Chip

Research Article

ARTIFICIAL INTELLIGENCE

DOI: 10.1016/j.chip.2021.100002

Low-depth optical neural networks

Xiao-Ming Zhang[1,2,*] & Man-Hong Yung[1,3,4,5,*]

[1]*Department of Physics, Southern University of Science and Technology, Shenzhen 518055, China* [2]*Department of Physics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China* [3]*Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China* [4]*Guangdong Provincial Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China* [5]*Shenzhen Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China*

E-mails: phyxmz@gmail.com (Xiao-Ming Zhang), yung@sustech.edu.cn (Man-Hong Yung)

Cite as: Zhang, X.-M. & Yung, M.-H. *Chip* **1**, 2 (2022).
https://doi.org/10.1016/j.chip.2021.100002

Received: 11 December 2021

Accepted: 12 December 2021

Published online: 31 January 2022

Optical neural network (ONNs) are emerging as attractive proposals for machine-learning applications. However, the stability of ONNs decreases with the circuit depth, limiting the scalability of ONNs for practical uses. Here we demonstrate how to compress the circuit depth to scale only logarithmically in terms of the dimension of the data, leading to an exponential gain in terms of noise robustness. Our low-depth (LD)-ONN is based on an architecture, called Optical CompuTing Of dot-Product UnitS (OCTOPUS), which can also be applied individually as a linear perceptron for solving classification problems. We present both numerical and theoretical evidence showing that LD-ONN can exhibit a significant improvement on robustness, compared with previous ONN proposals based on singular-value decomposition.

Keywords: optical neural networks, photonic chip, machine learning

INTRODUCTION

Photonic computation represents an emerging technology enabling high-speed information processing with low energy consumption[1]. Such technology can potentially be applied to solve many problems of machine learning, which has already created a significant impact on the physics community[2–10]. In particular, efforts have been made for decades in developing optical neural networks (ONNs) with different approaches[11–20]. Recently, much progress has been made in developing scalable on-chip photonic circuits[1,21–24], leading to a new avenue towards large-scale implementation of ONNs. Compared with its free-space counterpart, on-chip ONN has advantages in terms of programmability and integrability[13]. This unconventional hardware architecture could potentially revolutionize the field of AI computing.

In order to achieve scalable ONNs, various circuit designs have been proposed recently[13,17,18], and they share similar characteristics, such as the form of the multiport interferometers and the scaling complexity of the circuit depth. In particular, ONN-based deep learning has been experimentally demonstrated[13], by applying singular-value decomposition (SVD) for constructing any given linear transformation. Physically, these unitary transformations can be achieved with multiport interferometers[25,26], together with a set of diagonal attenuators.

However, the circuit structure of SVD-ONN is only applicable for linear transformation represented by a square matrix. Here and after, we denote the dimension of the input and output vectors as N and M respectively. The SVD approach of ONN requires $O(\max(N, M))$ layers of interferometers. As each layer will introduce errors to its output, the scalability of the SVD approach of ONN is limited by the errors scaling as $O(\max(N, M))$.

Moreover, for machine-learning tasks of practical interest, both cases of $N \gg M$ (e.g. image recognition[27]) and $M \gg N$ (e.g. generative model[28]) are very common. Therefore, the SVD approach would require appending a large number of ancillary modes to "square the matrix", increasing the spatial complexity of the ONN.

To surmount the problem of robustness and flexibility, we propose an alternative approach of ONN for performing machine-learning tasks. Our ONN is constructed by connecting basic optical units, called Optical Computation of dot-Product Units (OCTOPUS), which optically outputs the dot-product of two vectors; the resulting circuit depth scales logarithmically $O(\log N)$. Even a single OCTOPUS can be applied as an optical linear perceptron[29]. In addition, the noise robustness of the OCTOPUS exhibits an exponential advantage compared with the SVD approach (see Supplementary Materials for the theoretical analysis).

For constructing a deep neural network, we propose two variants of low depth ONN, called tree low depth (TLD) and recursive low depth (RLD) ONN. Both architectures involve OCTOPUS as basic optical computing units, and they are applicable to non-square transformation at each layer as well, i.e., $N \neq M$. The TLD-ONN requires fewer optical elements, but may cost more energy; the RLD-ONN involves a more complex structure, but it is more energy efficient. In terms of noise robustness, our numerical simulation suggests that TLD- and RTD-ONN have the same level of robustness, while both of them are significantly better than SVD-ONN.

RESULTS

We begin our discussion with linear transformation. Given a one-dimensional real vector $\boldsymbol{x}$ and an $N \times M$ real transformation matrix W, our goal is to optically achieve the following linear transformation

$$\boldsymbol{y} = W\boldsymbol{x}. \tag{1}$$

In the SVD approach[13], $N = M$ is assumed. Otherwise one needs to manually append many 0s to square the corresponding matrix and vectors. Then, the matrix is decomposed as $W = V^\dagger \Sigma U$ (see Fig. 1a), where U and $V^\dagger$ are unitary matrices, and Σ is a diagonal matrix containing all singular values of W. In optical implementation, U and $V^\dagger$ can be realized with multiport interferometers and Σ can be realized with a set of atten-

 DOI: 10.1016/j.chip.2021.100002

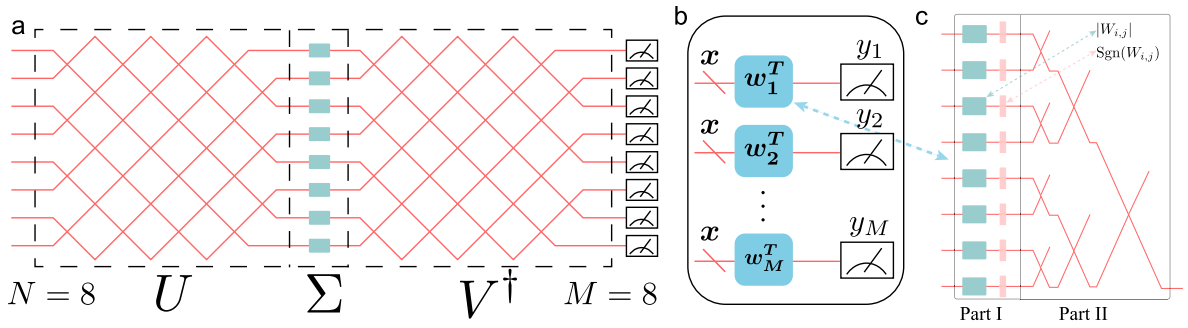

Fig. 1 | (a) The SVD approach for linear transformation. The 8×8 square matrix is decomposed into two unitaries U and $V^\dagger$, and a diagonal matrx Σ. The unitary matrices are realized by a set of M-Z interferometers, and Σ is realized with a set of attenuators. **(b)** Linear transformation with OCTOPUS. Each OPCTPUS corresponds to one row of the transformation matrix $\boldsymbol{w_i^T}$. **(c)** Sketch of the OCTOPUS calculating $y_i = \boldsymbol{w_i^T} \cdot \boldsymbol{x}$. The dimension of input vector $\boldsymbol{x}$ is $N = 8$. Part I: the attenuators (green) and the phase shifters (pink) encode values of $\boldsymbol{w_i^T}$. Part II: the interferometer tree performs the summation.

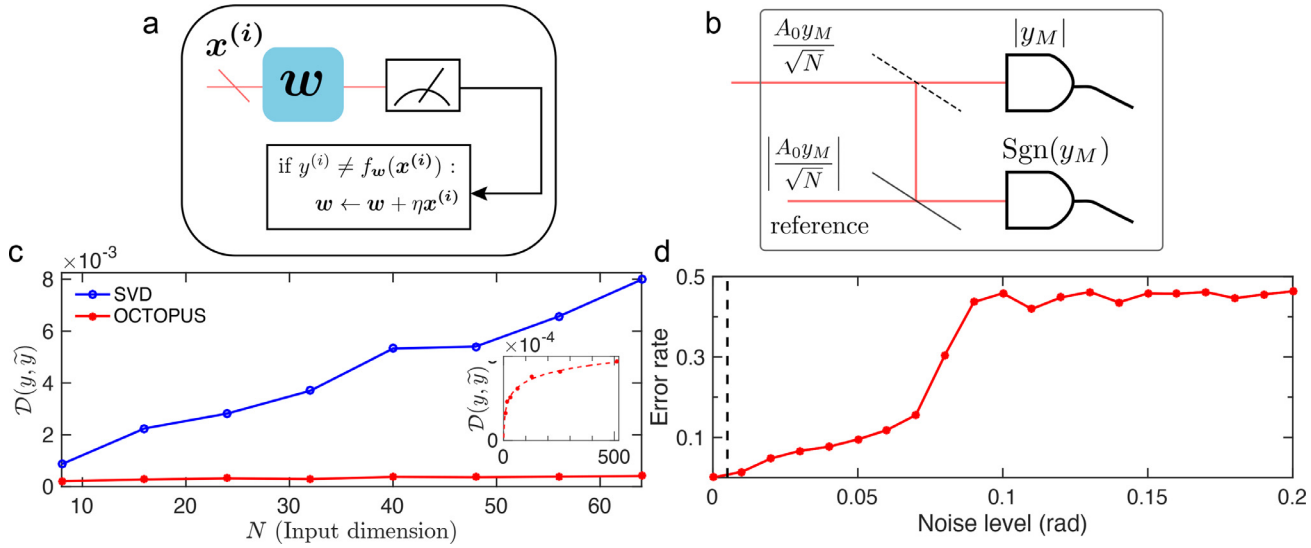

Fig. 2 | (a) Sketch of the training process of the optical linear perceptron. $\boldsymbol{x^{(i)}}$ and $y^{(i)}$ correspond to the training data and the label at the ith iteration, respectively. $\boldsymbol{w}^T \cdot \boldsymbol{x^{(i)}}$ is calculated with OCTOPUS, after which we obtain the value of $f_{\boldsymbol{w}}(\boldsymbol{x^{(i)}})$. If $y^{(i)} \neq f_{\boldsymbol{w}}(\boldsymbol{x^{(i)}})$, the weight $\boldsymbol{w}$ is updated. **(b)** Amplitude measurement. The black (dashed) lines represent 50/50 beam splitters. **(c)** Robustness of the optical linear transformation under both encoding noise and shot noise. We let the dimension of input and output vectors be identical, i.e., $M = N$. Main panel: comparison of the cosine distance $\mathcal{D}(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ for the SVD and the OCTOPUS approaches. Encoding error level is set to be $\sigma_I = \sigma_A = 0.005$, and the input amplitude along the path corresponding to x_i is set to be $2^8 \cdot x_i$. Inset: $\mathcal{D}(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ for the OCTOPUS approach when $\sigma_I = 0.005$, and $\sigma_A = 0$. The dots are simulation data, and the dashed line is fitted with $\mathcal{D}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = A \log N + B$. All results are averaged over 10 runs. **(d)** Linear perceptron simulation on the "Iris" data set. Red dots correspond to the error rate versus noise level $\sigma = \sigma_I = \sigma_A$ after 1000 iterations of training. The red line is the guide for the eye. The black dashed line corresponds to $\sigma = 0.005$. The results are averaged over 100 runs.

uators[13,30]. Because this scheme is valid only when the maximal singular value $s_{\max}$ is not larger than 1; when $s_{\max} > 1$, one should first rescale the matrix as $W \to 1/s_{\max} W$.

In contrast, our OCTOPUS (Fig. 1b and Fig. 1c) solves the same problem by calculating the elements of the output $\boldsymbol{y}$ "one by one". For vector $\boldsymbol{x}$, we require M copies of the optical input with amplitude $A_0\boldsymbol{x}$, where A_0 is a predetermined constant. Each copy serves as the input of one OCTOPUS. The ith OCTOPUS encodes the ith row of the matrix W (denoted with w_i), and aims at calculating the ith element of the result $y_i = \boldsymbol{w_i}^T \cdot \boldsymbol{x}$. Similar to the SVD approach, each OCTOPUS are constructed with a set of interferometers and attenuators. We refer the readers to Methods section and Supplementary Materials for more details.

The main noise source for optical computation is the encoding noise. The inaccurate length of each path of the interferometers and attenuators, which is mainly due to the finite precision of waveguide fabrication and the imperfect tuning process, will introduce an extra phase shift of the signals. During our simulation, this phase shift is assumed to follow normal distribution with zero mean. The computation result is affected by the encoding noise determined by the circuit depth, defined as the maximum number of interferometers and attenuators the signal should pass through from its input port to its output. The circuit depths of the SVD approach and the OCTOPUS approach are very different. The former scales linearly, $O(N)$, and the latter logarithmically, $O(\log N)$, leading to a dramatic difference in terms of the noise robustness against the encoding noise of the optical elements. Specifically, let us denote the output vector subject to noise with $\tilde{\boldsymbol{y}}$. The error can be quantified by *cosine distance*

$$\mathcal{D}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \equiv 1 - \frac{\boldsymbol{y} \cdot \tilde{\boldsymbol{y}}}{\|\boldsymbol{y}\| \|\tilde{\boldsymbol{y}}\|}, \tag{2}$$

which has been widely adopted in classification problems[31–33].

Our simulation results comparing the robustness of the SVD approach with that of OCTOPUS are shown in Fig. 2c. For the SVD approach,

DOI: 10.1016/j.chip.2021.100002

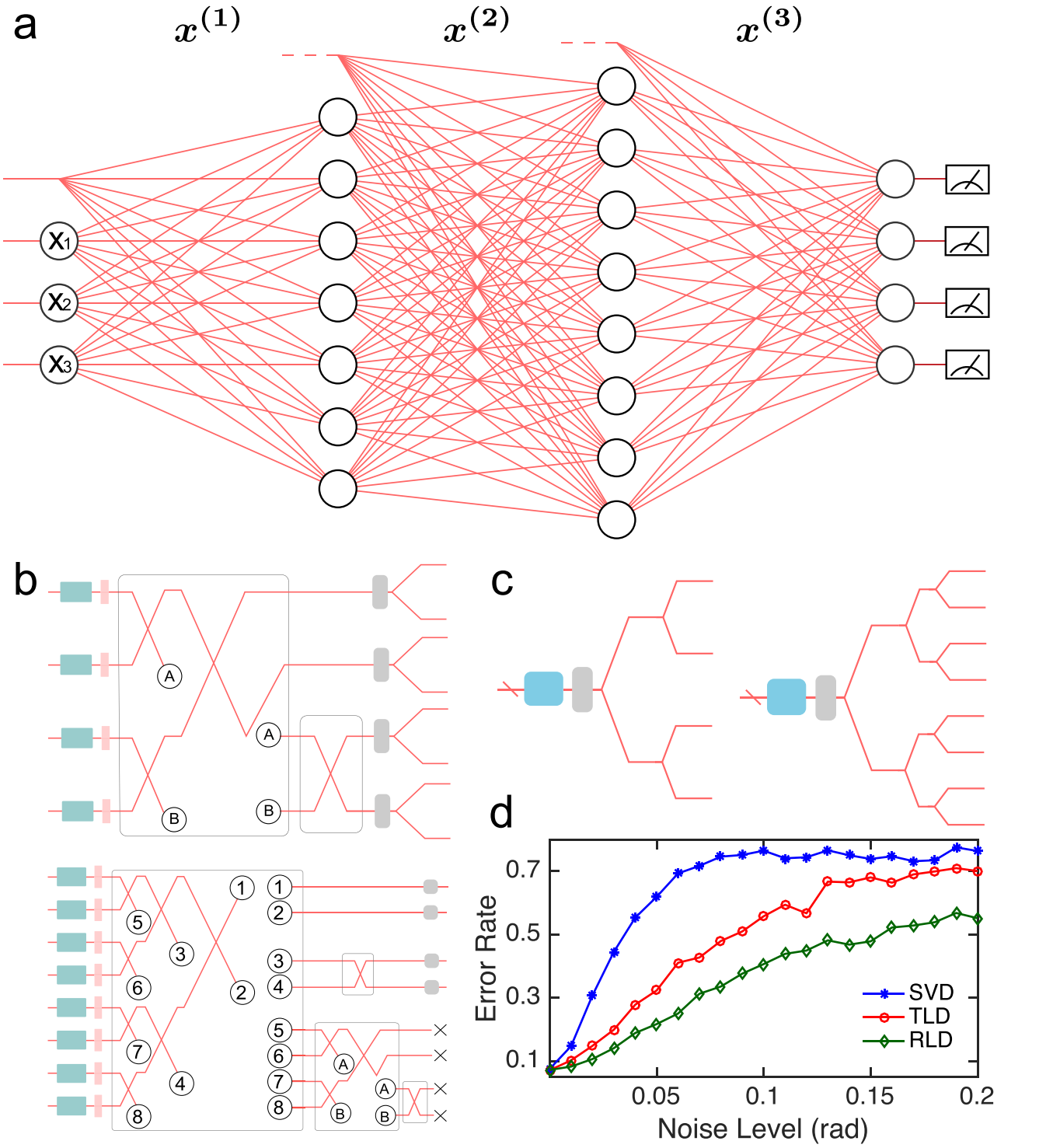

Fig. 3 | (a) General structure of TLD- and RLD-ONN with 3,7,8, and 4 neurons at different layers. The red lines represent optical paths. The input data denoted with $\boldsymbol{x} = [x_1, x_2, x_3]$ are encoded at the first layer. Neurons are represented with circles. **(b)** Realization of the neurons at the first (upper panel) and the second (lower panel) layers of RLD-ONN. The input signals pass through sets of tunable attenuators (green) and phase shifters (pink). Then, several interferometer trees are appended recursively, until all output ports are connected to all input ports. Here, paths with same labels are connected to each other. The grey panels represent nonlinear activation. The paths denoted with black crosses are discarded. **(c)** Realization of the neurons at the first (left panel) and the second (right panel) layers of TLD-ONN. The input signal first passes through an OCTOPUS and then a nonlinear activation. Finally, it is split uniformly into several paths. **(d)** Error rate comparison of the Letter Classification task.

$\mathcal{D}(\boldsymbol{y}, \widetilde{\boldsymbol{y}})$ increases linearly with N. In contrast, $\mathcal{D}(\boldsymbol{y}, \widetilde{\boldsymbol{y}})$ for OCTOPUS grows only very slowly. These results are consistent with the scaling of the circuit depths of the two approaches. In Supplementary Materials, we provide further theoretical analysis of the noise effect. For the SVD approach, the error scales linearly with the data size N, $\mathcal{D}(\boldsymbol{y}, \widetilde{\boldsymbol{y}}) \sim \sigma_I^2 N + \sigma_A^2$, where σ_I and σ_A represent noise level for the interferometers and the attenuators, respectively. However, for OCTOPUS, the error scales only logarithmically, $\mathcal{D}(\boldsymbol{y}, \widetilde{\boldsymbol{y}}) \sim \sigma_I^2 \log N + \sigma_A$. This exponential advantage of the OCTOPUS approach agrees well with our numerical results (see inset of Fig. 2c).

Another important noise source is shot noise: Instead of being a fixed value, the measurement outcome of the coherent light follows Poisson distribution. The signal to noise ratio (SNR), defined as the mean of the estimated amplitude divided by its standard deviation, is proportional to the signal intensity. In Supplementary Materials, we discuss the relation between A_0 and the average SNR ($\langle \mathrm{SNR} \rangle$) for linear transformation with random matrix and input vector of zero mean. In particular, when $N = M$, $\langle \mathrm{SNR} \rangle$ are proportional to A_0 and independent of N for both the OCTOPUS and the SVD approaches. We also found that if one requires $\langle \mathrm{SNR} \rangle = 2^8$, A_0 should be chosen as $A_0 = 433.7$ photon$^{1/2}$ for the SVD and $A_0 = 388.1$ photon$^{1/2}$ for the OCTOPUS approaches.

As OCTOPUS enables one to address each element of the input vector directly, it is possible to optically realize linear perceptrons[34,35] with OCTOPUS for solving classification problems in machine learning (see Fig. 2a and Method).

For the binary case, the goal of linear perceptron is to output a hyperplane separating two classes of data labeled by either 0 or 1, allowing us to make prediction about the unlabelled new data (see Methods). Fig. 2d shows the simulation results of our optical linear perceptron with "Iris Data Set"[36]. We define error rate as the rate of providing incorrect prediction on the label. Remarkably, the error rate remains under 0.1 when the noise level $\sigma \leqslant 0.05$.

It is well known that linear perceptron performs well with relatively simple tasks. However, for problems involving complicated nonlinear relations, one may consider deep neural networks. In the following, we present two variants of multi-layered ONN, namely, Tree Low-Depth (TLD) and Recursive Low-Depth (RLD) ONN. Both of them share a similar structure, as illustrated in Fig. 3a. Again, the input data $\boldsymbol{x}$ is encoded at the first layer containing an attenuator and a phase shifter at each node. Optical computation is performed at each neuron (denoted with circles), encapsulating the trainable parameters of the networks. Furthermore, the optical paths at the top of each layer represent the "biases" of the corresponding layer.

As shown in Fig. 3c, for TLD-ONN, each neuron consists of an OCTOPUS together with a nonlinear activation function, which can be physically realized with nonlinear crystal[13], optical-electronic conversion[37], measurement[17] or optical amplifier[38]. Then, each path is distributed uniformly to many paths, which are the inputs of the neurons at the next layer. More details are given in Supplementary Materials.

Note that in TLD-ONN, OCTOPUS only picks one path as its output; many paths are discarded. To realize a deep ONN, one may need a strong light source or amplify the signal at each layer. We may re-structure the ONN, which is the motivation for developing the RLD-ONN approach. Alternatively, another variance "sign based" TLD-ONN is introduced in Supplementary Materials, also to reduce the energy loss.

As shown in Fig. 3b, for RLD-ONN, the input signals first pass through a set of trainable attenuators and phase shifters, followed by a (3- or 2-layer) interferometer tree. Different from TLD-ONN, no signals are discarded after these steps. At this point, only two paths (such as "1" and "2" in Fig. 3b) are fully connected to the corresponding input ports. In order to connect all other output paths, the remaining paths are sent to interferometer trees with smaller size recursively. Then, the nonlinear activation is applied to all output paths. If the number of neurons at the next layer is larger than the current number of output paths, the output paths can be expanded with a 50/50 beam splitter; if it is less, one can just discard several output paths[39]. Note that the way of connecting the input and the output paths are not unique, so further optimization can be performed. For example, one can perform "pruning" [40] after the RLD-ONN is well trained. One can cut out all connections with little influence on the computation results, and the ONN structure may be significantly simplified. Although RLD-ONN requires more optical elements, the circuit depth remains logarithmic.

There are other low depth approaches for optical neural networks and optical computations[41–43] that have been proposed very recently. However, none of them have been proven to be universal, casting doubt on their ability to represent complicated, highly nonlinear functions. To substantiate the representational power of our schemes, in Supplementary Materials, we show that the transformation of TLD-ONN is equivalent to standard feed-forward nerual network. Thus, according to the universality approximation theorem[44] of neural network, TLD-ONN can represent any continuous function to an arbitrary accuracy. We also show in Supplementary Materials that for any given form of one-hidden-layer TLD-ONNs, there always exists a RLD-ONN that is equivalent to it. Therefore, both TLD- and RLD-ONN proposed in this work are universal.

To compare the performance of LD-ONNs with that of SVD-ONN[13], we perform numerical simulation on the "Letter Recognition" data set[36], classifying letters "A", "B", "C", and "D" (see Supplementary Materials for technical details). The ONN used in the simulation contains one hidden layer with 64 neurons. While the SVD- and TLD-ONN are trained with standard back-propagation Supplementary Materials, the RLD-ONN is trained with "forward propagation"[13]. In this work, we consider the training as pre-processing, i.e., the parameters of the network are first trained on a conventional computer. However, the training can also be realized optically with little assistance from electronic devices (see Supplementary Materials). As shown in Fig. 3d, as the noise level increases, TLD- and RLD-ONN have comparable error rates, while both of them are significantly lower than the error rate of SVD-ONN. It is worth noting that different from Fig. 2d, there are nonzero error rates when the noise level $\sigma = 0$. This is because the classification task here is more complicated, so more complex ONNs are required. The error rate can be further reduced by including more layers and a larger number of neurons.

DISCUSSION AND CONCLUSION

We compare the performance of our ONN scheme to the state-of-the-art Radeon Instinct MI60 7nm GPU (denoted as GPU in the following). To avoid ambiguity, we just focus on the inference of TLD-ONN with one layer. The performance of a deep neural network depends on the ways of implementing the nonlinear activation, and should be estimated on a case-by-case basis.

The throughput, in the unit of tera floating-point operations per second (TFLOPS), is an important metric for data processing speed. In an ideal case, this value is limited by the detection rate of the photon detector, which can be as high as 100 GHz[45]. Note that the magnitude and the sign of a computation result should be obtained by two separate detections, so one can compute 5×10^{10} linear transformations per second. The corresponding throughput is $0.1NM$ TFLOPS. For large N and M, it can be much higher than the peak throughput for GPU (59 TFLOPS).

The power consumption mainly results from the supply of input signal[13,37]. We are interested in the power efficiency, defined as the throughput divided by the total power. For example, for float8 (or int8) data type, the power efficiency of GPU can be estimated by dividing the

peak throughput by the thermal design power, which turns out to be 0.2 TFLOPS/W. In optical computation, to meet the accuracy requirement for float8 (or int8) data type, it is required that $\langle \mathrm{SNR} \rangle \geqslant 2^8$ when estimating the magnitude of the result. The $\langle \mathrm{SNR} \rangle$ required for estimating the sign is much smaller, so its energy cost is negligible. Again, the energy cost depends on the distribution of the input vector and the transformation matrix. We consider a simple example where the elements of both the matrix and vector are independently drawn from $\mathcal{N}(0, 1)$, and the wavelength is set to be 1550nm. The energy cost for a single linear transformation is $(2.1 \times 10^{-15} NM)$ J, and each linear transformation corresponds to $2NM$ floating-point operations. So the total power efficiency for linear transformation can be estimated to be 952 TFLOPS/W. We have also provided a "sign based" linear transformation method in Supplementary Material, which can further improve the power efficiency substantially. Whether the power efficiency of ONN is better than electronic devices also depends on the physical realization of the nonlinear activation. We take the measurement method[17] as an example, in which the output of linear transformation is first measured, and the nonlinear function is calculated in electronic devices. The power requirement of a photodetector is on the order of 1 W, and one can also assume a 100 GHz detection rate. As we require in total M detectors and two measurements for each output element (sign and amplitude), the energy cost for nonlinear activation can be estimated as $2 \times 10^{-11} M$ J. When $N \gg 9.5 \times 10^3$, this value is negligible compared to the energy cost for linear transformation $\left((2.1 \times 10^{-15} NM)\ \mathrm{J}\right)$.

Note that several simplifications can further be made on the LD-ONN structures. Firstly, as discussed in Supplementary Materials, the ONN can be binarized with an amplifier working within the saturation regime. With the binarization of the weight, the biases and the activation function, the attenuators (for magnitude encoding) at each OCTOPUS can be removed. Secondly, instead of encoding the parameters at the phase shifter and the attenuators, they can also be encoded at the interferometers. More specifically, it is possible to remove part I of OCTOPUS, and replace the Hadamard transformation at Part II by tunable interferometers. Thirdly, after the training process, the network can be "compressed" with the "pruning" technique[40], removing all paths with weights below a threshold. The improvements or revisions above are within the reach of current technologies, and could help reduce the complexity of the hardware architectures.

An integrable, scalable, strong, and stable nonlinear activation is also key to practical applications of ONNs. We have shown that the amplifier working in saturation region (Sec.4 of Supplementary Materials) can introduce strong nonlinear effects. In the literature, there are other possible approaches, including spiking with optical cavities[46] and cold atoms with electromagnetically induced transparency[47]. Besides using all-optical nonlinear effects, there are other approaches to nonlinear activation, such as non-destructive measurements. For example, Ref. 37 have suggested converting a small portion of photons into electrical signals, which is used to modulate the remaining portion of photon signals. The on-chip nonlinear activation is an open question and requires further studies.

A summary of SVD-, TLD- and RLD-ONN is given in Table 1, providing a comparison of the cost for an ONN layer with input dimension N and output dimension M. Both TLD- and RLD-ONN have logarithmic circuit depth, leading to exponential improvements on the error scalings over the SVD approach. Furthermore, TLD-ONN requires fewer number of optical elements, but we note that it also requires discarding more paths during the implementation. On the contrary, RLD-ONN requires discarding much fewer paths (same as SVD-ONN), but at the cost of a larger number of optical elements, and requiring waveguide cross.

Table 1 | Summary of the cost per layer for different ONN structures. All values correspond to the scaling order, $O(\cdot)$, for realizing one layer of a neural network with input dimension N and output dimension M.

	SVD	TLD	RLD
Circuit depth	$\max(N, M)$	$\log N$	$\log N$
Error scaling	$\max(N, M)$	$\log N$	$\log N$
Number of elements	$\max(N^2, M^2)$	NM	N^2M

To conclude, based on OCTOPUS, we presented a new architecture of ONN for machine learning, which provides exponential improvements on the robustness against encoding error. We discussed different schemes of optical linear transformation, linear perceptron, and two variants of multi-layered ONNs. Numerical simulations with random transformations and standard machine learning data sets are employed to illustrate the robustness of our schemes. The proposed LD-ONN can be directly implemented with current photonic circuits[1] technology. Our proposal, combined with appropriate realization of nonlinear activation, provides a possible solution to machine-learning tasks of industrial interest with robust, scalable, and flexible ONNs.

METHODS

Structures of OCTOPUS. As shown in Fig. 1c, our OCTOPUS contains two parts. At Part I, $\boldsymbol{w_i}^T$ is encoded with a set of tunable attenuators and phase shifters. In particular, the attenuators encode the magnitude of $\boldsymbol{w_i}^T$, while the phase shifters conditionally add a π phase to the signal when the elements are negative. At Part II, we require a set of optical Hadamard transformations[25,26]. After each Hadamard transformation, we only trace the output port corresponding to the "sum" of its input (other paths denoted with dashed lines are discarded). They are constructed as an interferometer tree of n layers with in total $N = 2^n$ input ports and 1 output port. The amplitude of the final output becomes $\frac{A_0}{\sqrt{N}} \boldsymbol{w_i}^T \cdot \boldsymbol{x} = \frac{A_0}{\sqrt{N}} y_i$, which is the desired outcome multiplied by a constant. The amplitude measurement is sketched at Fig. 2b. More details are provided in Supplementary Materials.

Linear perceptron. With a set of training data $\{\boldsymbol{x}\}$, one needs to determine the parameters $\boldsymbol{w}$ for the following function $f_{\boldsymbol{w}}(\boldsymbol{x})$:

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{w}\cdot\boldsymbol{x} \geqslant 0 \\ 0 & \boldsymbol{w}\cdot\boldsymbol{x} < 0 \end{cases} \tag{3}$$

which can be realized with by one-time usage of OCTOPUS followed by an appropriate measurement. As shown in Fig. 2a, at the ith iteration, we use training data $\boldsymbol{x^{(i)}}$ as the input, and determine whether its corresponding label $y^{(i)}$ is equal to $f_{\boldsymbol{w}}(\boldsymbol{x^{(i)}})$. If it is not, the weight is updated according to $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta \boldsymbol{x^{(i)}}$, where η is the learning rate.

REFERENCES

1. Flamini, F., Spagnolo, N. & Sciarrino, F. Photonic quantum information processing: a review. *Rep. Prog. Phys.* **82**, 016001 (2018). https://doi.org/10.1088/1361-6633/aad5b2.
2. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. Quantum machine learning. *Nature* **549**, 195 (2017). https://doi.org/10.1038/nature23474.
3. Mehta, P., Bukov, M., Wang, C.-H., Day, A. G., Richardson, C., Fisher, C. K. & Schwab, D. J.. *A high-bias, low-variance introduction to machine learning for physicists* (2018) arXiv preprint arXiv:1803.08823. https://doi.org/10.1016/j.physrep.2019.03.001.
4. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017). https://doi.org/10.1126/science.aag2302.
5. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Physics* **13**, 431 (2017). https://doi.org/10.1038/nphys4035.

6. Ma, Y.-C. & Yung, M. H. Transforming bell's inequalities into state classifiers with machine learning. *npj Quantum Inf* **4** (2018). https://doi.org/10.1038/s41534-018-0081-3.
7. Bukov, M., Day, A. G. R., Sels, D., Weinberg, P., Polkovnikov, A. & Mehta, P. Reinforcement learning in different phases of quantum control. *Phys. Rev. X* **8**, 031086 (2018). https://doi.org/10.1103/PhysRevX.8.031086.
8. Yang, X.-C., Yung, M.-H. & Wang, X. Neural-network-designed pulse sequences for robust control of singlet-triplet qubits. *Phys. Rev. A* **97**, 042324 (2018). https://doi.org/10.1103/PhysRevA.97.042324.
9. Zhang, X.-M., Cui, Z.-W., Wang, X. & Yung, M. H. Automatic spin-chain learning to explore the quantum speed limit. *Phys. Rev. A* **97**, 052333 (2018). https://doi.org/10.1103/PhysRevA.97.052333.
10. Gao, J., Qiao, L.-F., Jiao, Z.-Q., Ma, Y.-C., Hu, C.-Q., Ren, R.-J., Yang, A.-L., Tang, H., Yung, M.-H. & Jin, X. M. Experimental machine learning of quantum states. *Phys. Rev. Lett.* **120**, 240501 (2018). https://doi.org/10.1103/PhysRevLett.120.240501.
11. Wagner, K. & Psaltis, D. Multilayer optical learning networks. *Appl. Opt.* **26**, 5061–5076 (1987). https://doi.org/10.1364/AO.26.005061.
12. Jutamulia, S. & Yu, F. Overview of hybrid optical neural networks. *Opt. Laser Technol.* **28**, 59–72 (1996). https://doi.org/10.1016/0030-3992(95)00070-4.
13. Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H. Englund, D., *et al.* Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441 (2017). https://doi.org/10.1038/nphoton.2017.93.
14. Tait, A. N., Lima, T. F., Zhou, E., Wu, A. X., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017). https://doi.org/10.1038/s41598-017-07754-z.
15. Lin, X., Rivenson, Y., Yardimci, N. T., Veli, M., Luo, Y., Jarrahi, M. & Ozcan, A. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004 (2018). https://doi.org/10.1126/science.aat8084.
16. Chang, J., Sitzmann, V., Dun, X., Heidrich, W. & Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018). https://doi.org/10.1038/s41598-018-30619-y.
17. Hughes, T. W., Minkov, M., Shi, Y. & Fan, S. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5**, 864–871 (2018). https://doi.org/10.1364/OPTICA.5.000864.
18. Bagherian, H., Skirlo, S., Shen, Y., Meng, H., Ceperic, V. & Soljacic, M.. *On-chip optical convolutional neural networks* (2018). https://doi.org/10.48550/arXiv.1808.03303.
19. Penkovsky, B., Porte, X., Jacquot, M., Larger, L. & Brunner, D.. *Coupled nonlinear delay systems as deep convolutional neural networks* (2019). https://doi.org/10.1103/PhysRevLett.123.054101.
20. Feldmann, J., Youngblood, N., Wright, C., Bhaskaran, H. & Pernice, W. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208 (2019). https://doi.org/10.1038/s41586-019-1157-8.
21. Harris, N. C., Steinbrecher, G. R., Prabhu, M., Lahini, Y., Mower, J., Bunandar, D., Chen, C., Wong, F. N., Baehr-Jones, T. Hochberg, M., *et al.* Quantum transport simulations in a programmable nanophotonic processor. *Nat. Photonics* **11**, 447 (2017). https://doi.org/10.1038/nphoton.2017.95.
22. Wang, H., He, Y., Li, Y.-H., Su, Z.-E., Li, B., Huang, H.-L., Ding, X., Chen, M.-C., Liu, C. Qin, J., *et al.* High-efficiency multiphoton boson sampling. *Nat. Photonics* **11**, 361 (2017). https://doi.org/10.1038/nphoton.2017.63.
23. Carolan, J., Harrold, C., Sparrow, C., Martín-López, E., Russell, N. J., Silverstone, J. W., Shadbolt, P. J., Matsuda, N., Oguma, M. Itoh, M., *et al.* Universal linear optics. *Science* **349**, 711–716 (2015). https://doi.org/10.1126/science.aab3642.
24. Spring, J. B., Metcalf, B. J., Humphreys, P. C., Kolthammer, W. S., Jin, X.-M., Barbieri, M., Datta, A., Thomas-Peter, N., Langford, N. K. Kundys, D., *et al.* Boson sampling on a photonic chip. *Science* **339**, 798–801 (2013). https://doi.org/10.1126/science.1231692.
25. Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994). https://doi.org/10.1103/PhysRevLett.73.58.
26. Clements, W. R., Humphreys, P. C., Metcalf, B. J., Kolthammer, W. S. & Walmsley, I. A. Optimal design for universal multiport interferometers. *Optica* **3**, 1460–1465 (2016). https://doi.org/10.1364/OPTICA.3.001460.
27. He, K., Zhang, X., Ren, S. & Sun, J.. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778) (2016). https://doi.org/10.1109/CVPR.2016.90.
28. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y.. Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680) (2014).
29. Freund, Y. & Schapire, R. E. Large margin classification using the perceptron algorithm. *Machine learning* **37**, 277–296 (1999). https://doi.org/10.1023/A:1007662407062.
30. Steinbrecher, G. R., Olson, J. P., Englund, D. & Carolan, J.. *Quantum optical neural networks* (2018). https://doi.org/10.1038/s41534-019-0174-7.
31. Nair, V. & Hinton, G. E.. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814) (2010).
32. Nguyen, H. V. & Bai, L.. Cosine similarity metric learning for face verification. In *Asian conference on computer vision* (p. 709). Springer (2010). https://doi.org/10.1007/978-3-642-19309-5_55.
33. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. & Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* **19**, 788 (2011). https://doi.org/10.1109/TASL.2010.2064307.
34. Rosenblatt, F.. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory (1957).
35. MacKay, D. J. & Kay, D. J. M.. *Information theory, inference and learning algorithms*. Cambridge university press (2003).
36. C. Blake, 1998, http://archive.ics.uci.edu.
37. Williamson, I. A. D., Hughes, T. W., Minkov, M., Bartlett, B., Pai, S. & Fan, S.. *Reprogrammable electro-optic nonlinear activation functions for optical neural networks* (2019). https://doi.org/10.1109/JSTQE.2019.2930455.
38. Connelly, M. J.. *Semiconductor optical amplifiers*. Springer Science & Business Media (2007).
39. Since the last four output paths are untraced, many parts of the circuit are redundant. We still keep them in our illustration, in order to provide a more clear picture of the general idea of the circuit.
40. Han, S., Mao, H. & Dally, W. J.. *Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding* (2015). https://doi.org/10.48550/arXiv.1510.00149.
41. Flamini, F., Spagnolo, N., Viggianiello, N., Crespi, A., Osellame, R. & Sciarrino, F. Benchmarking integrated linear-optical architectures for quantum information processing. *Scientific reports* **7**, 1–10 (2017). https://doi.org/10.1038/s41598-017-15174-2.
42. Jing, L., Shen, Y., Dubcek, T., Peurifoy, J., Skirlo, S., LeCun, Y., Tegmark, M. & Soljačić, M.. Tunable efficient unitary neural networks (eunn) and their application to rnns. In *International Conference on Machine Learning* (pp. 1733–1741) (2017).
43. Pai, S., Williamson, I., Hughes, T. W., Minkov, M., Solgaard, O., Fan, S. & Miller, D. A. Parallel programming of an arbitrary feedforward photonic network. *IEEE Journal of Selected Topics in Quantum Electronics* (2020). https://doi.org/10.1109/JSTQE.2020.2997849.
44. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks* **4**, 251–257 (1991). https://doi.org/10.1016/0893-6080(91)90009-T.
45. Vivien, L., Polzer, A., Marris-Morini, D., Osmond, J., Hartmann, J. M., Crozat, P., Cassan, E., Kopp, C., Zimmermann, H. & Fédéli, J. M. Zero-bias 40gbit/s germanium waveguide photodetector on silicon. *Optics express* **20**, 1096–1101 (2012). https://doi.org/10.1364/OE.20.001096.
46. Xiang, J., Torchy, A., Guo, X. & Su, Y. All-optical spiking neuron based on passive microresonator. *Journal of Lightwave Technology* (2020). https://doi.org/10.1109/JLT.2020.2986233.
47. Zuo, Y., Li, B., Zhao, Y., Jiang, Y., Chen, Y.-C., Chen, P., Jo, G.-B., Liu, J. & Du, S. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019). https://doi.org/10.1364/OPTICA.6.001132.

MISCELLANEA

Supplementary material Supplementary material associated with this article can be found, in the online version, at 10.1016/j.chip.2021.100002.

Acknowledgments X.-M. Z. thanks Xi-Ming Wang for helpful discussion. This work is supported by the Natural Science Foundation of Guangdong Province (Grant No. 2017B030308003), the Key R&D Program of Guangdong province (Grant No. 2018B030326001), the Science,Technology and Innovation Commission of Shenzhen Municipality (Grant No. JCYJ20170412152620376 and No. JCYJ20170817105046702 and No. KYTDPT20181011104202253), National Natural Science Foundation of China (Grant No. 11875160 and No. U1801661), the Economy, Trade and Information Commission of Shenzhen Municipality (Grant No. 201901161512), and Guangdong Provincial Key Laboratory (Grant No. 2019B121203002).

Declaration of Competing Interest The authors declare that they have no conflict of interest.