Role of all-optical neural networks

M. Matuszewski[1,2,*] A. Prystupiuk[1,3] and A. Opala[1,3]

[1] *The Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, Warsaw PL-02-668, Poland*
[2] *Center for Theoretical Physics, Polish Academy of Sciences, Aleja Lotników 32/46, Warsaw PL-02-668, Poland*
[3] *Institute of Experimental Physics, Faculty of Physics, University of Warsaw, ul. Pasteura 5, Warsaw PL-02-093, Poland*

(Received 7 June 2023; revised 18 September 2023; accepted 4 December 2023; published 17 January 2024)

In light of recent achievements in optical computing and machine learning, we consider the conditions under which all-optical computing may surpass electronic and optoelectronic computing in terms of energy efficiency and scalability. When considering the performance of a system as a whole, the cost of memory access and data acquisition is likely to be one of the main efficiency bottlenecks not only for electronic, but also for optoelectronic and all-optical devices. However, we predict that all-optical devices will be at an advantage in the case of inference in large neural network models, and the advantage will be particularly large in the case of generative models. We also consider the limitations of all-optical neural networks, including footprint, strength of nonlinearity, optical signal degradation, limited precision of computations, and quantum noise.

DOI: 10.1103/PhysRevApplied.21.014028

I. INTRODUCTION

In recent years, remarkable strides have been made in the field of machine learning and artificial intelligence, heralding a new era of practical applications that are swiftly permeating various industries and our daily lives. However, this remarkable progress comes at the price of the rise in energy consumption, which is primarily driven by the exponential growth in the volume of data being processed [1] and to the apparent flattening of the improvement of computing performance. For many decades, progress was governed by remarkable principles of Moore's law and Dennard scaling. However, it seems that we are now entering a phase where these principles are gradually approaching a more stable plateau [2,3]. It has become apparent that the cost of data movement through electronic wires, requiring charging their capacity for each bit of information, dominates the energy budget for data-intensive applications such as large-scale machine learning [4,5].

To circumvent this limitation and make computations more efficient, a natural strategy is to reduce the physical distance between memory and processing units. This motivates interest in computing systems that go beyond the von Neumann architecture, such as in-memory computing [4]. Many physical implementations of machine learning have been realized with emulators of neural networks on specialized hardware, where the structure of the network is resembled physically [6–9]. In these cases, computing is often analog rather than digital. Since neural network models are themselves analog, this approach appears to be more adequate in the case of neural networks than in the case of traditional algorithm-based computing.

Another way to increase the efficiency of computations is to realize them on a non-standard physical platform [10–12]. A particularly promising approach is to use photons instead of electrons [12–23]. The advantage of optical systems is that they do not require charging the capacitance of communication channels, so data movement can be almost lossless. For this reason, optical systems are used for communications over large distances and at high data rates, when the energy cost of data movement is particularly important [24]. However, computation with photons, while being researched for many decades [25], has not found mainstream applications yet. Optical computing has been hampered by many factors, including the weakness of optical nonlinearity, bulkiness of optical elements, and the difficulty to regenerate optical signals and to integrate optical sources. It has been difficult to realize a device capable of realizing general digital operations with appropriate fidelity and at a low energy cost [26]. However, the recent resurgence of interest in optical computing led to breakthroughs that alleviated many of these limitations [14,15], and it seems that taking advantage of optical computing in practical devices is within reach.

A natural solution to overcome the disadvantages of electronic and optical computing is to combine them in a single system, using the advantages of both light

*mmatuszewski@cft.edu.pl

and matter. This approach typically assumes constructing an optoelectronic device where communication or linear operations are realized optically, while other operations, including nonlinear transformations and signal regeneration, FAN-IN, and FAN-OUT are taken care of by electronics. While this approach is very promising, it has its own limitations. One of them is the limited compatibility of electronic and optical systems and the difficulty to integrate the two. For example, typical length scales for state-of-the-art electronics are in the range of a few nanometers. In the case of optics, it is difficult to squeeze light below the micrometer length scale without incurring significant losses. On the other hand, if conversion from electronic to optical signals and vice versa and from analog to digital signals occur in the device, they may create additional energy costs and technological difficulties.

Here, we consider the viability of all-optical computing as an alternative to electronic and optoelectronic approaches, and attempt to identify the applications where it may excel. This topic has been considered previously, and it was pointed out that an all-optical approach may not be well suited for digital computations [26]. We look at this problem from an alternative perspective, taking into account recent advancements both in optical technology and in the field of machine learning. We assume that the main technological bottleneck of high-intensity computing is, as it appears, the efficiency of data movement via electronic wires. This is justified both by fundamental physical limitations, and by the observation that electronic computation efficiency is apparently saturating after decades of exponential progress. On the other hand, there is still room for improvement before the fundamental limits are reached for optics. Consequently, we consider the energy cost of operations requiring data movement by electronic channels, such as memory access, to be the most stringent limitation. Importantly, to provide a fair comparison, we consider the efficiency of the complete computing system, and not only a certain part of it. In particular, we take into account the cost of data acquisition, and the electronic memory access cost necessary to provide data, which is often overlooked in estimates of energy efficiency. It is important to mention that our considerations do not apply to the case where input data can be supplied in the form of optical signals, without accessing external electronic memory [27,28].

Based on these assumptions, we try to answer the question of the viability and practicality of all-optical neural networks. In other words, we consider whether there are applications in which all-optical neural networks can outperform their electronic and optoelectronic counterparts, and to what extent. We conclude that electronic memory access cost will likely be the main limitation not only for electronic and optoelectronic, but also for all-optical networks. However, we find that one application where all-optical networks will be at an advantage is the inference in large-scale neural networks, where the number of neurons in the hidden layer is much larger than the dimensionality of inputs and outputs. The advantage will be particularly large in the case of generative models, where input data can be reused in many subsequent inferences, reducing data acquisition costs. These conditions are fulfilled in many machine learning models used in practice.

In addition, we consider the limitations of all-optical neural networks that must be overcome before they become practical. We analyze optical neural networks, taking into account the specifics of information processing with light, such as the quantization of light. By performing numerical simulations, we show that all-optical neural networks can be accurate even if the precision of optical transformations is reduced by noise and fabrication errors. We discuss the issues of signal regeneration, network depth, and scalability of optical networks.

II. CAN ALL-OPTICAL NEURAL NETWORKS BE EFFICIENT?

In this section, we show the main motivation of our paper, that is, the advantage of using all-optical systems as an efficient platform for analog neural networks, as opposed to electronic or optoelectronic devices. We consider the energy cost of calculations, which is currently the most important limitation of computing systems [2,4]. We leave the considerations of footprint, speed, precision, and other limitations of optical systems to the next section.

The main assumption of our considerations is that the data movement cost in electronic wires will be difficult to improve in the future. This assumption can be justified by two arguments. One is the physical lower limit of the energy required to charge an electronic wire to send a single bit of information. The cost of charging wire capacitance per unit length is approximately independent of the wire cross section, and can be estimated as 100 fJ/bit per 1 mm of connection length [5]. Another argument for considering data movement cost as a physical lower limit is the apparent saturation of energy efficiency of computations and memory access costs [2–4], despite decades-long developments and huge investments in the complementary metal-oxide (CMOS) technology. In fact, it appears that state-of-the-art efficiencies are reaching the fundamental estimates. In machine learning applications, which require a great number of memory access operations to perform multiplications of large tensors, the cost of data movement is at least comparable to the cost of logic operations [4]. Therefore, it seems that room for improvement in the efficiency of current CMOS technology is limited, unless a significant technological breakthrough is achieved.

How can optics be advantageous from the perspective of hardware-implemented neural networks? As was mentioned, optical links do not require charging of communication lines. Optical energy dissipation corresponds to

effects such as optical absorption, light leakage in waveguides, optical dispersion, and spatial- or temporal-mode misalignment; however, these effects typically lead to a much lower dissipation than in the case of electronics. This is the reason why optical connections are used for long-haul communications, in data centers, and can be applied for communications even at short length scales [29].

In this work, we focus on the aspects specific to neural networks. The structure of neural networks and the specifics of the required computations make analog systems a much better match than in the case of the digital von Neumann architecture. In this context, one can point out several advantages that we list below.

A. Optical FAN-OUT and FAN-IN

In a typical artificial neural network, neurons perform two kinds of operations. Summation of neuron inputs x_i multiplied by weights w_{ij} is a linear operation (i.e., linear as a function of neuron inputs), which is followed by a nonlinear activation given by a function f_j such as the sigmoid or rectified linear unit (ReLU)

$$y_j = f_j\left(\sum_i^N w_{ij}x_i\right). \tag{1}$$

Note that f_j can act on vectors rather than scalars, as is the case in the softmax function. We refer to a recent review [19] for a discussion of other types of neural operations, including spiking neural networks. In the case of electronic systems, the efficiency is strongly tied to the cost of a single multiply and accumulate (MAC) operation. This operation occurs once for every multiplication of a neuron input x_i with the corresponding input (synaptic) weight w_{ij}. Performing such an operation in the von Neumann machine requires accessing memory for all the inputs and all the corresponding weights. If the number of neuron inputs is large, so is the required data movement. On the other hand, the nonlinear activation function is applied only once per neuron activation, and its energy cost can be much lower. In practice, input-weight multiplications are performed in batches, so the weights can be, to a large extent, reused if stored in a local memory. However, due to the scale of large machine learning models, the memory access cost is still a major source of energy dissipation.

On the other hand, in the case of optical neural networks, the summation of optical signals can be performed at almost no cost of data movement by directing or focusing optical pulses or beams to certain regions in space, which is the optical FAN-IN. This can take the form of either simple intensity addition in the case of mutually incoherent light pulses, or optical interference in the case of mutually coherent pulses. On the other hand, FAN-OUT of output optical signals to a very large number of copies can be realized with linear optical elements such as diffractive optical elements [30], spatial light modulators, microlens arrays [28], or in integrated circuits [31]. There is no fundamental lower limit for the energy cost of these operations. Therefore, it is the nonlinear activation function, rather than the weighted linear summation, that creates a bottleneck for the highest possible efficiency of all-optical devices. This is a particularly important limitation due to the weakness of nonlinear interactions between photons, which are much weaker than the interactions between electrons in semiconductors.

B. Static weights in neural network inference

In machine learning, the inference phase follows the training phase. From the point of view of energy consumption, the inference phase is often more important than the training phase since the trained model can be used for inference arbitrarily many times [32]. Specialized CMOS inference systems include Google TPUv1 and the nVidia inference platform. In the inference phase, the weights of neurons do not change. Therefore, if weights can be implemented in optical hardware without the need for external memory access, the cost of data movement can be greatly reduced. In the case of electronics, this approach is the basis of in-memory computing [4]. However, electronic chips with in-memory computing can require complicated wiring to connect computing units with each other [7]. There is also strong interest in analog electronic architectures, which can improve energy efficiency by 1 to 2 orders of magnitude [33]. In the case of optics, hardware-encoded static weights can be combined with almost dissipationless transport to drastically reduce the cost of weighted summation in Eq. (1). The multiplication of inputs by the corresponding weights can be implemented with linear optical elements that apply a certain amplitude or phase modulation to the optical signals. One of the widely used methods is implementing Mach-Zehnder interferometers, or a mesh of such interferometers that perform an arbitrary linear operation represented by a matrix [31]. Weights encoded in phase-change materials [34] do not require power to sustain their state. In the case of free-space propagating beams, spatial light modulators can be used for applying weights, with millions of independently tunable parameters [30].

C. Structure of large neural network models

One of the reasons for the recent progress in machine learning is that hardware, such as specialized parallel computing units, allowed the implementation of models with increasing complexity that could accommodate and process large datasets. Usually, to obtain high accuracy of predictions, it is necessary to use models with a large number of parameters and neurons. It was noticed that optics could be particularly effective in comparison to electronics in applications that require large-scale computations

[18,35]. This advantage is particularly large in the case of all-optical networks. Let us, for the moment, focus on the memory access cost as the main bottleneck and compare the potential efficiency of all-optical and optoelectronic devices (we justify this assumption later). Consider an optoelectronic device that stores the result of computation in digital memory after computation in each layer. In this case, data need to be converted to a digital signal and stored in memory after each layer of neural network computation, while in an all-optical system this is not necessary; see Fig. 1. We assume that the input to an all-optical device is provided in an electronic form since most of the data in our world are encoded electronically. They need to be read from a digital memory both in the case of an optoelectronic and an all-optical neural network. However, the advantage of an all-optical network is that once converted to the optical domain, they do not need to be reconverted to the electronic, digital domain until the entire computation on the data sample is finished. In Fig. 1, memory access cost occurs only at the input and the output of the all-optical network and at each hidden neuron of an optoelectronic network. Moreover, in the cases where data are provided in an optical form, they do not occur at the input layer [27,28].

An alternative solution is one where an optoelectronic device converts between optical and electronic domains in each layer, but keeps the analog form of the signal throughout the computation. This kind of device will avoid memory access cost, while optoelectronic conversion cost will still occur. In many respects, it will have the same advantages and disadvantages as all-optical devices.

D. Comparison of optoelectronic and all-optical network efficiencies

How does this lower memory access cost affect the efficiency of optical networks? This largely depends on the particular structure of the neural network model, and we analyze some examples here. Generally, in the case of large-scale neural networks, the size of hidden layers in Fig. 1, measured as the number of neurons or the number of parameters, is much larger than the size of input and output layers [36,37]. The energy cost of computations per sample in the inference stage can be very roughly estimated for an optoelectronic network as

$$E_{\mathrm{OE}} \geq E_{\mathrm{electronic}}(N_{\mathrm{input}} + N_{\mathrm{output}} + N_{\mathrm{hidden}}), \qquad (2)$$

and for an all-optical network as

$$E_{\mathrm{AO}} \geq E_{\mathrm{electronic}}(N_{\mathrm{input}} + N_{\mathrm{output}}) + E_{\mathrm{optical}}N_{\mathrm{hidden}}, \qquad (3)$$

where $E_{\mathrm{electronic}}$ is the average cost of electronic operations per neuron per inference, including memory access cost, and E_{optical} is the average cost of optical operations per neuron per inference, including the energy of both input and pump pulses, optical losses, and electronics necessary for an optical neuron to operate. Under the assumption that the memory access cost is the main bottleneck of computation efficiency, and considering the case where the majority of neurons are in the hidden layer, $N_{\mathrm{hidden}} \gg N_{\mathrm{input}} + N_{\mathrm{output}}$, the ratio of energy costs can be estimated as

$$\begin{aligned}\frac{E_{\mathrm{AO}}}{E_{\mathrm{OE}}} &= \frac{N_{\mathrm{input}} + N_{\mathrm{output}}}{N_{\mathrm{input}} + N_{\mathrm{output}} + N_{\mathrm{hidden}}} \\ &\quad + \frac{E_{\mathrm{optical}}}{E_{\mathrm{electronic}}}\frac{N_{\mathrm{hidden}}}{N_{\mathrm{input}} + N_{\mathrm{output}} + N_{\mathrm{hidden}}} \\ &\approx \frac{E_{\mathrm{optical}}}{E_{\mathrm{electronic}}}.\end{aligned}$$

If $E_{\mathrm{electronic}} \gg E_{\mathrm{optical}}$, i.e., the energy cost of computation per inference per neuron in the optoelectronic network is much larger than the cost of the same operation performed all-optically, the energy cost will be much lower in the case of an all-optical device.

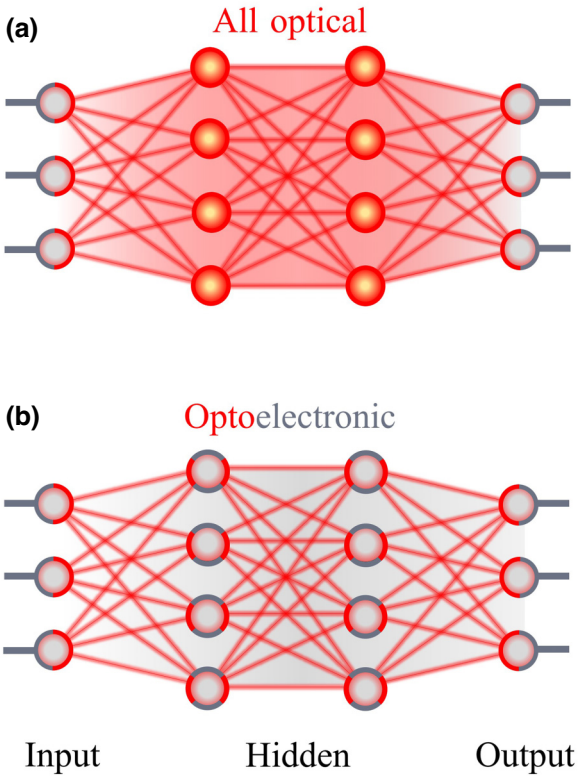

FIG. 1. All-optical and optoelectronic neural networks. (a) In an all-optical network, input data are transformed to optical form at the input layer, and all subsequent operations up to the output layer are realized all-optically. (b) In an optoelectronic network, the signal is transformed from optical to electronic and back at each layer of the network. If the size of the hidden (middle) layers is much larger than input and output layers, this results in a bottleneck of system efficiency.

To justify the above reasoning, we consider whether the above conditions, $N_{\text{hidden}} \gg N_{\text{input}} + N_{\text{output}}$ and $E_{\text{electronic}} \gg E_{\text{optical}}$, can be fulfilled in practice. To estimate the average cost of electronic operations $E_{\text{electronic}}$ in an optoelectronic device, one needs to take into account the costs of conversion from an optical signal to an electronic signal, analog-to-digital conversion, the costs of the reverse processes, and the cost of memory access. In certain optoelectronic devices, some of these costs may be absent, for example, if the electronic part of the computation is analog as well. Several recent realizations of optoelectronic neural networks went in this direction [13,38–40], achieving very low energy cost per operation. However, it appears to be difficult to reduce the cost of all these operations below the level of picojoules per data unit, such as a byte. In particular, the cost of memory access appears to be the main bottleneck. For an 8-bit input, it ranges from several picojoules to several nanojoules, depending on the technology used and the size of the memory. For example, single access to 100-MB memory requires around 10 pJ of energy [3,4]. Even if the system is designed in such a way that access to memory is not required for each neuron operation, the cost of analog-to-digital conversion and electronic-to-optical conversion results in a bottleneck [39,41,42]. It appears that to achieve the highest possible efficiency, both the cost of memory access and analog-digital conversion have to be avoided. In this case, the optoelectronic conversion cost may be reduced to the level of tens of femtojoules per nonlinear operation [5]. This may require a design with "near-receiverless," low-capacitance photodetectors.

On the other hand, all-optical neural networks in the inference mode do not require optoelectronic conversion and the energy E_{optical} is mainly bounded by the required power of the light source. This bound depends on the optical nonlinearity of the system, optical losses, and the sensitivity of detectors at the output layer. While weak nonlinear response is one of the main disadvantages of optical systems, the use of strong light-matter coupling characterized by ultrafast nonlinearities can result in high efficiency of nonlinear operations at high data rates [43]. In particular, exciton-polaritons in both inorganic and organic materials or two-dimensional materials can exhibit optical nonlinearity that is orders of magnitude stronger than in other materials [44,45]. Using exciton-polaritons, the energy cost of a single nonlinear operation can be as low as a few attojoules per neuron [43]. At the same time, since at the output layer light is collected from all hidden neurons, the average light intensity reaching an output detector scales proportionally to $N_{\text{hidden}}/N_{\text{output}}$. Accordingly, in the limit of large $N_{\text{hidden}}/N_{\text{output}}$ considered here, detector sensitivity will not be the main bottleneck.

The large size of a neural network model translates to a large size of hidden layers N_{hidden}. As a result, large-scale neural networks used in practice are usually characterized by very high ratios $N_{\text{hidden}}/N_{\text{input}}$. For example, one of the leading models in the ImageNet competition, AmoebaNet, has 10^9 hidden nodes and performs 10^{11} operations per inference, while the ImageNet input size is $165 \times 165 \times 3 \approx 10^5$, resulting in $N_{\text{hidden}}/N_{\text{input}} \approx 10^4$. In recent language models, this ratio is even higher, with the large BERT model consisting of approximately $24 \times 2 \times 512 \times 1024$ nonlinear nodes for a 512-long input token sequence, resulting in the ratio $N_{\text{hidden}}/N_{\text{input}} \approx 10^5$. Therefore, the condition $N_{\text{hidden}} \gg N_{\text{input}} + N_{\text{output}}$ is fulfilled in many practical large neural network models. It is interesting that the same relation appears to hold for the network of neurons in the human brain. The number of neurons in the brain is of the order of 10^{11}, which is likely to be much greater than the dimensionality of the input information from all stimuli [46].

To give a concrete example of potential efficiency, we estimate the energy cost per synaptic operation for a hypothetical large-scale neural network with $N_{\text{input}} + N_{\text{output}} = 10^3$ and $N_{\text{hidden}} = 10^8$, assuming that $E_{\text{electronic}} = 1$ pJ and $E_{\text{optical}} = 100$ aJ. In both cases, we assume the optical FAN-IN of 1000 inputs per neuron in the hidden layer. For a fair comparison, the energy cost per operation is calculated as the total energy cost of the complete network, including memory access for each electronic neuron operation. The number of operations is calculated as two operations (multiply and accumulate) per neuron input and one for nonlinear activation in a neuron. According to Eqs. (2) and (3), the lower bounds for the energy cost are estimated to be 500 aJ for an optoelectronic network and 55 zJ for an all-optical network, almost 4 orders of magnitude lower.

These estimations are not complete unless we consider the cost of acquiring data. This includes the cost of access to external memory, such as DRAM memory, from where input data have to be retrieved. In the case of data that need to be transmitted over a distance, as is usually the case in cloud computing, the cost of accessing input data may be further increased. The costs of both reading from DRAM memory and fiber link data transmission are in the range of 1–100 pJ per bit, or up to 1 nJ per byte [4,47]. These costs may be significantly higher in the case of wireless communication or less efficient data transmission channels. The cost of acquiring input data may also be high in the case of edge computing, for example, if input information is gathered by a camera with a high energy cost per pixel. In all of these cases, the input data costs may dominate over all other costs of computations.

However, there is an important class of practical machine learning tasks where input data costs can be drastically reduced by "recycling" input data acquisition. All generative tasks in which most of the input information used at one step of computation can be used for the next step belong to this category. These include applications such as text completion, language translation, question answering, chatbots, image and sound synthesis. In these

TABLE I. Estimates of average energy cost per synaptic operation in inference for the system as a whole, including the data acquisition cost. Parameters are $E_{\text{electronic}} = 2$ pJ, $E_{\text{optical}} = 100$ aJ, $E_{\text{memory}} = 10$ pJ, $E_{\text{acquisition}} = 100$ pJ. We assume 10^3 inputs per neuron on average and the possibility to reuse input data $M = 10^3$ times in a generative network ($M = 1$ in other cases). For simplicity, we assume that $N_{\text{input}} \gg N_{\text{output}}$.

Network type	Small	Large	Large generative
$N_{\text{input}} + N_{\text{output}}$	100	1000	1000
N_{hidden}	1000	10^8	10^8
Electronic	1 pJ	1 pJ	1 pJ
Optoelectronic	7 fJ	1 fJ	1 fJ
All optical	6 fJ	600 zJ	100 zJ

cases, input information known as "context" can often be reused to a great extent across inferences, for example, 4×10^3 times in the case of large language models. As a result, the cost of data acquisition may be orders of magnitude smaller than the cost of local memory access for input neurons, which is already included in our estimations.

In Table I, we present examples of complete estimates for electronic, optoelectronic, and all-optical neural networks in the case of various machine learning model sizes. We take into account the costs of all contributions to energy usage, including optical, optoelectronic, memory, and data acquisition. To this end, in the calculation of energy cost per inference, we include additional terms in addition to those present in Eqs. (2) and (3):

$$E^{\text{total}}_{\text{OE,AO}} = E_{\text{OE,AO}} + N_{\text{input}}\left(\frac{E_{\text{acquisition}}}{M} + E_{\text{memory}}\right). \quad (4)$$

Here $E_{\text{acquisition}}$ is the cost of acquiring input data, which is divided by the number of inferences M where it is reused, and E_{memory} is the cost of accessing local memory that stores the input data. It is clear from Table I that all-optical neural networks will have the advantage in the case of large models, and in particular in the case of generative models that require less input data. Our estimates for optoelectronic devices are in line with recent analysis [48], which predicted analogous efficiency for similar-scale optical transformers with state-of-the-art electronics.

III. LIMITATIONS OF ALL-OPTICAL COMPUTING

In this section, we discuss the possible limitations of all-optical computing. We consider the footprint and speed of operations, cascadability and signal degradation, implementing useful nonlinear transformations, precision of computations, fabrication errors, and quantum noise.

A. Footprint

One of the arguments commonly raised against using optics for computing is the footprint of optical systems. The rationale of this argument is that the wavelength of visible light is of the order of a single micrometer, while electronic systems can be integrated in chips with nanometer-sized transistors. While the footprint is certainly a limitation for optics, it is actually not as severe as it may seem. In the case of neural network implementations in electronics, the nanometer size of a transistor does not directly translate into nanometer-sized neurons. For example, in the IBM TrueNorth and Intel Loihi neuromorphic chips [7], fabricated in 28-nm process technology, the footprint is respectively approximately 200 μm^2 per neuron and 1 μm^2 per synaptic weight. In the recent Intel Loihi 2 chip, a footprint of 0.3 μm^2 per synaptic weight was achieved [8]. Such length scales result from the complicated circuitry that must be implemented in an electronic chip to emulate a neuron.

Moreover, it is important to realize that there exists a direct relation between the energy efficiency of a chip and its footprint, which results from the need to dissipate heat generated by the computation. Heat removal is space consuming. In CMOS chips, the circuit structure is usually two dimensional, and the third dimension is sacrificed for a heat sink. One exception to this rule are memory chips, which often have a multilayer stacked structure with more than 100 layers. This is possible due to the reduced amount of heat generation as compared to information processing chips. As a result, the reduction of energy dissipation leads to the reduction of the footprint.

Moreover, there have been great advancements in the miniaturization of optical systems. Integrated silicon photonics chips can be fabricated and processed in large quantities by specialized foundries. A typical size of an element of a photonic chip, such as a Mach-Zehnder interferometer, is of the order of micrometers to hundreds of micrometers [31,34,38]. If energy dissipation in such optical chips is lower than dissipation in electronic chips, stacking of optical chip layers should be possible, thus reducing the footprint. On the other hand, the free-space approach to computing, while requiring the third dimension for light propagation, also permits achieving a very low footprint. For example, commercially available spatial light modulators with a few micrometer pixel pitches are able to encode synaptic weight information with a density comparable to the density in electronic chips. It could be further increased if holographic encoding of some form was used. Assuming a conservative estimate of encoding a single weight parameter on 10 μm^2 of surface, the size of a weight bank encoding the full BERT language model with 110 million parameters would require a surface of only 11 cm^2.

A fundamental limit for thickness for a given number of optical connections results from the consideration of diffraction [49]. For example, connection of N optical neurons to M optical neurons requires a dividing surface containing of the order of $N \times M$ optical modes. It can be deduced that the thickness of a single optical layer is of the

same order as the linear size of the weight bank encoding $N \times M$ connections.

B. Speed

The speed of a neural network inference can be measured in several ways. While the number of operations per second is a valid measure of computational power, probably more important ones from the practical point of view are latency and performance density, which is the number of operations per second per area [3]. In terms of latency, all-optical networks can certainly outperform electronic and optoelectronic networks in most applications, since apart from input generation and output detection, they require only propagation of light across the network layers at the speed of light. For a centimeter-sized system, this results in a latency of the order of picoseconds, which is many orders of magnitude lower than millisecond latency typical for electronics [32]. Performance density of all-optical networks can be estimated by considering the number of synaptic weight multiplications per second, taking into account the 10-μm^2 weight footprint as estimated above and the 10-GHz inference rate corresponding to commercial optical modulators. The resulting performance density of the order of 10^6 GOP s^{-1} mm^{-2} is 3 orders of magnitude higher than in state-of-the-art standard electronics [3], with memristor-based electronic devices being a promising platform to increase these figures [50,51].

C. Strength of nonlinearity

An efficient all-optical neural network requires strong optical nonlinearity. This nonlinearity has to be characterized by a fast response time, ideally in the gigahertz to terahertz range, since the optical pulse energy required for realizing an operation scales linearly with its duration at the same light intensity. A range of optical materials has been considered for this purpose [34,52–54]. In particular, semiconductor materials possess characteristics that make them good candidates for nonlinear elements of artificial neurons [52,53,55].

In this context, microcavity exciton-polaritons [16,17,43,56] are a particularly promising alternative. These quasiparticles are half-light, half-matter excitations existing in semiconductors, which induce optical nonlinearity orders of magnitude stronger than in standard semiconductor materials [43]. They can operate at room temperature [57–59] and have response times of hundreds of femtoseconds to nanoseconds [60]. Moreover, the nonlinearity of polaritons is significantly enhanced in two-dimensional materials [61], in the case of trion-polaritons [62], or Rydberg polaritons [63].

D. Activation functions

Electronic implementations of neural networks make it possible to realize virtually any nonlinear activation function at a very low energy cost. In all-optical networks, one does not have such flexibility and usually has to deal with a nonlinear response of the system that is either fixed or exhibits some limited tunability. Moreover, it is usually not possible to realize the activation function that is optimal for a particular network model. However, as is well known in the field of machine learning, the particular form of activation function often has only limited impact on system performance. On the other hand, the use of real and imaginary parts of the complex optical field amplitude may lead to certain improvements in accuracy [64]. We checked the impact of the type of activation function and the complex nature of the light field by considering a simple example of a feed-forward neural network with one hidden layer performing the MNIST handwritten digit recognition task. We consider the following real and complex activation functions:

$$f_j(x) = \begin{cases} \mathrm{ReLU}(x), \\ \dfrac{1}{1+e^{-x}}, \end{cases} \tag{5}$$

$$f_j(z) = \begin{cases} |z|, \\ \dfrac{1}{1+e^{-|z|}}. \end{cases} \tag{6}$$

The first two functions are standard activation functions used in machine learning. Note that in the case of the functions in Eq. (6), which take complex arguments, both the inputs x_i and the weights w_{ij} in Eq. (1) can be complex valued, which reflects the amplitude and phase of the optical field. The first function in Eq. (6) is one of the simplest nonlinear functions that takes advantage of the complexity of variables. To justify the form of the second function in Eq. (6), we consider the simple optical setup shown in Fig. 2. This setup is based on the nonlinear refractive index change induced by both the optical control beams and the signal beam, which is transmitted through the nonlinear medium. The optical nonlinearity can be enhanced by enclosing the medium in a microcavity and achieving strong light-matter coupling [43]. At the point of the optical bistability threshold [Fig. 2(b), red line], the dependence of the transmitted signal intensity on the incident light amplitude is strongly nonlinear. It exhibits an S shape analogous to the sigmoid activation function known from machine learning models. We assume that the control beams and the input beam are not coherent (e.g., formed by different lasers), which allows us to discard the effects of interference. On the other hand, control beams are assumed to be coherent with each other. This assumption is natural if coherent light is used in the linear vector-matrix multiplication setup, which precedes the nonlinear activation stage [25,31]. The phase of the transmitted beam is therefore not related to the phases of control beams, but its intensity is

strongly modulated by the intensity of the superposition of control beams. Here, we assume a simplified, complex sigmoid dependence of the transmitted light intensity at the threshold

$$I_{\text{out}} \sim I_{\text{pump}} \frac{1}{1 + e^{-|z|}}, \tag{7}$$

where $z = \sum A_i$ is the sum of complex amplitudes of all of the control beams corresponding to this nonlinear node. These beams can be treated as synaptic inputs to the nonlinear node. Therefore, we consider the situation where the intensity of the pump beam is tuned to the middle of the sigmoidal dependence near the optical bistability threshold; see the red line in Fig. 2(b).

In Fig. 3, we present the estimated accuracy of fully connected complex-valued and real-valued neural network models with different nonlinear activation functions. A detailed description of the neural network structure and parameters can be found in the Appendix. Two conclusions can be drawn from these results. First, complex-valued networks can perform slightly better than real-valued networks with the corresponding activation functions and the same number of parameters. This is the case even if biases are not used in the complex-valued networks, which simplifies the implementation with optics. Second, the particular form of activation functions can have some influence on the accuracy, but there is no substantial difference between "optimal" functions such as the rectified linear unit function and the sigmoid. In particular, the physically relevant complex sigmoid activation controlled by complex-valued inputs gives optimal results.

E. Precision

Limited precision of analog systems is a potential obstacle for applications. In the context of neural networks, it is known that low precision can be good enough to perform machine learning tasks with very high accuracy as long as it is kept above a certain, task-dependent level. Examples include quantized and binarized neural networks [65,66]. In the context of analog computing, there are examples where 2–6-bit precision is sufficient to achieve accuracy close to an optimal one [7,66,67]. It appears that the required precision of computations strongly depends on the task to be solved and must be considered on a case-by-case basis.

F. Fabrication errors

The influence of fabrication variability can strongly impact system performance. Ideally, robustness would mean that a neural network model trained *in silico* can perform equally well in a physical system where the parameters are not fully controllable. This can be achieved by reducing device variability, but it is not always possible.

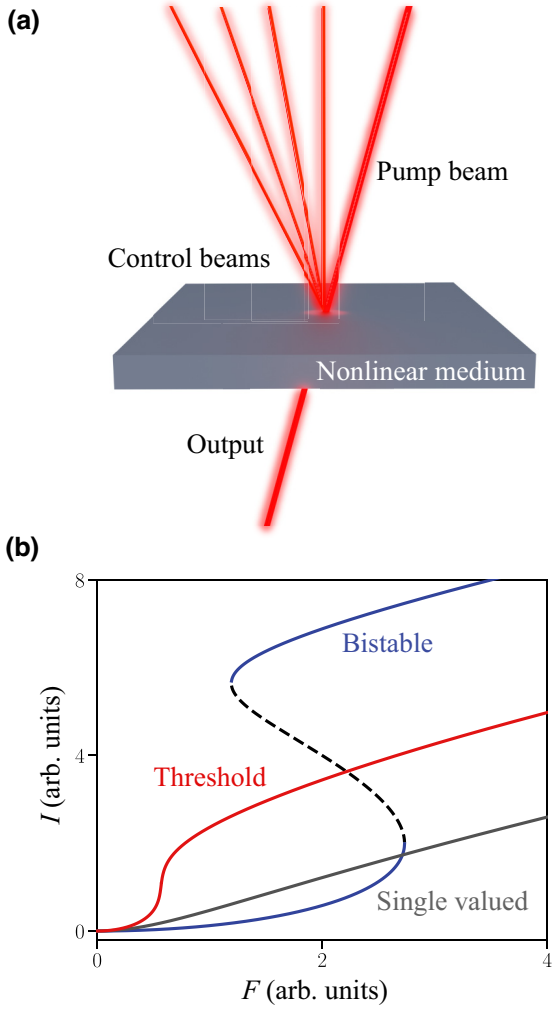

FIG. 2. A possible realization of an optical neuron. (a) A nonlinear system is tuned close to the bistability threshold, which results in a sigmoidlike response to the total intensity of control beams. The transmittance of the strong pump beam depends on the nonlinear index change induced by weak control beams. As a result, both nonlinear activation and signal amplification can be realized. (b) Schematic examples of output intensity I, as a function of total incident amplitude F in the bistable, threshold, and single-valued cases.

However, additional postprocessing correction methods or tunable "control knobs" may be used to adjust the system. For example, in the scheme shown in Fig. 2, such knobs are the phase of the pump beam, and the weights of the linear vector-matrix multiplication, which can correct for the variability of the nonlinear response of the sample. Another method is to fine-tune a pretrained model, taking into account the imperfections existing in a particular device, or specific training methods that directly take into account the response of the physical system [68,69]. If the system is to be used many times in the inference phase,

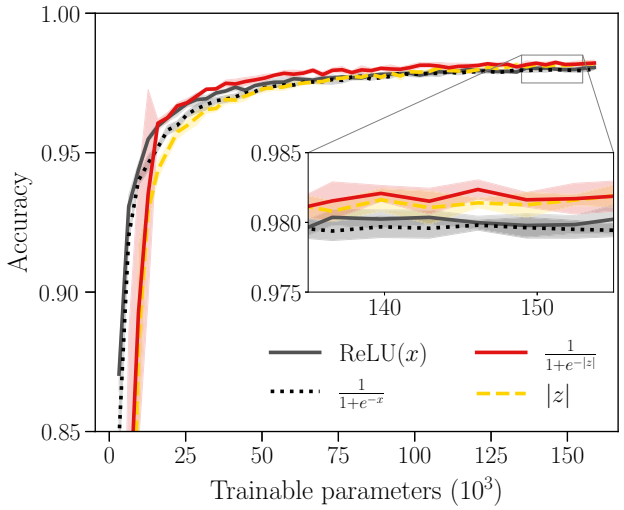

FIG. 3. Accuracy of handwritten digit recognition for fully connected complex- and real-valued networks with a single hidden layer as a function of the number of trainable parameters. Complex-valued networks with the same number of parameters as real-valued ones have roughly half as many neurons. Lines correspond to optimal accuracy of electronic (gray and black) and optoelectronic and all-optical (red and yellow) neural networks after 50–150 epochs of training, depending on the activation function. Shaded regions correspond to the estimated uncertainty based on multiple training iterations. Neurons in the hidden layer with complex activation functions were modeled without biases due to the possible difficulty of their optical implementation, which did not significantly degrade their accuracy.

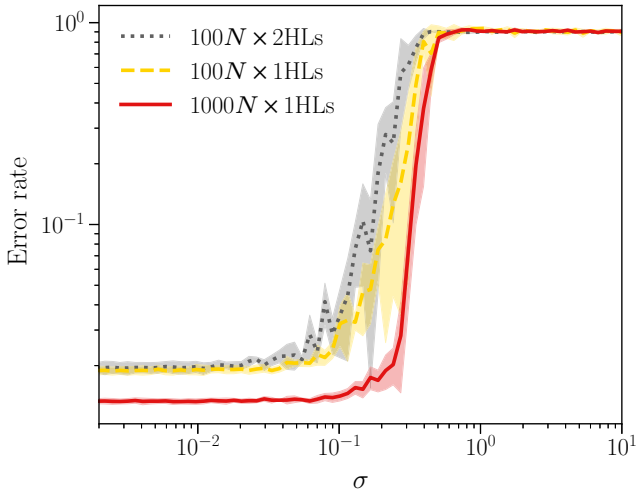

FIG. 4. Influence of imperfections on the performance of optical neural networks. A static disorder of relative amplitude σ is applied to each neuron, as described in Eq. (8). Networks with one and two hidden layers (HLs) are considered, with 100 neurons ($100N$) or 1000 neurons ($1000N$) in each hidden layer. The error rate is defined here as $1-\eta$, where η is the accuracy of the model.

performing such procedures once for each device may be reasonable, even if they are lengthy or expensive.

Apart from these correction methods, neural networks are characterized by an intrinsic robustness to imperfections. To investigate the robustness of optical networks, we analyze the accuracy of the neural network used for the MNIST dataset classification in the case when an additional disorder is introduced to individual neurons. In Fig. 4, we show the accuracy in the function of disorder strength, which is perturbing the response of hidden neurons only in the inference phase, according to

$$f_j(z) = \frac{a_j}{b_j + c_j e^{-|z|}}, \tag{8}$$

where a_j, b_j, c_j are parameters chosen individually for each neuron from the distributions described by a Gaussian probability density centered around unity, i.e., $p(x) = (1/\sqrt{2\pi\sigma^2})e^{-[(x-1)/\sigma]^2/2}$, where $p(x)$ is the probability density of a_j, b_j, c_j taking the value of x and σ is the factor describing the width of the distribution and thus the strength of disorder. The results shown in Fig. 4 indicate that, up to a certain disorder strength, the accuracy of the network inference does not suffer significantly. Robustness certainly increases with the number of neurons, but it decreases with the number of layers, which can be interpreted as the propagation of errors. This shows that even in the case when correction of device imperfections is not possible, reducing the disorder below a certain level may be sufficient.

G. Quantum noise

One of the benefits of all-optical computing is that, in principle, thermal noise can be avoided at the intermediate layers of computation, which is inevitably present in electronic elements whenever optoelectronic conversion occurs. This is true as long as optical signals created by laser beams are well described by a nonthermal coherent state and secondary thermal effects, associated, e.g., with thermal fluctuations in optical elements, are negligible. The effects of thermal noise and its influence on the power efficiency of neural networks were investigated in detail in Refs. [35,41]. The fundamental limit of energy efficiency of all-optical computing is related to quantum noise, or shot noise, which becomes significant near the single-photon level. In the context of optoelectronic networks, it was shown both theoretically [18] and experimentally [70,71] that vector-matrix multiplication operations in neural networks can be performed, even below the single photon per operation level. The reason for this surprising result is that, when many such operations contribute to the result of the weighted summation in a single neuron as

in Eq. (1), the signal-to-noise ratio of the sum is approximately a factor of $\sqrt{N}$ higher than the ratio for individual elements of the sum. This property can greatly increase the energy efficiency of neural networks if the number of neuron inputs is large, which is the case in many practical neural network models.

We analyze to what extent all-optical neural networks, where information is encoded with coherent light amplitudes, can benefit from a similar quantum noise reduction. We assume that the weighted input of an optical neuron j, $w_{ij}x_i$ in Eq. (1). is encoded by a coherent optical laser beam of amplitude proportional to $w_{ij}x_i$. Recall that both the weights w_{ij} and the inputs x_i are complex valued. A superposition of such beams results in optical amplitude proportional to the weighted summation as in Eq. (1) for a given neuron j. In the following, we focus on a particular neuron and drop index j for convenience.

In our quantum treatment, weighted inputs are optical laser pulses represented by coherent photon states $|\alpha_i\rangle$ such that $\alpha_i = \beta w_i x_i$, where β is a factor that relates the coherent state amplitude to the amplitude of the neuron input. For a given neural network model, it can be chosen arbitrarily, with higher values of β resulting in stronger light intensities and a higher signal-to-noise ratio. The approximation of treating inputs as coherent states may not be correct when one is dealing with input states that are quantum themselves, for example, when they have been affected by a strong single-photon nonlinearity in the previous computation layer. In the following, we exclude this possibility, which is consistent with the fact that the nonlinearity of optical materials does not allow us to achieve such a strong single-photon nonlinearity except for very specific configurations [72,73].

For convenience, we denote weighted inputs with $a_i = w_i x_i$. Thus, we take input states of the neuron as $|\alpha_i\rangle = \beta a_i$, $i = 1, \ldots, N$, and assume that the output state is approximately a coherent state. The weighted sum of inputs, i.e., the state of light in the spatial and temporal mode corresponding to the neuron, is a superposition of N input states, i.e., $|\alpha\rangle = \sum_{i=1}^{N} |\alpha_i\rangle$, and it is also a coherent state with $\alpha = \sum_i \alpha_i$. We neglect phase factors such as $e^{i(kr-\omega t)}$ since we can select the basis for input coherent states in such a way that they are eliminated.

We can now determine all the quantum properties of light, in particular its intensity and fluctuations. To determine the fluctuations, it is convenient to use quadratures $\hat{X}_1$ and $\hat{X}_2$, with $\alpha = \langle\alpha|\hat{X}_1|\alpha\rangle + i\langle\alpha|\hat{X}_2|\alpha\rangle$. In our neural network model, we simulate quantum noise by introducing $a_i' = a_i + \delta a_i$ with δa_i being random variables reproducing quantum shot noise, with appropriate statistics. It is clear that the expectation value of a_i' should be equal to $\bar{a_i'} = a_i = \alpha_i/\beta = (\langle\alpha_i|\hat{X}_1|\alpha_i\rangle + i\langle\alpha_i|\hat{X}_2|\alpha_i\rangle)/\beta$. On a similar basis, the fluctuations δa_i will also scale proportionally to $1/\beta$, so we finally get $a_i' = a_i + \delta a_i$, where δa_i is a complex Gaussian noise with variance

$$(\Delta \,\mathrm{Re}(\delta a_i))^2 = ((\Delta\hat{X}_1)^2)/\beta^2 = 1/4\beta^2, \tag{9}$$

$$(\Delta \,\mathrm{Im}(\delta a_i))^2 = ((\Delta\hat{X}_2)^2)/\beta^2 = 1/4\beta^2. \tag{10}$$

This defines the statistical properties of a_i', which we use in numerical simulations. At the same time, we can determine the average energy of input light pulses from the formula $E_i = \hbar\omega|\alpha_i|^2$, which scales proportionally to $|\beta|^2$.

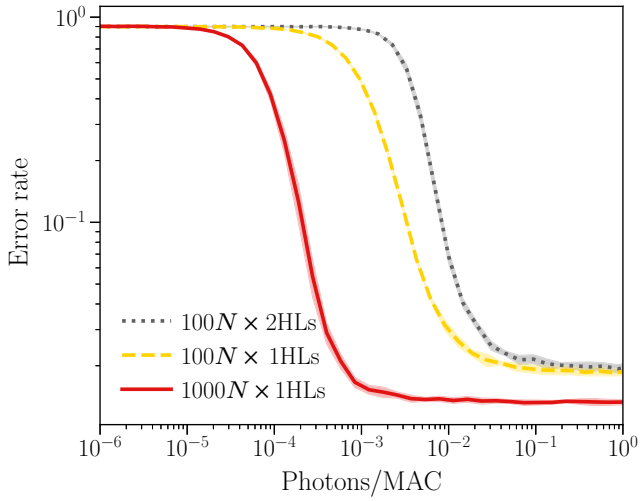

FIG. 5. Error rate of neural networks as a function of the number of photons per synaptic operation, with quantum noise included. Results are shown for networks with a single hidden layer and two hidden layers. It was assumed that the β factor relating the coherent state amplitude to the amplitude of the neuron input is the same in all layers.

In Fig. 5 we present the results of simulations of optical networks with quantum noise included. As in the optoelectronic case [18], we find that the error rate of predictions can remain low even in the case when the number of photons per operation is lower than unity. In the optical range, this corresponds to hundreds of zeptojoules per operation. As a result, we may expect that quantum noise will not be a limiting factor up to this level of energy efficiency.

H. Signal degradation and network depth

While optical signal regeneration and amplification is possible [52,53], signal decay and degradation is one of the most important challenges for all-optical systems, especially in the case of multilayer networks. Although the cascadability of optical neurons has been achieved [40], full cascadability in a large-scale system may be difficult to realize in practice. Moreover, elements of the optical setup necessary for light beam steering may lead to significant losses [74]. In the case where full regeneration is not viable, or signal distortion at each layer is significant, these

factors will limit the possible number of network layers. Since the most successful applications of machine learning are based on deep networks, with up to thousands of layers in networks such as ResNet, this is an obstacle that could limit the practical use of all-optical networks.

Recent results in the field of machine learning suggest that the number of neural network layers can often be greatly reduced without the loss of accuracy if model designs are appropriately modified. Examples include shallow networks for speech and image recognition [75,76], nondeep networks achieving state-of-the-art results with a reduced number of layers [77], and shallow transformer networks for language models that successfully compete with recurrent neural networks [78]. In many of these cases, one or two hidden layers are enough to obtain high accuracy of predictions. Some authors suggested that in the case of fully connected or convolutional networks, making networks deeper beyond a relatively shallow level does not improve accuracy [79,80]. These observations are aligned with the arguments considering models of physical systems, which are usually described by Hamiltonians that are low-order polynomials [81].

I. Neural network architectures

All-optical networks have limited flexibility of possible architectures. This also concerns the structure of computations. Some neural network architectures are easier to implement than others. For example, it may be straightforward to design an optical feed-forward neural network with scalar nonlinear activation functions, but more complex models require vectorial nonlinear operations. It is known that vector-matrix multiplications can be implemented optically as long as the vector component is encoded optically, while the matrix component is encoded in the material part of the device [25]. The same is true for convolutions [82–84]. However, it is not known how to implement all-optically some other nonlinear transformations that are important for neural network models. These include vector-matrix multiplications and softmax activations, which are key components of attention layers [36,48], where both the vector and the matrix are to be encoded optically. It is important to either find a way to realize these functions all-optically, or to determine alternative models that do not require these functions, but are able to perform the same tasks with comparable accuracy.

J. Constant radiance theorem

The constant radiance theorem imposes a fundamental limitation on the geometry and optical energy required for performing computations with light [85]. This theorem states that in the case of linear propagation of light, the generalized étendue, which measures the spread of light in real and momentum space, remains constant. In the case of neural networks, this condition imposes a limit on the optical energy per neuron, which inversely scales with the number of neurons. In particular, if the number of neurons in a hidden layer is much higher than in the input layer, the energy per neuron in the hidden layer is allowed to be much lower than the average energy of optical inputs. As a result, if the condition stated in Sec. II is fulfilled, that is, $N_{\text{hidden}} \gg N_{\text{input}} + N_{\text{output}}$, high energy efficiency of operations in the hidden layer is not excluded by the constant radiance theorem.

Moreover, the constant radiance theorem applies strictly only in the linear regime. For example, in the setup shown in Fig. 2(a), the power of the output beam, controlled by multiple input beams, is not limited to $1/N_{\text{in}}$ of the power of inputs, where N_{in} is the number of input beams for this neuron. This is because the nonlinear interaction occurs only in a small region in space where the light is focused. We also assume that the interaction is phase insensitive, as in the case of the optical Kerr effect in the case when control and pump beams are not coherent with each other. In this setting, it is the volume in real space (and not in the generalized real-momentum space, as in the case of étendue) that is physically relevant for the power redistribution.

On the other hand, in the case of small neural networks or networks in which this condition is not fulfilled, the constant radiance theorem will impose a limit on the achievable energy efficiency. For all-optical networks, this has to be considered as an important factor in system design.

IV. CONCLUSIONS

In conclusion, under certain plausible assumptions about the limitations of electronics, we showed that all-optical neural networks can find an important role in the applications of machine learning. It is estimated that all-optical devices could outperform both electronic and optoelectronic devices by orders of magnitude in energy efficiency in the case of inference in large neural network models. This estimate takes into account all the components of the complete system, including the cost of memory access and data acquisition from remote resources. All-optical networks are predicted to give the biggest advantage in the case of generative models, where the cost of data acquisition and memory access is reduced due to the possibility to reuse input data.

On the other hand, it is clear that there are still important issues that need to be solved before all-optical networks become practical. These include scalability of optical neurons, signal decay and distortion, the strength of nonlinearity, and the nonuniversal character of optical computing. To overcome these obstacles, studies on both the physical implementations of optical networks and on accommodating neural network models to the capabilities of optical systems may be necessary. It is likely that an

interdisciplinary approach will be the key to successful implementations.

ACKNOWLEDGMENTS

We acknowledge support from the National Science Center, Poland under Grants No. 2019/35/N/ST3/01379, No. 2020/37/B/ST3/01657, and No. 2021/43/B/ST3/00752, and the Foundation for Polish Science (FNP).

APPENDIX: NEURAL NETWORK MODELS

In order to solve the MNIST task and address the questions presented in the main text of the manuscript, we constructed two types of fully connected feed-forward neural networks: real-valued neural networks (RVNNs) and complex-valued neural networks (CVNNs).

We used Tensorflow for the RVNNs and the cvnn library [64] for CVNNs to implement network models. Input layers consisted of 28^2 neurons corresponding to the image resolution of the MNIST dataset. Depending on the case, one or two hidden layers were added with nonlinear activation functions given by Eqs. (5) and (6). Finally, the output layer was composed of ten neurons with softmax activation functions. In the case of CVNNs, we used the softmax function applied to the modulus of the complex number at each neuron.

1. Hyperparametrization

During the training procedure, the learning rate used in both algorithms was set to 0.001, and the "Adam" optimizer was selected. The batch size was set to 1000 samples (the training set contains 60 000 images, while the testing set has 10 000). The number of neurons in the hidden layer(s) for both real and complex networks was different to ensure that the comparison between the networks is fair (comparing networks with the same number of trainable parameters). We observed that all four activation functions reached saturation of accuracy after about 150 epochs of training; see Fig. 6. In Fig. 3, we selected the optimal number of epochs for each activation function (and each number of neurons in the hidden layer) separately to ensure a fair comparison. The results of each simulation were averaged over five runs.

2. Noise and disorder simulations

For simulations with disorder and noise, we focused on complex-valued networks with sigmoid activation functions. The plots were obtained by saving the weights of the trained "ideal" neural network and "manually" testing the network on the MNIST test set with additional modifications corresponding to fabrication errors (Sec. III F) or quantum noise (Sec. III G).

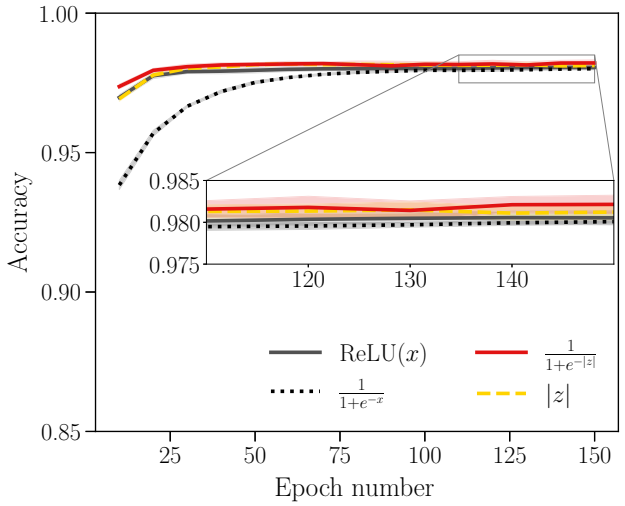

FIG. 6. Accuracy of handwritten digit recognition for fully connected complex- and real-valued networks with 100 and 200 neurons in the hidden layer, respectively, as a function of the number of epochs. We see that all four networks reach saturation before 150 epochs of training.

In the case of fabrication error analysis, parameters a_j, b_j, c_j of Eq. (8) were chosen from the probability distribution $p(x)$ once for each neuron (for the entire testing procedure) for a given value of σ and the number of neurons in the hidden layer. This is because fabrication errors are meant to represent the deviations of physical (optical) neurons from the ideal neuron.

In the quantum noise analysis, random fluctuations of activations δa_i were chosen from a Gaussian distribution independently for each image from the MNIST test set. The number of photons per operation was calculated in the following way. First, we summed all absolute values of weighted inputs squared, $|a_i|^2$, of all connections in the network. We then multiplied them by the factor $|\beta|^2$ in order to obtain the total number of photons. In this way, we take into account both input pulses and pump pulses in hidden layers. Finally, we divided the result by the number of operations. Since the number of neuron inputs is large in the considered cases, the number of synaptic operations is almost equal to the number of MAC operations, with the remaining nonlinear operations being a small fraction.

[1] A. Mehonic and A. J. Kenyon, Brain-inspired computing needs a master plan, Nature **604**, 255 (2022).

[2] M. M. Waldrop, The chips are down for Moore's law, Nat. News **530**, 144 (2016).

[3] 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).

[4] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, and P. Deaville, In-memory computing:

Advances and prospects, IEEE Solid-State Circuits Mag. **11**, 43 (2019).
[5] D. A. Miller, Attojoule optoelectronics for low-energy information processing and communications, J. Lightwave Technol. **35**, 346 (2017).
[6] J. Misra and I. Saha, Artificial neural networks in hardware: A survey of two decades of progress, Neurocomputing **74**, 239 (2010).
[7] P. A. Merolla, *et al.*, A million spiking-neuron integrated circuit with a scalable communication network and interface, Science **345**, 668 (2014).
[8] M. Davies, *et al.*, Loihi: A neuromorphic manycore processor with on-chip learning, IEEE Micro **38**, 82 (2018).
[9] 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).
[10] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, Physics for neuromorphic computing, Nat. Rev. Phys. **2**, 499 (2020).
[11] J. Zhu, T. Zhang, Y. Yang, and R. Huang, A comprehensive review on emerging artificial neuromorphic devices, Appl. Phys. Rev. **7**, 011312 (2020).
[12] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, Recent advances in physical reservoir computing: A review, Neural. Netw. **115**, 100 (2019).
[13] 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).
[14] 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).
[15] 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).
[16] 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).
[17] 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).
[18] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, Large-scale optical neural networks based on photoelectric multiplication, Phys. Rev. X **9**, 021032 (2019).
[19] N. Stroev and N. G. Berloff, Analog photonics computing for information processing, inference, and optimization, Adv. Quantum Technol. **6**, 2300055 (2023).
[20] A. Hirose, Applications of complex-valued neural networks to coherent optical computing using phase-sensitive detection scheme, Inf. Sci.-Appl. **2**, 103 (1994).
[21] H. Zhang, M. Gu, X. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, and M. Yung, *et al.*, An optical neural chip for implementing complex-valued neural network, Nat. Commun. **12**, 457 (2021).
[22] 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).
[23] K. Tyszka, M. Furman, R. Mirek, M. Król, A. Opala, B. Seredyński, J. Suffczyński, W. Pacuski, M. Matuszewski, J. Szczytko, and B. Piętka, Leaky integrate-and-fire mechanism in exciton–polariton condensates for photonic spiking neurons, Laser Photonics Rev. **17**, 2100660 (2023).
[24] E. Agrell, M. Karlsson, A. R. Chraplyvy, D. J. Richardson, P. M. Krummrich, P. Winzer, K. Roberts, J. K. Fischer, S. J. Savory, B. J. Eggleton, M. Secondini, F. R. Kschischang, A. Lord, J. Prat, I. Tomkos, J. E. Bowers, S. Srinivasan, M. Brandt-Pearce, and N. Gisin, Roadmap of optical communications, J. Opt. **18**, 063002 (2016).
[25] 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).
[26] D. A. B. Miller, Are optical transistors the logical next step?, Nat. Photonics **4**, 3 (2010).
[27] F. Ashtiani, A. J. Geers, and F. Aflatouni, An on-chip photonic deep neural network for image classification, Nature **606**, 501 (2022).
[28] T. Wang, M. M. Sohoni, L. G. Wright, M. M. Stein, S.-Y. Ma, T. Onodera, M. G. Anderson, and P. L. McMahon, Image sensing with multilayer nonlinear optical neural networks, Nat. Photonics **17**, 408 (415).
[29] D. A. B. Miller, Optical interconnects to electronic chips, Appl. Opt. **49**, F59 (2010).
[30] L. Bernstein, A. Sludds, C. Panuski, S. Trajtenberg-Mills, R. Hamerly, and D. Englund, Single-shot optical neural network, Sci. Adv. **9**, eadg7904 (2023).
[31] 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).
[32] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, and A. Borchers, *et al.*, in *Proceedings of the 44th Annual International Symposium on Computer Architecture* (ACM, Toronto, 2017), p. 1.
[33] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, A. Nomura, T. Yasuda, A. Chen, and A. Friz, *et al.*, An analog-AI chip for energy-efficient speech recognition and transcription, Nature **620**, 768 (2023).
[34] 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).
[35] 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).
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. **30**, 1 (2017).

[37] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33 (2019), p. 4780.

[38] A. N. Tait, T. F. 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).

[39] Z. Chen, A. Sludds, R. Davis, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein, R. Hamerly, and D. Englund, Deep learning with coherent VCSEL neural networks, Nat. Photonics **17**, 723 (2023).

[40] S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, Single chip photonic deep neural network with accelerated training, ArXiv:2208.01623 (2022).

[41] A. N. Tait, Quantifying power in silicon photonic neural networks, Phys. Rev. Appl. **17**, 054029 (2022).

[42] C. Demirkiran, F. Eris, G. Wang, J. Elmhurst, N. Moore, N. C. Harris, A. Basumallik, V. J. Reddi, A. Joshi, and D. Bunandar, An electro-photonic system for accelerating deep neural networks, J. Emerg. Technol. Comput. Syst. **19**, 1 (2023).

[43] M. Matuszewski, A. Opala, R. Mirek, M. Furman, M. Król, K. Tyszka, T. C. H. Liew, D. Ballarini, D. Sanvitto, J. Szczytko, and B. Piętka, Energy-efficient neural network inference with microcavity exciton polaritons, Phys. Rev. Appl. **16**, 024045 (2021).

[44] E. Estrecho, T. Gao, N. Bobrovska, D. Comber-Todd, M. D. Fraser, M. Steger, K. West, L. N. Pfeiffer, J. Levinsen, 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).

[45] D. W. Snoke, V. Hartwell, J. Beaumariage, S. Mukherjee, Y. Yoon, D. M. Myers, M. Steger, Z. Sun, K. A. Nelson, and L. N. Pfeiffer, Reanalysis of experimental determinations of polariton-polariton interactions in microcavities, Phys. Rev. B **107**, 165302 (2023).

[46] A. Snyder, S. Laughlin, and D. Stavenga, Information capacity of eyes, Vision. Res. **17**, 1163 (1977).

[47] C. A. Thraskias, E. N. Lallas, N. Neumann, L. Schares, B. J. Offrein, R. Henker, D. Plettemeier, F. Ellinger, J. Leuthold, and I. Tomkos, Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications, IEEE Commun. Surv. Tutor. **20**, 2758 (2018).

[48] M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, and P. L. McMahon, Optical transformers, ArXiv:2302.10360 (2023).

[49] D. A. Miller, Why optics needs thickness, Science **379**, 41 (2023).

[50] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, and B. Gao, *et al.*, A compute-in-memory chip based on resistive random-access memory, Nature **608**, 504 (2022).

[51] R. Khaddam-Aljameh, *et al.*, HERMES-core–A 1.59-TOPS/mm^2 PCM on 14-nm CMOS in-memory compute core using 300-ps/LSB linearized CCO-based ADCs, IEEE J. Solid-State Circuits **57**, 1027 (2022).

[52] M. T. Hill, E. E. Frietman, H. de Waardt, G.-D. Khoe, and H. J. Dorren, All fiber-optic neural network using coupled SOA based ring lasers, IEEE Trans. Neural Netw. **13**, 1504 (2002).

[53] D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, A high performance photonic pulse processing device, Opt. Express **17**, 22767 (2009).

[54] Q. Guo, R. Sekine, L. Ledezma, R. Nehra, D. J. Dean, A. Roy, R. M. Gray, S. Jahani, and A. Marandi, Femtojoule femtosecond all-optical switching in lithium niobate nanophotonics, Nat. Photonics **16**, 625 (2022).

[55] C.-W. Shih, I. Limame, S. Krüger, C. C. Palekar, A. Koulas-Simos, D. Brunner, and S. Reitzenstein, Low-threshold lasing of optically pumped micropillar lasers with $Al_{0.2}Ga_{0.8}As/Al_{0.9}Ga_{0.1}As$ distributed Bragg reflectors, Appl. Phys. Lett. **122**, 151111 (2023).

[56] R. Mirek, A. Opala, P. Comaron, M. Furman, M. Król, K. Tyszka, B. Seredyński, 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).

[57] 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).

[58] 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).

[59] 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).

[60] N. Bobrovska, M. Matuszewski, K. S. Daskalakis, S. A. Maier, and S. Kéna-Cohen, Dynamical instability of a nonequilibrium exciton-polariton condensate, ACS Photonics **5**, 111 (2018).

[61] B. Datta, M. Khatoniar, P. Deshmukh, F. Thouin, R. Bushati, S. De Liberato, S. K. Cohen, and V. M. Menon, Highly nonlinear dipolar exciton-polaritons in bilayer MoS_2, Nat. Commun. **13**, 6341 (2022).

[62] R. Emmanuele, M. Sich, O. Kyriienko, V. Shahnazaryan, F. Withers, A. Catanzaro, P. Walker, F. Benimetskiy, M. Skolnick, and A. Tartakovskii, *et al.*, Highly nonlinear trion-polaritons in a monolayer semiconductor, Nat. Commun. **11**, 3589 (2020).

[63] J. Gu, V. Walther, L. Waldecker, D. Rhodes, A. Raja, J. C. Hone, T. F. Heinz, S. Kéna-Cohen, T. Pohl, and V. M. Menon, Enhanced nonlinear interaction of polaritons via excitonic Rydberg states in monolayer WSe_2, Nat. Commun. **12**, 2269 (2021).

[64] J. A. Barrachina, NEGU93/cvnn: Complex-valued neural networks (2022).

[65] 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.

[66] 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.

[67] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, Wide reduced-precision networks, ArXiv:1709.01134 (2017).

[68] L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, and P. L. McMahon, Deep physical neural networks trained with backpropagation, Nature **601**, 549 (2022).

[69] J. Spall, X. Guo, and A. I. Lvovsky, Hybrid training of optical neural networks, Optica **9**, 803 (2022).

[70] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, An optical neural network using less than 1 photon per multiplication, Nat. Commun. **13**, 123 (2022).

[71] S.-Y. Ma, T. Wang, J. Laydevant, L. G. Wright, and P. L. McMahon, Quantum-noise-limited optical neural networks operating at a few quanta per activation, ArXiv:2307.15712 (2023).

[72] Y. Arakawa and M. J. Holmes, Progress in quantum-dot single photon sources for quantum information technologies: A broad spectrum overview, Appl. Phys. Rev. **7**, 021309 (2020).

[73] G. Muñoz-Matutano, A. Wood, M. Johnsson, X. Vidal, B. Q. Baragiola, A. Reinhard, A. Lemaître, J. Bloch, A. Amo, G. Nogues, B. Besga, M. Richard, and T. Volz, Emergence of quantum correlations from interacting fibre-cavity polaritons, Nat. Mater. **18**, 213 (2019).

[74] F. B. McCormick, T. J. Cloonan, F. Tooley, A. L. Lentine, J. M. Sasian, J. Brubaker, R. L. Morrison, S. Walker, R. Crisci, and R. Novotny, *et al.*, Six-stage digital free-space optical switching network using symmetric self-electro-optic-effect devices, Appl. Opt. **32**, 5153 (1993).

[75] J. Ba and R. Caruana, in *Advances in Neural Information Processing Systems*, Vol. 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Curran Associates, Inc., 2014), p. 1.

[76] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, A. Mohamed, M. Philipose, M. Richardson, and R. Caruana, in *International Conference on Learning Representations* (2017), https://openreview.net/forum?id=r10FA8Kxg.

[77] A. Goyal, A. Bochkovskiy, J. Deng, and V. Koltun, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022), p. 6789.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), p. 4171.

[79] D. A. Winkler and T. C. Le, Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR, Mol. Inform. **36**, 1600118 (2017).

[80] A. J. Thomas, M. Petridis, S. D. Walters, S. M. Gheytassi, and R. E. Morgan, in *Engineering Applications of Neural Networks: 18th International Conference, EANN 2017, Athens, Greece, August 25–27, 2017, Proceedings* (Springer, Athens, 2017), p. 279.

[81] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well?, J. Stat. Phys. **168**, 1223 (2017).

[82] 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).

[83] 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).

[84] 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).

[85] J. W. Goodman, Fan-in and fan-out with optical interconnections, Opt. Acta: Int. J. Opt. **32**, 1489 (1985).