nature photonics

ARTICLES

PUBLISHED ONLINE: 12 JUNE 2017 | DOI: 10.1038/NPHOTON.2017.93

Deep learning with coherent nanophotonic circuits

Yichen Shen[1*†], Nicholas C. Harris[1*†], Scott Skirlo[1], Mihika Prabhu[1], Tom Baehr-Jones[2], Michael Hochberg[2], Xin Sun[3], Shijie Zhao[4], Hugo Larochelle[5], Dirk Englund[1] and Marin Soljačić[1]

Artificial neural networks are computational network models inspired by signal processing in the brain. These models have dramatically improved performance for many machine-learning tasks, including speech and image recognition. However, today's computing hardware is inefficient at implementing neural networks, in large part because much of it was designed for von Neumann computing schemes. Significant effort has been made towards developing electronic architectures tuned to implement artificial neural networks that exhibit improved computational speed and accuracy. Here, we propose a new architecture for a fully optical neural network that, in principle, could offer an enhancement in computational speed and power efficiency over state-of-the-art electronics for conventional inference tasks. We experimentally demonstrate the essential part of the concept using a programmable nanophotonic processor featuring a cascaded array of 56 programmable Mach–Zehnder interferometers in a silicon photonic integrated circuit and show its utility for vowel recognition.

Computers that can learn, combine and analyse vast amounts of information quickly, efficiently and without the need for explicit instructions are emerging as a powerful tool for handling large data sets. 'Deep learning' algorithms have received an explosion of interest in both academia and industry for their utility in image recognition, language translation, decision-making problems and more[1–4]. Traditional central processing units (CPUs) are suboptimal for implementing these algorithms[5], and a growing effort in academia and industry has been directed towards the development of new hardware architectures tailored to applications in artificial neural networks (ANNs) and deep learning[6]. Graphical processing units (GPUs), application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs)[2,5,7–11], including IBM TrueNorth[5] and Google TPU[11], have improved both energy efficiency and speed enhancement for learning tasks. In parallel, hybrid optical–electronic systems that implement spike processing[12–14] and reservoir computing[15–18] have been demonstrated.

Fully optical neural networks (ONNs) offer a promising alternative approach to microelectronic and hybrid optical–electronic implementations. ANNs are a promising fully optical computing paradigm for several reasons. (1) They rely heavily on fixed matrix multiplications. Linear transformations (and certain nonlinear transformations) can be performed at the speed of light and detected at rates exceeding 100 GHz (ref. 19) in photonic networks and, in some cases, with minimal power consumption[20,21]. For example, it is well known that a common lens performs a Fourier transform without any power consumption and that certain matrix operations can also be performed optically without consuming power. (2) They have weak requirements on nonlinearities. Indeed, many inherent optical nonlinearities can be directly used to implement nonlinear operations in ONNs. (3) Once a neural network is trained, the architecture can be passive, and computation on the optical signals will be performed without additional energy input. These features could enable ONNs that are substantially more energy-efficient and faster than their electronic counterparts. However, implementing such transformations with bulk optical components (such as fibres and lenses) has been a major barrier so far because of the need for phase stability and large neuron counts[22]. Integrated photonics addresses this problem by providing a scalable solution to large, phase-stable optical transformations[23].

Here, we begin with a theoretical proposal for a fully optical architecture for implementing general deep neural network algorithms using nanophotonic circuits that process coherent light. The speed and power efficiency of our proposed architecture is largely enabled by coherent, fully optical matrix multiplication (a cornerstone of neural network algorithms). Under the assumption of practical, centimetre-scale silicon photonic die sizes and low waveguide losses, we estimate that such an ONN would enable forward propagation that is at least two orders of magnitude faster than state-of-the-art electronic or hybrid optical–electronic systems, and with a power consumption that is nearly proportional (instead of quadratic, as in electronics) to the number of neurons (for more details see the discussion about scaling in the Methods). Next, we experimentally demonstrate the essential component of our scheme by embedding our proposed optical interference unit (OIU) and diagonal matrix multiplication unit within a subset of the programmable nanophotonic processor (PNP), a photonic integrated circuit developed for applications in quantum information processing[23]. To test the practical performance of our theoretical proposal, we benchmarked the PNP on a vowel recognition problem, which achieved an accuracy comparable to a conventional 64-bit computer using a fully connected neural network algorithm.

ONN device architecture

An ANN[1] consists of a set of input artificial neurons (represented as circles in Fig. 1a) connected to at least one hidden layer and the output layer. In each layer (depicted in Fig. 1b), information propagates by a linear combination (for example, matrix multiplication) followed by the application of a nonlinear activation function. ANNs can be trained by feeding training data into the input layer and then computing the output by forward propagation; matrix entries (weights) are subsequently optimized using back propagation[24].

The ONN architecture is depicted in Fig. 1b,c. As shown in Fig. 1c, the task (an image, a vowel or a sentence to be recognized)

[1]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Elenion, 171 Madison Avenue, Suite 1100, New York, New York 10016, USA. [3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [5]Université de Sherbrooke, Administration, 2500 Boulevard de l'Université, Sherbrooke, Quebec J1K 2R1, Canada. †These authors contributed equally to this work. *e-mail: ycshen@mit.edu; n_h@mit.edu

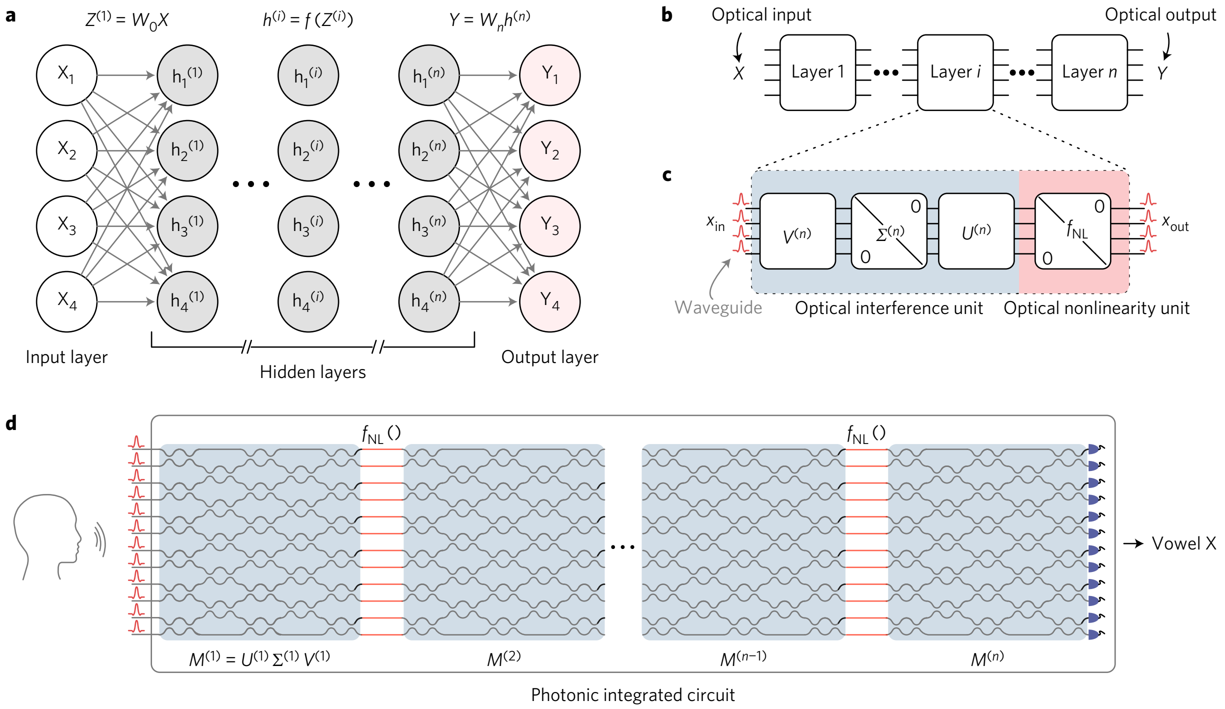

Figure 1 | General architecture of the ONN. a, General artificial neural network architecture composed of an input layer, a number of hidden layers and an output layer. **b**, Decomposition of the general neural network into individual layers. **c**, Optical interference and nonlinearity units that compose each layer of the artificial neural network. **d**, Proposal for an all-optical, fully integrated neural network.

is first preprocessed to a high-dimensional vector on a computer with a standard algorithm (this step is computationally inexpensive compared with inference). The preprocessed signals are then encoded in the amplitude of optical pulses propagating in the photonic integrated circuit, which implements a many-layer ONN. Each layer of the ONN is composed of an OIU that implements optical matrix multiplication and an optical nonlinearity unit (ONU) that implements the nonlinear activation. In principle, the ONN can implement an ANN of arbitrary depth and dimensions fully in the optical domain.

To realize an OIU that can implement any real-valued matrix, we first note that a general, real-valued matrix (M) may be decomposed as $M = U\Sigma V^{\dagger}$ through singular value decomposition (SVD)[25], where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal and $V^{\dagger}$ is the complex conjugate of the $n \times n$ unitary matrix V. It has been shown theoretically that any unitary transformations $U, V^{\dagger}$ can be implemented with optical beamsplitters and phase shifters[26,27]. Finally, Σ can be implemented using optical attenuators—optical amplification materials such as semiconductors or dyes could also be used[28]. Matrix multiplication with unitary matrices implemented in the manner above consumes, in principle, no power. The fact that a major proportion of ANN calculations involve matrix products enables the extreme energy efficiency of the ONN architecture presented here.

The ONU can be implemented using common optical nonlinearities such as saturable absorption[29–31] and bistability[32–36], which have all been demonstrated previously in photonic circuits. For an input intensity I_{in}, the optical output intensity is given by a nonlinear function $I_{out} = f(I_{in})$[37]. In this Article, we will consider an f that models the mathematical function associated with a realistic saturable absorber (such as a dye, semiconductor or graphene saturable absorber or saturable amplifier) that could, in future implementations, be directly integrated into waveguides after each OIU stage of the circuit. For example, graphene layers integrated on nanophotonic waveguides have already been demonstrated as saturable absorbers[38]. Saturable absorption is modelled as[29] (Supplementary Section 2)

$$\sigma\tau_s I_0 = \frac{1}{2}\frac{\ln(T_m/T_0)}{1 - T_m} \qquad (1)$$

where σ is the absorption cross-section, τ_s is the radiative lifetime of the absorber material, T_0 is the initial transmittance (a constant that only depends on the design of the saturable absorbers), I_0 is the incident intensity and T_m is the transmittance of the absorber. Given an input intensity I_0, one can solve for $T_m(I_0)$ from equation (1) and the output intensity can be calculated as $I_{out} = I_0 \cdot T_m(I_0)$. A plot of the saturable absorber's response function $I_{out}(I_{in})$ is shown in Supplementary Section 2.

A schematic diagram of the proposed fully optical neural network is shown in Fig. 1d.

Experiment

We evaluated the practicality of our proposal by experimentally implementing a two-layer neural network trained for vowel recognition. To prepare the training and testing data sets, we used 360 data points, each consisting of four log area ratio coefficients[39] of one phoneme. The log area ratio coefficients, or feature vectors, represent the power contained in different logarithmically spaced frequency bands and are derived by computing the Fourier transform of the voice signal multiplied by a Hamming window function. The 360 data points were generated by 90 different people speaking four different vowel phonemes[40]. We used half of these data points for training and the remaining half to test the performance of the trained ONN. We trained the matrix parameters used in the ONN with the standard back-propagation algorithm using a stochastic gradient descent method[41] on a conventional computer. Further details on the data set and back-propagation procedure are included in Supplementary Section 3.

The OIU was implemented using a PNP[23]—a silicon photonic integrated circuit fabricated in the OPSIS foundry[42]. This was

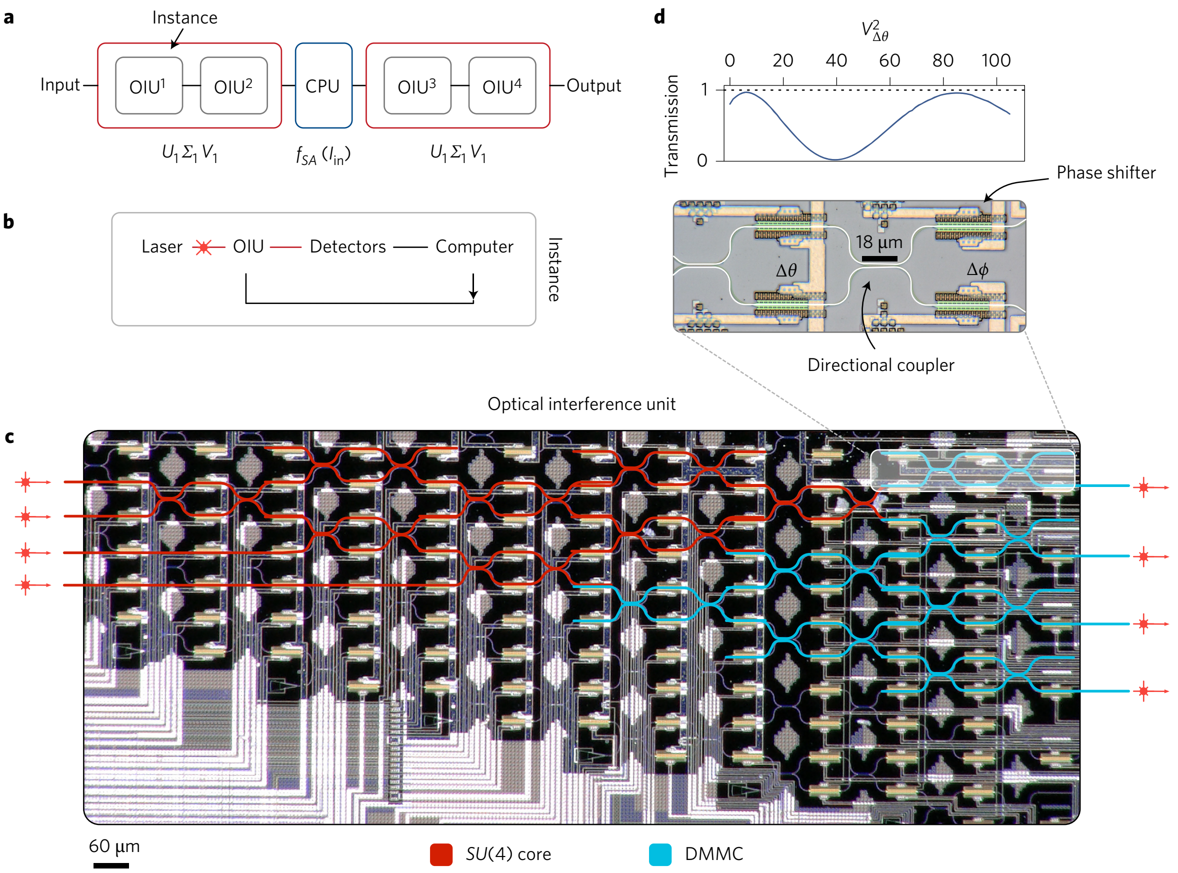

Figure 2 | Illustration of OIU. a, Schematic representation of our two-layer ONN experiment. The programmable nanophotonic processor is used four times to implement the deep neural network protocol. After the first matrix is implemented, a nonlinearity associated with a saturable absorber is simulated in response to the output of layer 1. **b**, Experimental feedback and control loop used in the experiment. Laser light is coupled to the OIU, transformed, measured on a photodiode array, and then read on a computer. **c**, Optical micrograph illustration of the experimentally demonstrated OIU, which realizes both matrix multiplication (highlighted in red) and attenuation (highlighted in blue) fully optically. The spatial layout of MZIs follows the Reck proposal[27], enabling arbitrary *SU*(4) rotations by programming the internal and external phase shifters of each MZI (θ_i, ϕ_i). **d**, Schematic illustration of a single phase shifter in the MZI and the transmission curve for tuning the internal phase shifter. DMMC, diagonal matrix multiplication core.

composed of 56 programmable Mach–Zehnder interferometers (MZIs), each of which comprised a thermo-optic phase shifter[43] (θ) between two 50% evanescent directional couplers, followed by another phase shifter (ϕ). The MZI splitting ratio was controlled with an internal phase shifter (Fig. 2d) and the differential output phase was controlled with the external phase shifter.

Unitary matrices with rank N (including U,V considered here) can be decomposed into sets of $SU(2)$ rotations implemented by cascaded programmable MZIs[27] (Supplementary Section 2). The highlighted region of the circuit in Fig. 2c, in particular, implements an arbitrary $SU(4)$ transformation.

Diagonal matrices of dimension N, such as Σ, can be implemented as shown by the blue-highlighted region of the PNP in Fig. 2c. Each of the four output ports of the $SU(4)$ core is coupled to an MZI, which can be programmed to rotate light to an untracked mode. Each entry of Σ is programmed by solving for θ, which gives $\Sigma_{ii} = \sin(\theta_i/2)$.

We used four instances of the OIU to realize the following matrix transformations in the spatial-mode basis: (1) $U^{(1)}\Sigma^{(1)}$, (2) $V^{(1)}$, (3) $U^{(2)}\Sigma^{(2)}$ and (4) $V^{(2)}$. Transformations (1) and (2) realize the first matrix $M^{(1)}$, and (3) and (4) implement $\boldsymbol{M}^{(2)}$. Given the number of MZI columns we have currently, we could only implement U and Σ on a single pass through the chip. With a larger or specially purposed chip, one could straightforwardly include the full matrix decomposition. The measured fidelity (defined in equation (2) in the Methods) for the 720 OIU instances used in the experiment was 99.8 ± 0.003, corresponding to ~2.24% measurement uncertainty for each output port (see Methods for further details).

As shown in Fig. 2a,b, we reprogrammed the PNP to realize all of the required OIUs and simulated the nonlinear transfer function of a saturable absorber (equation (1)) on a computer. This proof-of-concept demonstration required photodetection and re-injection of light into the PNP modes between the layers of the neural network. However, given the compactness of the required section of the PNP, all five layers of the PNP could be integrated on a chip less than a centimetre in length.

After programming the nanophotonic processor to implement our ONN architecture, which consisted of four layers of OIUs with four neurons in each layer (requiring training of a total of $4 \times 6 \times 2 = 48$ phase shifter settings), we evaluated it on the vowel recognition test set. Our ONN correctly identified 138/180 cases (76.7%, see Methods for more details on the repeatability of measurement), compared to the correctness of 165/180 (91.7%) computed with a conventional 64-bit digital computer. The difference between the ONN and the digital computed results is mainly caused by the difference in their computational resolution. As can be seen from Fig. 3a,b, both systems are good at classifying vowels A and B, but even the 64-bit computer had some difficulty classifying C and D, showing that these two vowels are relatively close in the parameter space we used (Fig. 3d). As a result, our ONN has even more misclassification on these two vowels due to its limited resolution.

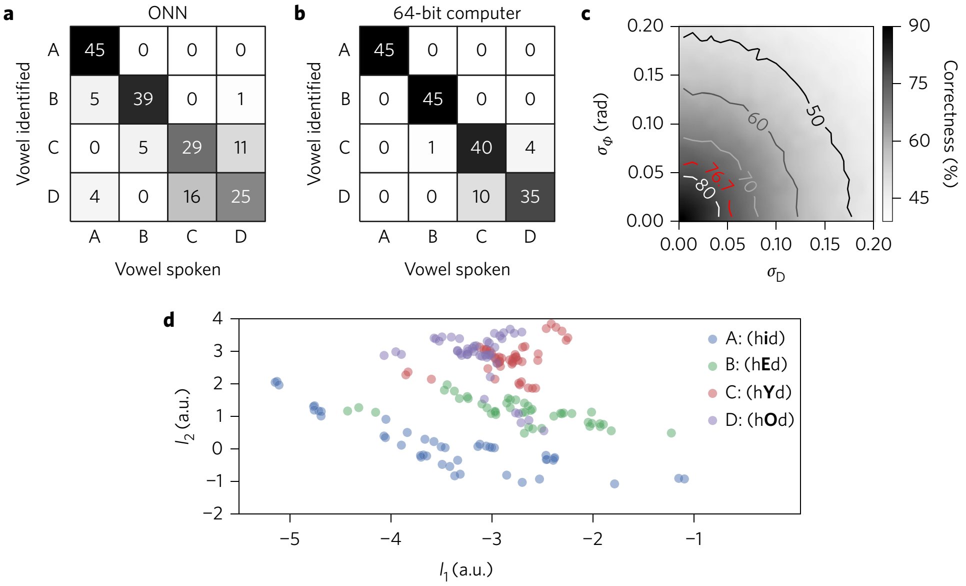

Figure 3 | Vowel recognition. a,**b**, Correlation matrices for the ONN and a 64-bit electronic computer, respectively, implementing two-layer neural networks for vowel recognition. Each row of the correlation matrices is a histogram of the number of times the ONN or 64-bit computer identified vowel X when presented with vowel Y. Perfect performance for the vowel recognition task would result in a diagonal correlation matrix. **c**, Correct identification ratio in percent for the vowel recognition problem with phase-encoding (σ_Φ) and photodetection error (σ_D). The definitions of these two variables are provided in the Methods. Solid lines are contours for different correctness ratios. In our experiment, $\sigma_D \simeq 0.1\%$. The contour line shown in red marks an isoline corresponding to the correct identification ratio for our experiment. **d**, Two-dimensional projection (log area ratio coefficient 1 on the *x* axis and 2 on the *y* axis) of the testing data set, which shows the large overlap between spoken vowel C and D. This large overlap leads to lower classification accuracy for both a 64-bit computer and the experimental ONN.

Discussion

Resolution analysis. As with digital floating-point computations, values are represented to a number of bits of precision. The finite dynamic range and noise in the optical intensities result in effective truncation errors in our ONN. The computational resolution of ONNs is limited by practical non-idealities, including (1) thermal crosstalk between phase shifters in interferometers, (2) optical coupling drift, (3) the finite precision with which an optical phase can be set (16 bits in our case), (4) photodetection noise and (5) finite photodetection dynamic range (30 dB in our case). Photodetection and phase encoding are the dominant sources of error in our experimental set-up. A detailed analysis of finite precision and low-flux photon shot noise is provided in Supplementary Section 1.

To understand the role of phase-encoding noise and photodetection noise in our ONN hardware architecture, we numerically simulated the performance of our trained matrices with varying degrees of phase-encoding noise (σ_Φ) and normalized photodetection noise (σ_D) (for detailed simulation steps see Methods). In this experiment, $\sigma_D \simeq 0.1\%$. The simulated distribution of correctness percentage versus σ_Φ and σ_D, plotted in Fig. 3c, indicates the tradeoff between encoding and photodetector noise that currently limits our experimental correctness to 76.7%. In experiments on individual MZIs, we obtained far lower noise values, $\sigma_\Phi \rightarrow 5 \times 10^{-3}$ (Supplementary Sections 1 and 7), which would result in a correctness of 90% (comparable to a digital computer at 91.7%). We attribute the additional excess noise in the full ONN to thermal crosstalk, which can be compensated in future experiments through additional calibration steps, or reduced altogether by adding thermal isolation trenches. Moreover, in a static ONN fabricated only for inference, thermal crosstalk would be eliminated. Even after these steps, any finite encoding error may still limit the effectiveness of the training step, that is, the optimality of the ONN parameters obtained by conventional back propagation. In these cases, generally slower, but more error-tolerant simulated annealing algorithms[44] could be used to train a more error-tolerant parameter set.

Computation speed and energy efficiency. Processing big data at high speeds and with low power is a central challenge in the field of computer science. Slow forward propagation and large power consumption limits the applications of ANNs in many fields, including self-driving cars, which require high speed and parallel image recognition.

Our ONN architecture takes advantage of high-detection-rate, high-sensitivity photon detectors to enable high-speed, energy-efficient neural networks compared to state-of-the-art electronic computer architectures. Once all parameters have been trained and programmed on the nanophotonic processor, forward-propagation computing is performed optically on a passive system. In our implementation, maintaining the phase modulator settings requires some (small) power of ~10 mW per modulator, on average. However, in future implementations, the phases could be set with nonvolatile phase-change materials[45], which would require no power to maintain. With this change, the total power consumption would be limited only by the physical size, the spectral bandwidth of dispersive components (THz) and the photodetection rate (100 GHz). In principle, such a system can be at least two orders of magnitude faster than electronic neural networks (which are restricted to a GHz clock rate). In addition, the ONN could have significantly lower latency (the time it takes from receiving input signals to computing an inference result) than electronic digital computers; this could be very useful for applications that require fast response times (such as autonomous driving or missile tracking). Assuming our ONN has N nodes, implementing m layers of $N \times N$ matrix multiplication and operating at a typical 100 GHz photodetection rate, these transformations correspond to 10^{11} N-dimensional matrix-vector multiplications in one second. Because the number of operations required to execute N-dimensional matrix-vector multiplications on a conventional digital computer scales as $O(N^2)$, the number of operations (floating point operations, or FLOPs) per second to match the optical network would be given by

$$R = 2m \times N^2 \times 10^{11} \text{FLOPs}$$

ONN power consumption during computation is dominated by the optical power necessary to trigger an optical nonlinearity and achieve a sufficiently high signal-to-noise ratio (SNR) at the photodetectors (assuming shot-noise-limited detection on n photons per pulse, SNR $\simeq \sqrt{1/n}$). We assume a saturable absorber threshold of $p \simeq 1$ MW cm^{-2}, which is valid for many dyes, semiconductors and graphene[29,30]. Because the cross-section for the waveguide is $A = 0.2 \times 0.5$ μm^2, the total power needed for forward propagation is estimated to be $P \approx N$ mW. Therefore, the energy per FLOP of the ONN will scale as $R/P = 2m \times N \times 10^{14}$ FLOPs J^{-1} (or $P/R = 5/mN$ fJ per FLOP). Almost the same energy performance and speed can be obtained if optical bistability[32,35,46] is used instead of saturable absorption as the enabling nonlinear phenomenon. Even for small ONNs, this power efficiency is already at least five orders of magnitude better than conventional GPUs, where $P/R \approx 100$ pJ per FLOP (shown in fig. 1.1.8 of ref. 47), or at least three orders of magnitude better than an 'ideal' (see Method for a detailed definition of 'ideal') electronic computer, where $P/R \approx 1$ pJ per FLOP assuming low-energy operations (by doing a 16 bit FLOP instead of the conventional 64 bit FLOP) and locality (no energy is used on data movement). Note that conventional image recognition tasks require tens of millions of training parameters and thousands of neurons ($mN \approx 10^5$) (ref. 4). These considerations suggest that the ONN approach may be far more efficient than conventional computers for standard problem sizes. In fact, the larger the neural network, the bigger the advantage of using optics: this arises largely from the fact that evaluating an $N \times N$ matrix in electronics requires $O(N^2)$ energy, whereas in optics it requires in principle no energy. Further details on power efficiency calculations are provided in Supplementary Section 3.

On-chip training. ONNs can also enable new ways to train ANN parameters. On a conventional computer, parameters are trained with back propagation and gradient descent. However, for certain ANNs where the effective number of parameters substantially exceeds the number of distinct parameters (including recurrent neural networks (RNNs) and convolutional neural networks (CNNs)), training using back propagation is notoriously inefficient. Specifically, the recurrent nature of RNNs makes them effectively an extremely deep ANN (depth = sequence length), while in CNNs the same weight parameters are used repeatedly in different parts of an image for extracting features. Here, we propose an alternative approach to directly obtain the gradient of each distinct parameter without back propagation, using forward propagation on an ONN and the finite difference method. It is well known that the gradient for a particular distinct weight parameter ΔW_{ij} in an ANN can be obtained with two forward-propagation steps that compute $J(W_{ij})$ and $J(W_{ij} + \delta_{ij})$, followed by the evaluation of $\Delta W_{ij} = (J(W_{ij} + \delta_{ij}) - J(W_{ij}))/\delta_{ij}$ (this step only takes two operations). On a conventional computer, this scheme is not favoured because forward propagation (evaluating $J(W)$) is computationally expensive. In an ONN, each forward-propagation step is computed in a constant time (limited by the photodetection rate, which can exceed 100 GHz)[19], with a power consumption that is only proportional to the number of neurons, making the above scheme tractable and capable of being executed at rates similar to or faster than conventional back propagation in some cases of interest (for example, very deep RNNs). Furthermore, with this on-chip training scheme, one can readily parameterize and train unitary matrices, an approach known to be particularly useful for deep neural networks[48]. As a proof of concept, we carried out the unitary-matrix-on-chip training scheme for our vowel recognition problem (Supplementary Section 4).

Scaling up the ONN. Regarding the physical size of the proposed ONN, current technologies should be capable of realizing ONNs exceeding the 1,000-neuron regime. Photonic circuits with up to 4,096 optical devices have been demonstrated[49]. Three-dimensional photonic integration could enable even larger ONNs by adding another (vertical) spatial degree of freedom[50]. Furthermore, by feeding in input signals (for example, an image) via multiple patches over time (instead of all at once)—an algorithm that has been increasingly adopted by the deep learning community[51]—the ONN should be able to realize much bigger effective neural networks with a relatively small number of physical neurons.

Conclusion

The proposed architecture could be applied to other ANN algorithms where matrix multiplications and nonlinear activations are heavily used, including CNNs and RNNs. Furthermore, the superior forward-propagation speed and power efficiency of our ONN can potentially allow the neural network to be trained directly on the photonics chip, using only forward propagation. Finally, it needs to be emphasized that another major portion of power dissipation in current neural network architectures is associated with data movement—an outstanding challenge that remains to be addressed. However, recent dramatic improvements in optical interconnects using integrated photonics technology has the potential to significantly reduce this energy cost[52]. Further integration of optical interconnects and optical computing units needs to be explored to realize the full advantage of all-optical neural networks.

Methods

Methods and any associated references are available in the online version of the paper.

Received 5 October 2016; accepted 12 May 2017;
published online 12 June 2017

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
2. Silver, D. *et al.* Mastering the game of go with deep neural networks and tree search. *Nature* **529,** 484–489 (2016).
3. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518,** 529–533 (2015).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Proc. NIPS* 1097–1105 (2012).
5. Esser, S. K. *et al.* Convolutional networks for fast, energy efficient neuromorphic computing. *Proc. Natl Acad. Sci. USA* **113,** 11441–11446 (2016).
6. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78,** 1629–1636 (1990).
7. Poon, C.-S. & Zhou, K. Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities. *Front. Neurosci.* **5,** 108 (2011).
8. Shafiee, A. *et al.* ISAAC: a convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars. *Proc. ISCA* **43,** 14–26 (2016).
9. Misra, J. & Saha, I. Artificial neural networks in hardware: a survey of two decades of progress. *Neurocomputing* **74,** 239–255 (2010).
10. Chen, Y. H., Krishna, T., Emer, J. S. & Sze, V. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE J. Solid-State Circuits* **52,** 127–138 (2017).
11. Graves, A. *et al.* Hybrid computing using a neural network with dynamic external memory. *Nature* **538,** 471–476 (2016).
12. Tait, A. N., Nahmias, M. A., Tian, Y., Shastri, B. J. & Prucnal, P. R. in *Nanophotonic Information Physics* (ed. Naruse, M.) 183–222 (Springer, 2014).
13. Tait, A. N., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Lightw. Technol.* **32,** 3427–3439 (2014).
14. Prucnal, P. R., Shastri, B. J., de Lima, T. F., Nahmias, M. A. & Tait, A. N. Recent progress in semiconductor excitable lasers for photonic spike processing. *Adv. Opt. Phot.* **8,** 228–299 (2016).
15. Vandoorne, K. *et al.* Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Commun.* **5,** 3541 (2014).
16. Appeltant, L. *et al.* Information processing using a single dynamical node as complex system. *Nat. Commun.* **2,** 468 (2011).
17. Larger, L. *et al.* Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing. *Opt. Express* **20,** 3241–3249 (2012).
18. Paquot, Y. *et al.* Optoelectronic reservoir computing. *Sci. Rep.* **2,** 287 (2011).
19. Vivien, L. *et al.* Zero-bias 40gbit/s germanium waveguide photodetector on silicon. *Opt. Express* **20,** 1096–1101 (2012).

20. Cardenas, J. *et al.* Low loss etchless silicon photonic waveguides. *Opt. Express* **17,** 4752–4757 (2009).
21. Yang, L., Zhang, L. & Ji, R. On-chip optical matrix-vector multiplier. In *SPIE Optical Engineering + Applications*, 88550F (International Society for Optics and Photonics, 2013).
22. Farhat, N. H., Psaltis, D., Prata, A. & Paek, E. Optical implementation of the Hopfield model. *Appl. Opt.* **24,** 1469–1475 (1985).
23. Harris, N. C. *et al.* Bosonic transport simulations in a large-scale programmable nanophotonic processor. Preprint at http://arXiv.org/abs/1507.03406 (2015).
24. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61,** 85–117 (2015).
25. Lawson, C. L. & Hanson, R. J. *Solving Least Squares Problems* Vol. 15 (SIAM, 1995).
26. Miller, D. A. B. Perfect optics with imperfect components. *Optica* **2,** 747–750 (2015).
27. Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73,** 58–61 (1994).
28. Connelly, M. J. *Semiconductor Optical Amplifiers* (Springer Science & Business Media, 2007).
29. Selden, A. Pulse transmission through a saturable absorber. *Br. J. Appl. Phys.* **18,** 743 (1967).
30. Bao, Q. *et al.* Monolayer graphene as a saturable absorber in a mode-locked laser. *Nano Res.* **4,** 297–307 (2010).
31. Schirmer, R. W. & Gaeta, A. L. Nonlinear mirror based on two-photon absorption. *J. Opt. Soc. Am. B* **14,** 2865–2868 (1997).
32. Soljačić, M., Ibanescu, M., Johnson, S. G., Fink, Y. & Joannopoulos, J. Optimal bistable switching in nonlinear photonic crystals. *Phys. Rev. E* **66,** 055601 (2002).
33. Xu, B. & Ming, N.-B. Experimental observations of bistability and instability in a two-dimensional nonlinear optical superlattice. *Phys. Rev. Lett.* **71,** 3959–3962 (1993).
34. Centeno, E. & Felbacq, D. Optical bistability infinite-size nonlinear bidimensional photonic crystals doped by a microcavity. *Phys. Rev. B* **62,** R7683–R7686 (2000).
35. Nozaki, K. *et al.* Sub-femtojoule all-optical switching using a photonic-crystal nanocavity. *Nat. Photon.* **4,** 477–483 (2010).
36. Ríos, C. *et al.* Integrated all-photonic non-volatile multilevel memory. *Nat. Photon.* **9,** 725–732 (2015).
37. Krizhevsky, A., Sutskever, I. & Hinton, G. E. in *Imagenet Classification with Deep Convolutional Neural Networks* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, 2012).
38. Cheng, Z., Tsang, H. K., Wang, X., Xu, K. & Xu, J.-B. In-plane optical absorption and free carrier absorption in graphene-on-silicon waveguides. *IEEE J. Sel. Top. Quantum Electron.* **20,** 43–48 (2014).
39. Chow, D. & Abdulla, W. H. in *PRICAI 2004: Trends in Artificial Intelligence* (eds Booth, R. & Zhang, M.-L.) 901–908 (Springer, 2004).
40. Deterding, D. H. *Speaker Normalisation for Automatic Speech Recognition*. PhD thesis, Univ. Cambridge (1990).
41. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313,** 504–507 (2006).
42. Baehr-Jones, T. *et al.* A 25 Gb/s silicon photonics platform. Preprint at http://arXiv.org/abs/1203.0767 (2012).
43. Harris, N. C. *et al.* Efficient, compact and low loss thermo-optic phase shifter in silicon. *Opt. Express* **22,** 10487–10493 (2014).
44. Bertsimas, D. & Nohadani, O. Robust optimization with simulated annealing. *J. Global Optim.* **48,** 323–334 (2010).
45. Wang, Q. *et al.* Optically reconfigurable metasurfaces and photonic devices based on phase change materials. *Nat. Photon.* **10,** 60–65 (2016).
46. Tanabe, T., Notomi, M., Mitsugi, S., Shinya, A. & Kuramochi, E. Fast bistable all-optical switch and memory on a silicon photonic crystal on-chip. *Opt. Lett.* **30,** 2575–2577 (2005).
47. Horowitz, M. Computing's energy problem. In *2014 IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers (ISSCC)* 10–14 (IEEE, 2014).
48. Arjovsky, M., Shah, A. & Bengio, Y. Unitary evolution recurrent neural networks. In *Int. Conf. Machine Learning* (2016).
49. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493,** 195–199 (2013).
50. Rechtsman, M. C. *et al.* Photonic Floquet topological insulators. *Nature* **496,** 196–200 (2013).
51. Jia, Y. *et al.* Caffe: convolutional architecture for fast feature embedding. In *Proc. 22nd ACM Int. Conf. Multimedia* (MM '14), 675–678 (ACM, 2014).
52. Sun, C. *et al.* Single-chip microprocessor that communicates directly using light. *Nature* **528,** 534–538 (2015).

Acknowledgements

The authors thank Y. LeCun, M. Tegmark, G. Pratt, I. Chuang and V. Sze for discussions. This work was supported in part by the Army Research Office through the Institute for Soldier Nanotechnologies under contract no. W911NF-13-D0001 and in part by the National Science Foundation under grant no. CCF-1640012 and in part by the Air Force Office of Scientific Research (AFOSR) Multidisciplinary University Research Initiative (FA9550-14-1-0052) and the Air Force Research Laboratory RITA programme (FA8750-14-2-0120). M.H. acknowledges support from AFOSR STTR grants, numbers FA9550-12-C-0079 and FA9550-12-C-0038 and G. Pomrenke, of AFOSR, for his support of the OpSIS effort, through both a PECASE award (FA9550-13-1-0027) and funding for OpSIS (FA9550-10-1-0439). N.H. acknowledges support from National Science Foundation Graduate Research Fellowship grant no. 1122374.

Author contributions

Y.S., N.C.H., S.S., X.S., S.Z., D.E. and M.S. developed the theoretical model for the optical neural network. N.H. designed the photonic chip and built the experimental set-up. N.H., Y.S. and M.P. performed the experiment. Y.S., S.S. and X.S. prepared the data and developed the code for training MZI parameters. T.B.-J. and M.H. fabricated the photonic integrated circuit. All authors contributed to writing the paper.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.S. and N.C.H.

Competing financial interests

The authors declare no competing financial interests.

Methods

Fidelity analysis. We evaluated the performance of the $SU(4)$ core with the fidelity metric

$$f = \sum_i \sqrt{p_i q_i} \qquad (2)$$

where p_i, q_i are experimental and simulated (on a 64-bit electronic computer) normalized ($\Sigma_i x_i = 1$ where $x \in \{p, q\}$) optical intensity distributions across the waveguide modes, respectively. Assuming $p_i = q_i + \varepsilon_i$, where $\varepsilon_i = q_i$ is the error (note $\Sigma_i \varepsilon_i = 0$), we have $f = \sum_i q_i \sqrt{(1 + (\varepsilon_i/q_i))} \approx \sum_i q_i (1 + (\varepsilon_i/2q_i) - (\varepsilon_i^2/4q_i^2)) = 1 + (1/2)\sum_i \varepsilon_i - (1/4)\sum_i (\varepsilon_i^2/q_i^2) = 1 - (1/4)\sum_i (\varepsilon_i^2/q_i^2)$. Given that $f = 0.998$, we have $|\sum_i (\varepsilon_i^2/q_i^2)| = 0.008$. Assuming $q_i \approx 0.25$ (uniform output power distribution), then $\sum_i \varepsilon_i^2 \approx 0.0005$. Finally, because our system performs the calculation in the amplitude of signal ($\sqrt{p_i}$ and $\sqrt{q_i}$), our experimental error for each output port of the $SU(4)$ core is on the order of

$$\delta = |\sqrt{p_i} - \sqrt{q_i}| \approx \sqrt{\frac{\varepsilon_i^2}{4q_i}} \approx 0.0112$$

and the percentage error for each output port is approximately $(\delta/\sqrt{q_i}) = 2.24\%$.

Simulation method for noise in ONN. We carried out the following steps to numerically simulate the performance of our trained matrices with varying degrees of phase encoding (σ_Φ) and detection (σ_D) noise:

(1) For each of the four trained 4 × 4 unitary matrices U^k, we calculate a set $\{\theta_i^k, \phi_i^k\}$ that encodes the matrix.
(2) We add a set of random phase-encoding errors, $\{\delta\theta_i^k, \delta\phi_i^k\}$ to the old calculated phases $\{\theta_i^k, \phi_i^k\}$, where we assume each $\delta\theta_i^k$ and $\delta\phi_i^k$ is a random variable sampled from a Gaussian distribution $G(\mu,\sigma)$ with $\mu = 0$ and $\sigma = \sigma_\Phi$. We obtain a new set of perturbed phases $\{\theta_i^{k\prime}, \phi_i^{k\prime}\} = \{\theta_i^k + \delta\theta_i^k, \phi_i^k + \delta\phi_i^k\}$.
(3) We encode the four perturbed 4 × 4 unitary matrices $U^{k'}$ based on the new perturbed phases $\{\theta_i^{k\prime}, \phi_i^{k\prime}\}$.
(4) We carry out the forward-propagation algorithm based on the perturbed matrices $U^{k'}$ with our test data set. During forward propagation, every time a matrix multiplication is performed (let us say when we compute $\vec{v} = U^{k\prime} \cdot \vec{u}$), we add a set of random photodetection errors $\vec{\delta v}$ to the resulting $\vec{v}$, where we assume each entry of $\vec{\delta v}$ is a random variable sampled from a Gaussian distribution $G(\mu,\sigma)$ with $\mu = 0$ and $\sigma = \sigma_D \cdot |v|$. We obtain the perturbed output vector $\vec{v}' = \vec{v} + \vec{\delta v}$.
(5) With the modified forward propagation scheme above, we calculate the correctness percentage for the perturbed ONN.
(6) Steps (2) to (5) are repeated 50 times to obtain the distribution of correctness percentage for each phase-encoding noise (σ_Φ) and photodetection noise (σ_D).

Scaling of power consumption. A shallow exponential scaling caused by propagation loss is applied to the power consumption estimation. Propagation through the chip is dominated by waveguide scattering losses, which have been shown experimentally to be as low as 0.3 dB cm^{-1} (ref. 20). Given this propagation loss, a single MZI would have a transmission of $(10^{-0.3/10})^{1/100} = 0.9993$, assuming that there are 100 MZIs per cm (that is, a conservative MZI length of 100 μm). In other words, if our neural network had 1,000 neurons (a useful size for commonly used neural networks), then the loss through the entire chip would be ~50%. This loss is negligible when the energy efficiency improvement for the ONN is orders of magnitude better than that of electronic neural networks. Furthermore, 0.3 dB cm^{-1} is not the fundamental limit of optical losses, and it can be improved by better fabrication. In summary, for mid-sized ONNs, the exponential loss due to optical scattering is negligible, so the linear approximation of energy consumption in the matrix multiplication applies.

Repeatability of measurement. To assess the repeatability of the experiment we carried out the entire testing run three times. The reported result (76.7% correctness percentage) is associated with the best calibrated run (highest measured fidelity). The other two less calibrated runs exceeded 70% correctness. For further discussion on enhancing the correctness percentage see Supplementary Section 6).

'Ideal' benchmark of electronic computing. To further illustrate this 'ideal' benchmark (1 pJ per FLOP), it is the minimum power needed for 45 nm technology-based electronic transistors to perform FLOPs (such as additions and multiplications). Note, here, that we do not count the power used for memory access and interconnects. Achieving this 1 pJ per FLOP performance requires a very specific combination of very low energy operations and extreme locality. Current machines (even GPUs) cannot yet do this, because they are designed to maximize performance and do not yet leverage locality as much as they could. For example, the best GPU now (NVIDIA TITAN X) has a power efficiency of 100 pJ per FLOP, which is still two orders of magnitude away from the efficiency limit we consider (1 pJ per FLOP).

Data availability. The data that support the plots within this paper and other findings of this study are available from the corresponding authors upon reasonable request.