detectors or emitters in these nodes, we do not need to be concerned with the detection of these weak signals. As already noted, output pulses arriving at the detectors in the final layer can have intensities orders of magnitude higher than the pulses that perform individual operations in the hidden-layer nodes.

VIII. DISCUSSION

In this work we analyze the potential energy efficiency of all-optical exciton-polariton-based systems for information processing and propose a design of a simple neural

network for data classification. The advantage of performance density and energy efficiency is achieved in the inference, while the training stage is assumed to be performed in software. The software model of a neural network resulting from training can be implemented in optical hardware by setting the appropriate optical weights. In this sense, the proposed system is an example of an optical neural network accelerator [8,90,91], where a previously trained software model is implemented in hardware to increase the performance of inference. Such systems are best suited to machine-learning tasks where analysis of a large number of samples is required, while retraining or replacement of the model is not necessary or performed infrequently. Such tasks include image, speech and video recognition, natural language processing, automatic detection, and control systems.

A practical system is required to operate at room temperature. While the majority of laboratory experiments with exciton polaritons are performed at cryogenic environments, there are already many experiments with room-temperature exciton-polariton systems, including inorganic [72,76] and organic [101–103] semiconductors, two-dimensional materials [104] and perovskites [73,75]. These materials are typically characterized by a more prominent sample disorder, but if scattering on disorder is not strong enough to make the signal in the output too weak, it can be part of the internal losses that do not alter the efficiency of the operation. On the other hand, the variation of node response due to disorder and variability from one sample to another can be corrected for by appropriately adjusting optical weights at the input and output layers. At the same time, we expect that heating will be a negligible effect due to the low optical powers required. Effects of heating are not typically observable in experiments with polaritons, except for the highest pumping rates.

The possible sources of optical losses are expected to be associated with imperfection of cavity mirrors and irreversible (polariton unrelated) absorption of photons in the sample. The technology of microcavity fabrication is sufficiently advanced to produce very clean mirrors, which results in extremely high optical mode quality factors. Irreversible absorption may result in sample heating, however this effect is expected to be negligible at low optical pulse energies. The creation of a long-lived exciton reservoir by resonant laser pulses is also possible [105], but this effect becomes less relevant for ultrashort pulses. Nevertheless, we accounted for the exciton reservoir lifetime in our estimation of maximum possible data rate in Sec. III E. The use of picosecond pulse laser sources is required, which contributes to the overall size of the system. The possible alternative to bulky ultrashort pulse lasers is the use of compact vertical-cavity surface-emitting lasers together with ultrafast optical modulators to shape picosecond input pulses [106].

ACKNOWLEDGMENTS

M.M. acknowledges support from National Science Center, Poland Grant No. 2017/25/Z/ST3/03032 under the QuantERA program. A.O. acknowledges support from National Science Center, Poland Grant No. 2016/22/E/ST3/00045. R.M. acknowledges support from National Science Center, Poland Grant No. 2019/33/N/ST3/02019. B.P. acknowledges support from National Science Center, Poland Grant No. 2020/37/B/ST3/01657. K.T. acknowledges support from National Science Center, Poland Grant No. 2020/04/X/ST7/01379. T.L. acknowledges the support of the Singapore Ministry of Education, via the Academic research fund project MOE2019-T2-1-004.

-
- [1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature* **521**, 436 (2015).
- [2] J. Misra and I. Saha, Artificial neural networks in hardware: A survey of two decades of progress, *Neurocomputing* **74**, 239 (2010).
- [3] A. S. Andrae and T. Edler, On global electricity usage of communication technology: Trends to 2030, *Challenges* **6**, 117 (2015).
- [4] E. Strubell, A. Ganesh, and A. McCallum, Energy and policy considerations for modern deep learning research, *Proc. AAAI Conf. Artif. Intelligence* **34**, 13693 (2020).
- [5] A. Canziani, A. Paszke, and E. Culurciello, An analysis of deep neural network models for practical applications, [arXiv:1605.07678](https://arxiv.org/abs/1605.07678) (2017).
- [6] D. Li, X. Chen, M. Becchi, and Z. Zong, in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, (IEEE, 2016), p. 477.
- [7] M. M. Waldrop, The chips are down for Moore's law, *Nat. News* **530**, 144 (2016).
- [8] K.-i. Kitayama, M. Notomi, M. Naruse, K. Inoue, S. Kawakami, and A. Uchida, Novel frontier of photonics for data processing—photonic accelerator, *APL Photonics* **4**, 090901 (2019).
- [9] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, Physics for neuromorphic computing, *Nat. Rev. Phys.* **2**, 499 (2020).
- [10] W. Bogaerts, D. Pérez, J. Capmany, D. A. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni, Programmable photonic circuits, *Nature* **586**, 207 (2020).
- [11] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, Inference in artificial intelligence with deep optics and photonics, *Nature* **588**, 39 (2020).
- [12] B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, Photonics for artificial intelligence and neuromorphic computing, *Nat. Photonics* **15**, 102 (2021).
- [13] A. N. Tait, T. Ferreira De Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Silicon Photonic Modulator Neuron, *Phys. Rev. Appl.* **11**, 064043 (2019).
- [14] R. Amin, J. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, and T. El-Ghazawi, *et al.*, ITO-based electro-absorption modulator for photonic neural activation function, *APL Mater.* **7**, 081112 (2019).
- [15] 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. Lightwave Technol.* **32**, 3427 (2014).
- [16] M. Brunstein, A. M. Yacomotti, I. Sagnes, F. Raineri, L. Bigot, and A. Levenson, Excitability and self-pulsing in a photonic crystal nanocavity, *Phys. Rev. A* **85**, 031803(R) (2012).
- [17] J. W. Goodman, A. Dias, and L. Woody, Fully parallel, high-Speed incoherent optical method for performing discrete Fourier transforms, *Opt. Lett.* **2**, 1 (1978).
- [18] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, Holography in artificial neural networks, *Nature* **343**, 325 (1990).
- [19] N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, Optical implementation of the Hopfield model, *Appl. Opt.* **24**, 1469 (1985).
- [20] T. Lu, S. Wu, X. Xu, and T. Francis, Two-dimensional programmable optical neural network, *Appl. Opt.* **28**, 4908 (1989).
- [21] D. Psaltis and N. Farhat, Optical information processing based on an associative-memory model of neural nets with thresholding and feedback, *Opt. Lett.* **10**, 98 (1985).
- [22] D. Psaltis, D. Brady, and K. Wagner, Adaptive optical networks using photorefractive crystals, *Appl. Opt.* **27**, 1752 (1988).
- [23] A. N. Tait, T. Ferreira 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 (2017).
- [24] T. W. Hughes, I. A. Williamson, M. Minkov, and S. Fan, Wave physics as an analog recurrent neural network, *Sci. Adv.* **5**, eaay6946 (2019).
- [25] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities, *Nature* **569**, 208 (2019).
- [26] B. Shi, N. Calabretta, and R. Stabile, Deep neural network through an InP SOA-based photonic integrated cross-connect, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1 (2019).
- [27] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, All-optical neural network with nonlinear activation functions, *Optica* **6**, 1132 (2019).
- [28] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Experimental demonstration of reservoir computing on a silicon photonics chip, *Nat. Commun.* **5**, 3541 (2014).
- [29] J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, Reinforcement learning in a large-scale photonic recurrent neural network, *Optica* **5**, 756 (2018).

- [30] D. Ballarini, A. Gianfrate, R. Panico, A. Opala, S. Ghosh, L. Dominici, V. Ardizzone, M. De Giorgi, G. Lerario, G. Gigli, T. C. H. Liew, M. Matuszewski, and D. Sanvitto, Polaritonic neuromorphic computing outperforms linear classifiers, *Nano Lett.* **20**, 3506 (2020).
- [31] R. Mirek, A. Opala, P. Comaron, M. Furman, M. Król, K. Tyszka, B. Sereżyński, D. Ballarini, D. Sanvitto, T. C. H. Liew, W. Pacuski, J. Suffczyński, J. Szczytko, M. Matuszewski, and B. Piętka, Neuromorphic binarized polariton networks, *Nano Lett.* **21**, 3715 (2021).
- [32] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, *Science* **361**, 1004 (2018).
- [33] P. Antonik, N. Marsal, and D. Rontani, Large-scale spatiotemporal photonic reservoir computer for image classification, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1 (2019).
- [34] P. Antonik, N. Marsal, D. Brunner, and D. Rontani, Human action recognition with a large-scale brain-inspired photonic computer, *Nat. Machine Intelligence* **1**, 530 (2019).
- [35] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit, *Nat. Photonics* **15**, 367 (2021).
- [36] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, Deep learning with coherent nanophotonic circuits, *Nat. Photonics* **11**, 441 (2017).
- [37] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, and A. Raja, *et al.*, Parallel convolutional processing using an integrated photonic tensor core, *Nature* **589**, 52 (2021).
- [38] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification, *Sci. Rep.* **8**, 1 (2018).
- [39] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, and R. Morandotti, *et al.*, 11 TOPS photonic convolutional accelerator for optical neural networks, *Nature* **589**, 44 (2021).
- [40] Rise of the learning machines, *Nat. Photon.* **15**, 65 (2021).
- [41] S. P. Rodrigues, Z. Yu, P. Schmalenberg, J. Lee, H. Iizuka, and E. M. Dede, Weighing in on photonic-based machine learning for automotive mobility, *Nat. Photonics* **15**, 66 (2021).
- [42] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, in *2011 38th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2011), p. 365.
- [43] D. A. Miller, Attojoule optoelectronics for low-energy information processing and communications, *J. Lightwave Technol.* **35**, 346 (2017).
- [44] C. Kachris and I. Tomkos, A survey on optical interconnects for data centers, *IEEE Commun. Surv. Tutorials* **14**, 1021 (2012).
- [45] A. Opala, S. Ghosh, T. C. H. Liew, and M. Matuszewski, Neuromorphic Computing in Ginzburg-Landau Polariton-Lattice Systems, *Phys. Rev. Appl.* **11**, 064029 (2019).
- [46] A. Kavokin, J. J. Baumberg, G. Malpuech, and F. P. Laussy, *Microcavities* (Oxford University Press, Oxford, 2017).
- [47] I. Carusotto and C. Ciuti, Quantum fluids of light, *Rev. Mod. Phys.* **85**, 299 (2013).
- [48] P. M. Walker, L. Tinkler, D. V. Skryabin, A. Yulin, B. Royall, I. Farrer, D. A. Ritchie, M. S. Skolnick, and D. N. Krizhanovskii, Ultra-low-power hybrid light-matter solitons, *Nat. Commun.* **6**, 8317 (2015).
- [49] A. Dreismann, H. Ohadi, Y. del Valle-Inclan Redondo, R. Balili, Y. G. Rubo, S. I. Tsintzos, G. Deligeorgis, Z. Hatzopoulos, P. G. Savvidis, and J. J. Baumberg, A sub-femtojoule electrical spin-switch based on optically trapped polariton condensates, *Nat. Mater.* **15**, 1074 (2016).
- [50] J. Kasprzak, M. Richard, S. Kundermann, A. Baas, P. Jeambrun, J. M. J. Keeling, F. M. Marchetti, M. H. Szymanska, R. André, J. L. Staehli, V. Savona, P. B. Littlewood, B. Deveaud, and L. S. Dang, Bose-Einstein condensation of exciton polaritons, *Nature* **443**, 409 (2006).
- [51] A. Amo, J. Lefrère, S. Pigeon, C. Adrados, C. Ciuti, I. Carusotto, R. Houdré, E. Giacobino, and A. Bramati, Superfluidity of polaritons in semiconductor microcavities, *Nat. Phys.* **5**, 805 (2009).
- [52] R. T. Juggins, J. Keeling, and M. H. Szymańska, Coherently driven microcavity-polaritons and the question of superfluidity, *Nat. Commun.* **9**, 4062 (2018).
- [53] D. Caputo, D. Ballarini, G. Dagvadorj, C. Sánchez-Muñoz, M. De Giorgi, L. Dominici, K. West, L. N. Pfeiffer, G. Gigli, F. P. Laussy, M. H. Szymanska, and D. Sanvitto, Topological order and thermal equilibrium in polariton condensates, *Nat. Mater.* **17**, 145 (2017).
- [54] K. G. Lagoudakis, F. Manni, B. Pietka, M. Wouters, T. C. H. Liew, V. Savona, A. V. Kavokin, R. André, and B. Deveaud-Plédran, Probing the Dynamics of Spontaneous Quantum Vortices in Polariton Superfluids, *Phys. Rev. Lett.* **106**, 115301 (2011).
- [55] L. Dominici, R. Carretero-González, A. Gianfrate, J. Cuevas-Maraver, A. S. Rodrigues, D. J. Frantzeskakis, G. Lerario, D. Ballarini, M. De Giorgi, G. Gigli, P. G. Kevrekidis, and D. Sanvitto, Interactions and scattering of quantum vortices in a polariton fluid, *Nat. Commun.* **9**, 1467 (2018).
- [56] P. St-Jean, V. Goblot, E. Galopin, A. Lemaître, T. Ozawa, L. Le Gratiet, I. Sagnes, J. Bloch, and A. Amo, Lasing in topological edge states of a one-dimensional lattice, *Nat. Photonics* **11**, 651 (2017).
- [57] S. Klembt, T. Harder, O. Egorov, K. Winkler, R. Ge, M. Bandres, M. Emmerling, L. Worschech, T. Liew, and M. Segev, *et al.*, Exciton-polariton topological insulator, *Nature* **562**, 552 (2018).
- [58] A. Amo, T. C. H. Liew, C. Adrados, R. Houdré, E. Giacobino, A. V. Kavokin, and A. Bramati, Exciton-polariton spin switches, *Nat. Photonics* **4**, 361 (2010).
- [59] M. De Giorgi, D. Ballarini, E. Cancellieri, F. M. Marchetti, M. H. Szymanska, C. Tejedor, R. Cingolani, E. Giacobino, A. Bramati, G. Gigli, and D. Sanvitto, Control and Ultrafast Dynamics of a Two-Fluid Polariton Switch, *Phys. Rev. Lett.* **109**, 266407 (2012).

- [60] D. Ballarini, M. D. Giorgi, E. Cancellieri, R. Houdré, E. Giacobino, R. Cingolani, A. Bramati, G. Gigli, and D. Sanvitto, All-optical polariton transistor, *Nat. Commun.* **4**, 1778 (2013).
- [61] A. V. Zasedatelev, A. V. Baranikov, D. Urbonas, F. Scafirimuto, U. Scherf, T. Stöferle, R. F. Mahrt, and P. G. Lagoudakis, A room-temperature organic polariton transistor, *Nat. Photonics* **13**, 378 (2019).
- [62] T. Gao, P. S. Eldridge, T. C. H. Liew, S. I. Tsintzos, G. Stavrinidis, G. Deligeorgis, Z. Hatzopoulos, and P. G. Savvidis, Polariton condensate transistor switch, *Phys. Rev. B* **85**, 235102 (2012).
- [63] A. V. Baranikov, A. V. Zasedatelev, D. Urbonas, F. Scafirimuto, U. Scherf, T. Stöferle, R. F. Mahrt, and P. G. Lagoudakis, All-optical cascaded universal logic gate with sub-picosecond operation, arXiv:2005.04802 (2020).
- [64] T. C. H. Liew, A. V. Kavokin, and I. A. Shelykh, Optical Circuits Based on Polariton Neurons in Semiconductor Microcavities, *Phys. Rev. Lett.* **101**, 016402 (2008).
- [65] T. Espinosa-Ortega and T. C. H. Liew, Perceptrons with Hebbian Learning Based on Wave Ensembles in Spatially Patterned Potentials, *Phys. Rev. Lett.* **114**, 118101 (2015).
- [66] N. G. Berloff, M. Silva, K. Kalinin, A. Askitopoulos, J. D. Töpfer, P. Cilibrizzi, W. Langbein, and P. G. Lagoudakis, Realizing the classical XY hamiltonian in polariton simulators, *Nat. Mater.* **16**, 1120 (2017).
- [67] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, Scaling for edge inference of deep neural networks, *Nat. Electron.* **1**, 216 (2018).
- [68] ITRS (International Technology Roadmap for Semiconductors) 2.0. 2015 beyond CMOS (2015).
- [69] E. Estrecho, T. Gao, N. Bobrovska, D. Comber-Todd, M. D. Fraser, M. Steger, K. West, L. N. Pfeiffer, J. Levinson, M. M. Parish, T. C. H. Liew, M. Matuszewski, D. W. Snoke, A. G. Truscott, and E. A. Ostrovskaya, Direct measurement of polariton-polariton interaction strength in the Thomas-Fermi regime of exciton-polariton condensation, *Phys. Rev. B* **100**, 035306 (2019).
- [70] Y. Sun, P. Wen, Y. Yoon, G. Liu, M. Steger, L. N. Pfeiffer, K. West, D. W. Snoke, and K. A. Nelson, Bose-Einstein Condensation of Long-Lifetime Polaritons in Thermal Equilibrium, *Phys. Rev. Lett.* **118**, 016602 (2017).
- [71] F. Tassone and Y. Yamamoto, Exciton-exciton scattering dynamics in a semiconductor microcavity and stimulated scattering into polaritons, *Phys. Rev. B* **59**, 10830 (1999).
- [72] S. Christopoulos, G. B. H. Von Högersthal, A. Grundy, P. Lagoudakis, A. Kavokin, J. Baumberg, G. Christmann, R. Butté, E. Feltn, and J.-F. Carlin, *et al.*, Room-Temperature Polariton Lasing in Semiconductor Microcavities, *Phys. Rev. Lett.* **98**, 126405 (2007).
- [73] A. Fieramosca, L. Polimeno, V. Ardizzone, L. De Marco, M. Pugliese, V. Maiorano, M. De Giorgi, L. Dominici, G. Gigli, and D. Gerace, *et al.*, Two-dimensional hybrid perovskites sustaining strong polariton interactions at room temperature, *Sci. Adv.* **5**, eaav9967 (2019).
- [74] R. Su, A. Fieramosca, Q. Zhang, H. S. Nguyen, E. Deleporte, Z. Chen, D. Sanvitto, T. C. Liew, and Q. Xiong, Perovskite semiconductors for room-temperature exciton-polaritonics, *Nat. Mater.* **1**, (2021).
- [75] R. Su, S. Ghosh, J. Wang, S. Liu, C. Diederichs, T. C. H. Liew, and Q. Xiong, Observation of exciton polariton condensation in a perovskite lattice at room temperature, *Nat. Phys.* **16**, 301 (2020).
- [76] F. Li, L. Orosz, O. Kamoun, S. Bouchoule, C. Brimont, P. Disseix, T. Guillet, X. Lafosse, M. Leroux, and J. Leymarie, *et al.*, From Excitonic to Photonic Polariton Condensate in a ZnO-Based Microcavity, *Phys. Rev. Lett.* **110**, 196406 (2013).
- [77] G. Muñoz-Matutano, A. Wood, M. Johnsson, X. Vidal, B. Q. Baragiola, A. Reinhard, A. Lemaître, J. Bloch, A. Amo, G. Noguees, B. Besga, M. Richard, and T. Volz, Emergence of quantum correlations from interacting fibre-cavity polaritons, *Nat. Mater.* **18**, 213 (2019).
- [78] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, A million spiking-neuron integrated circuit with a scalable communication network and interface, *Science* **345**, 668 (2014).
- [79] ITRS (International Technology Roadmap for Semiconductors) 2.0. 2015 executive report (2015).
- [80] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016) p. 4107.
- [81] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, in *Computer Vision - ECCV 2016*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling (Springer International Publishing, Cham, 2016), p. 525.
- [82] M. Gruber, J. Jahns, and S. Sinzinger, Planar-integrated optical vector-matrix multiplier, *Appl. Opt.* **39**, 5367 (2000).
- [83] J. Spall, X. Guo, T. D. Barrett, and A. Lvovsky, Fully reconfigurable coherent optical vector-matrix multiplication, *Opt. Lett.* **45**, 5752 (2020).
- [84] K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions, *Nat. Photonics* **13**, 454 (2019).
- [85] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. S. Hosseini, A. Biberman, and M. R. Watts, An ultralow power athermal silicon modulator, *Nat. Commun.* **5**, 1 (2014).
- [86] V. J. Sorger, R. Amin, J. B. Khurgin, Z. Ma, H. Dalir, and S. Khan, Scaling vectors of attojoule per bit modulators, *J. Opt.* **20**, 014012 (2017).
- [87] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, Fully hardware-implemented memristor convolutional neural network, *Nature* **577**, 641 (2020).
- [88] T. Chen, J. van Gelder, B. van de Ven, S. V. Amitonov, B. de Wilde, H.-C. Ruiz Euler, H. Broersma, P. A. Bobbert, F. A. Zwanenburg, and W. G. van der Wiel, Classification with a disordered dopant-atom network in silicon, *Nature* **577**, 341 (2020).

- [89] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**, 2278 (1998).
- [90] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, in *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '14* (Association for Computing Machinery, New York, NY, USA, 2014), p. 269.
- [91] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H. Yoo, Unpu: An energy-efficient deep neural network accelerator with fully variable weight bit precision, *IEEE J. Solid-State Circuits* **54**, 173 (2019).
- [92] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, Equivalent-accuracy accelerated neural-network training using analogue memory, *Nature* **558**, 60 (2018).
- [93] In Ref. [88] 1 SOP = 10 FLOPs conversion ratio was assumed.
- [94] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Photonic multiply-accumulate operations for neural networks, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1 (2019).
- [95] D. A. Miller, Device requirements for optical interconnects to silicon chips, *Proc. IEEE* **97**, 1166 (2009).
- [96] M. Vladimirova, S. Cronenberger, D. Scalbert, K. V. Kavokin, A. Miard, A. Lemaître, J. Bloch, D. Solnyshkov, G. Malpuech, and A. V. Kavokin, Polariton-polariton interaction constants in microcavities, *Phys. Rev. B* **82**, 075301 (2010).
- [97] I. A. Shelykh, R. Johne, D. D. Solnyshkov, and G. Malpuech, Optically and electrically controlled polariton spin transistor, *Phys. Rev. B* **82**, 153303 (2010).
- [98] D. A. B. Miller, Are optical transistors the logical next step? *Nat. Photonics* **4**, 3 (2010).
- [99] E. Wertz, A. Amo, D. D. Solnyshkov, L. Ferrier, T. C. H. Liew, D. Sanvitto, P. Senellart, I. Sagnes, A. Lemaître, A. V. Kavokin, G. Malpuech, and J. Bloch, Propagation and Amplification Dynamics of 1D Polariton Condensates, *Phys. Rev. Lett.* **109**, 216404 (2012).
- [100] A. Delteil, T. Fink, A. Schade, S. Höfling, C. Schneider, and A. Imamoglu, Towards polariton blockade of confined exciton-polaritons, *Nat. Mater.* **18**, 219 (2019).
- [101] J. D. Plumhof, T. Stöferle, L. Mai, U. Scherf, and R. Mahrt, Room-temperature Bose-Einstein condensation of cavity exciton-polaritons in a polymer, *Nat. Mater.* **13**, 247 (2014).
- [102] K. S. Daskalakis, S. A. Maier, R. Murray, and S. Kéna-Cohen, Nonlinear interactions in an organic polariton condensate, *Nat. Mater.* **13**, 271 (2014).
- [103] M. Dusel, S. Betzold, O. A. Egorov, S. Klembt, J. Ohmer, U. Fischer, S. Höfling, and S. C., Room temperature organic exciton-polariton condensate in a lattice, *Nat. Commun.* **11**, 2863 (2020).
- [104] S. Dufferwiel, S. Schwarz, F. Withers, A. A. P. Trichet, F. Li, M. Sich, O. Del Pozo-Zamudio, C. Clark, A. Nalitov, D. D. Solnyshkov, G. Malpuech, K. S. Novoselov, J. M. Smith, M. S. Skolnick, D. N. Krizhanovskii, and A. I. Tartakovskii, Exciton-polaritons in van der Waals heterostructures embedded in tunable microcavities, *Nat. Commun.* **6**, 8579 (2015).
- [105] P. M. Walker, L. Tinkler, B. Royall, D. V. Skryabin, I. Farrer, D. A. Ritchie, M. S. Skolnick, and D. N. Krizhanovskii, Dark Solitons in High Velocity Waveguide Polariton Fluids, *Phys. Rev. Lett.* **119**, 097403 (2017).
- [106] J. Robertson, M. Hejda, J. Bueno, and A. Hurtado, Ultrafast optical integration and pattern classification for neuromorphic photonics based on spiking vcsel neurons, *Sci. Rep.* **10**, 6098 (2020).