nature photonics

ARTICLES

https://doi.org/10.1038/s41566-021-00796-w

Check for updates

Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit

Tiankuang Zhou[1,2,3,4], Xing Lin[1,2,5]✉, Jiamin Wu[1,2], Yitong Chen[1,2], Hao Xie[1,2], Yipeng Li[1,2], Jingtao Fan[1,2], Huaqiang Wu[5,6,7], Lu Fang[2,3,6]✉ and Qionghai Dai[1,2,6]✉

There is an ever-growing demand for artificial intelligence. Optical processors, which compute with photons instead of electrons, can fundamentally accelerate the development of artificial intelligence by offering substantially improved computing performance. There has been long-term interest in optically constructing the most widely used artificial-intelligence architecture, that is, artificial neural networks, to achieve brain-inspired information processing at the speed of light. However, owing to restrictions in design flexibility and the accumulation of system errors, existing processor architectures are not reconfigurable and have limited model complexity and experimental performance. Here, we propose the reconfigurable diffractive processing unit, an optoelectronic fused computing architecture based on the diffraction of light, which can support different neural networks and achieve a high model complexity with millions of neurons. Along with the developed adaptive training approach to circumvent system errors, we achieved excellent experimental accuracies for high-speed image and video recognition over benchmark datasets and a computing performance superior to that of cutting-edge electronic computing platforms.

Computing processors driven by electronics have evolved dramatically over the past decade, from general-purpose central processing units (CPUs)[1] to custom computing platforms, for example, graphics processing units (GPUs)[2], field-programmable gate arrays[3] and application-specific integrated circuits[4], to meet the globally increasing demand of computing resources. The progress of these silicon computing hardware platforms has contributed greatly to the resurgence of artificial intelligence (AI) by allowing the training of larger scale and more complicated models[5,6]. We have witnessed the extensive application of various neural computing architectures, for example, convolutional neural networks (CNNs)[2,6], recurrent neural networks (RNNs)[7], spiking neural networks[8] and reservoir computing[9], in a broad range of fields. However, electronic hardware implementations have reached unsustainable performance growth as the exponential scaling of electronic transistors embodied by Moore's law approaches its physical limit, where the speed and energy are fundamentally limited by parasitic capacitance, the tunnelling effect and crosstalk[10,11]. Therefore, despite the current dominance of electronic processors, the development of the next-generation computing modality is especially anticipated.

Optical computing uses photons instead of electrons for computation, and this process can overcome the inherent limitations of electronics and improve the energy efficiency, processing speed and computational throughput by orders of magnitude[12,13]. Such extraordinary properties have been exploited to build application-specific optical processors for solving fundamental mathematical[14–17] and signal processing[18,19] problems with performances far beyond those of existing electronic processors. In particular, artificial neural networks (ANNs) are one of the most promising optical computing models, where the neuron functionality and its dense interconnectivity can be effectively implemented with optoelectronic devices and the nature of light propagation[20–22]. Substantial recent progress has been made in optically accelerated neural information processing to accomplish some advanced AI tasks[23–38]. Nevertheless, existing optical AI accelerators can only support a single functionality that is customized for a specific neural network architecture or task and cannot be adapted to diverse AI algorithms for different tasks. In addition, the model complexity and experimental performance of current optical neural networks are much lower, resulting in a large gap in network performance (for example, classification accuracy) compared with that of state-of-the-art electronic ANNs. The reasons for this are due mainly to the limited flexibility of the network design space in optics; examples include the difficulty of integrating ideal non-linear operations and flexibly controlling the complex dataflow, and the imperfections of optical systems, which cause the deviation of a model and the accumulation of computing errors in practice.

Here, we propose a reconfigurable diffractive processing unit (DPU) for large-scale neuromorphic optoelectronic computing that can be programmed to change its functionality and construct different types of ANN architectures. The proposed optoelectronic processor combines the advantages of optics and electronics with the availability of extremely high data-throughput optical modulators and photodetectors; almost all of its computational operations are effectively implemented with optical diffraction, and its programmability is controlled electronically. We demonstrate the flexible design of complex diffractive neural networks (DNNs) with a DPU that can support millions of neurons. During the practical

[1]Department of Automation, Tsinghua University, Beijing, China. [2]Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing, China. [3]Department of Electronic Engineering, Tsinghua University, Beijing, China. [4]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. [5]Beijing Innovation Center for Future Chips, Tsinghua University, Beijing, China. [6]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. [7]Institute of Microelectronics, Tsinghua University, Beijing, China. ✉e-mail: lin-x@tsinghua.edu.cn; fanglu@tsinghua.edu.cn; qhdai@tsinghua.edu.cn

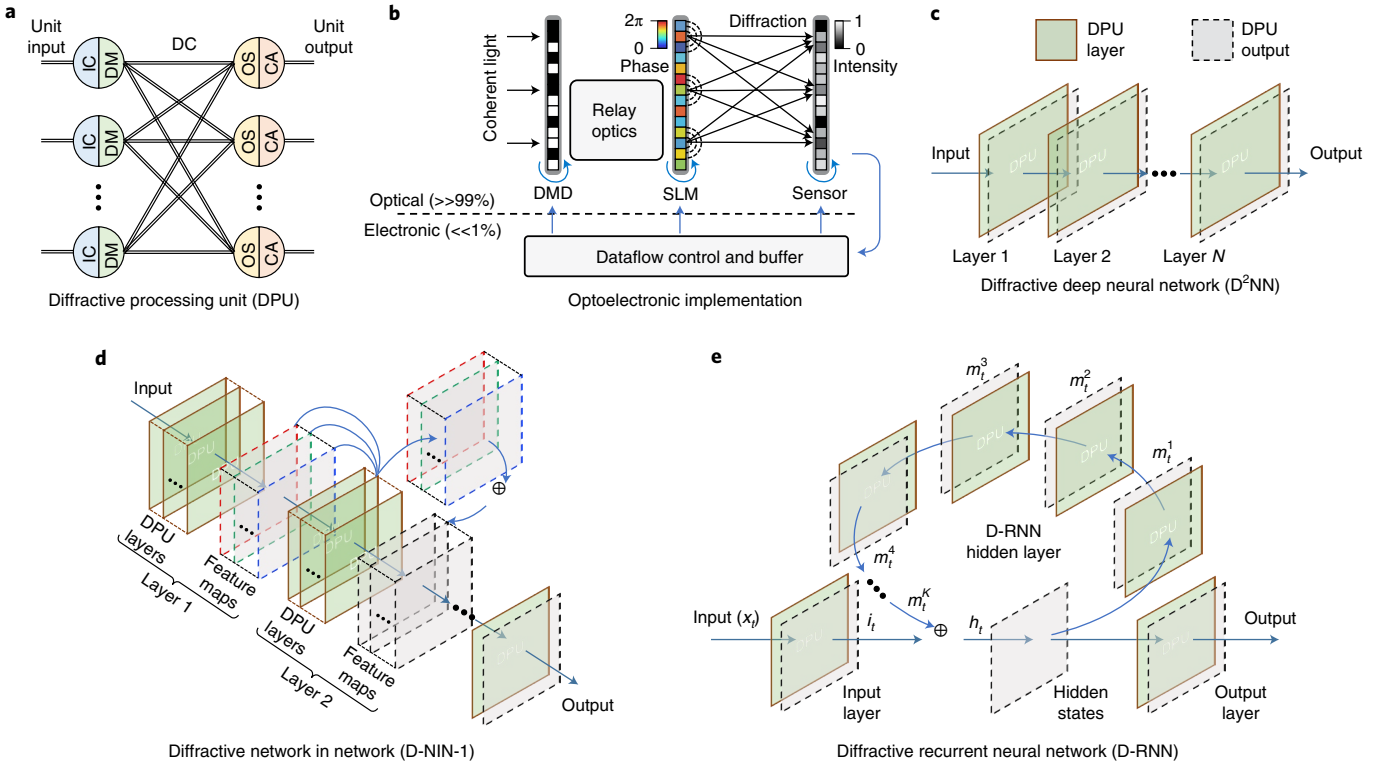

Fig. 1 | Reconfigurable diffractive optoelectronic processor. a, The DPU is a large-scale perceptron-like optoelectronic computing building block that can be programmed to construct different DNNs. IC, information coding; DM, diffractive modulation; DC, diffractive connection; OS, optical field summation; CA, complex activation. **b**, The programmable optoelectronic devices were adopted to implement the DPU, that is, a DMD, a phase SLM and a CMOS sensor, in this example. **c**–**e**, Three different types of neural network architectures were constructed, including the D²NN (refs. [27–30]) (**c**), D-NIN-1 (**d**) and D-RNN (**e**), where x_t, i_t, m_t and h_t correspond to the input sequence, input state, memory state (with K DPU layers) and hidden state, respectively, at the time step of t. ⊕ denotes the weighted summation.

implementation of the designed models with optoelectronic devices, we develop an adaptive training method to address the deviations of models caused by different error sources, such as alignment errors and non-ideal device characteristics, and achieve experimental results that demonstrate a high performance (see Supplementary Table 1).

Results

The principle and optoelectronic implementation of the proposed DPU are illustrated in Fig. 1a,b with forward model details in Supplementary Note 1. The DPU is an optoelectronic neuromorphic processor comprising large-scale diffractive neurons and weighted optical interconnections. It represents a fundamental building block that can be programmed for establishing various types of ANN with high model complexity and accuracy. The unit input data are quantized and electro-optically converted to a complex-valued optical field, encoded on the amplitude or phase component, with an information-coding module. Different input nodes are physically connected to individual output neurons with the light diffractive connections, where the synaptic weights that control the strength of the connections are determined by the diffractive modulation of the wavefront. Each diffractive optoelectronic neuron performs the optical field summation of its weighted inputs and generates the unit output by applying the complex activation function on the calculated optical field that occurs naturally during the photoelectric conversion.

We adopt programmable optoelectronic devices with a high data throughput, that is, on the order of a gigabit per second, to implement the DPU that allows high-speed neural network configurations and achieves video-rate inference capabilities (Fig. 1b). In this work, our system is designed to process large-scale visual signals that feed in images and videos. Thus, a digital micromirror device (DMD) and a spatial light modulator (SLM) are selected as the optical modulators to implement the input nodes, and a complementary metal–oxide–semiconductor (CMOS) sensor is used as the photodetector to implement the optoelectronic neurons. The schematic and experimental setup of our system are shown in Supplementary Fig. 1. By controlling and buffering the massively parallel optoelectronic dataflow of the unit, the DPU is allowed to be temporally multiplexed and programmed for customizing different types of ANN architectures (Fig. 1c–e). Since almost all of the computational operations are accomplished optically, the proposed optoelectronic AI architecture dramatically improves the computing speed and system energy efficiency compared with cutting-edge GPU processors (see Methods, Supplementary Note 2 and Supplementary Table 2).

We validate the effectiveness of the proposed adaptive training approach (see Methods, Supplementary Note 3 and Supplementary Fig. 2) and the functionality of the DPU by constructing a three-layer optoelectronic diffractive deep neural network[27–30], or D²NN, configured with 0.94 million optoelectronic neurons (see Methods), for classifying the MNIST (Modified National Institute of Standards and Technology) handwritten digits[39] (Fig. 2). The model was trained in silico using an MNIST training set and achieved a blind-testing accuracy of 97.6% on 10,000 digit images in a test set (see Methods and Supplementary Table 1). The training target was set to individually map input handwritten digits, from '0' to '9', into ten predefined regions on the output layer (that is, layer 3), where

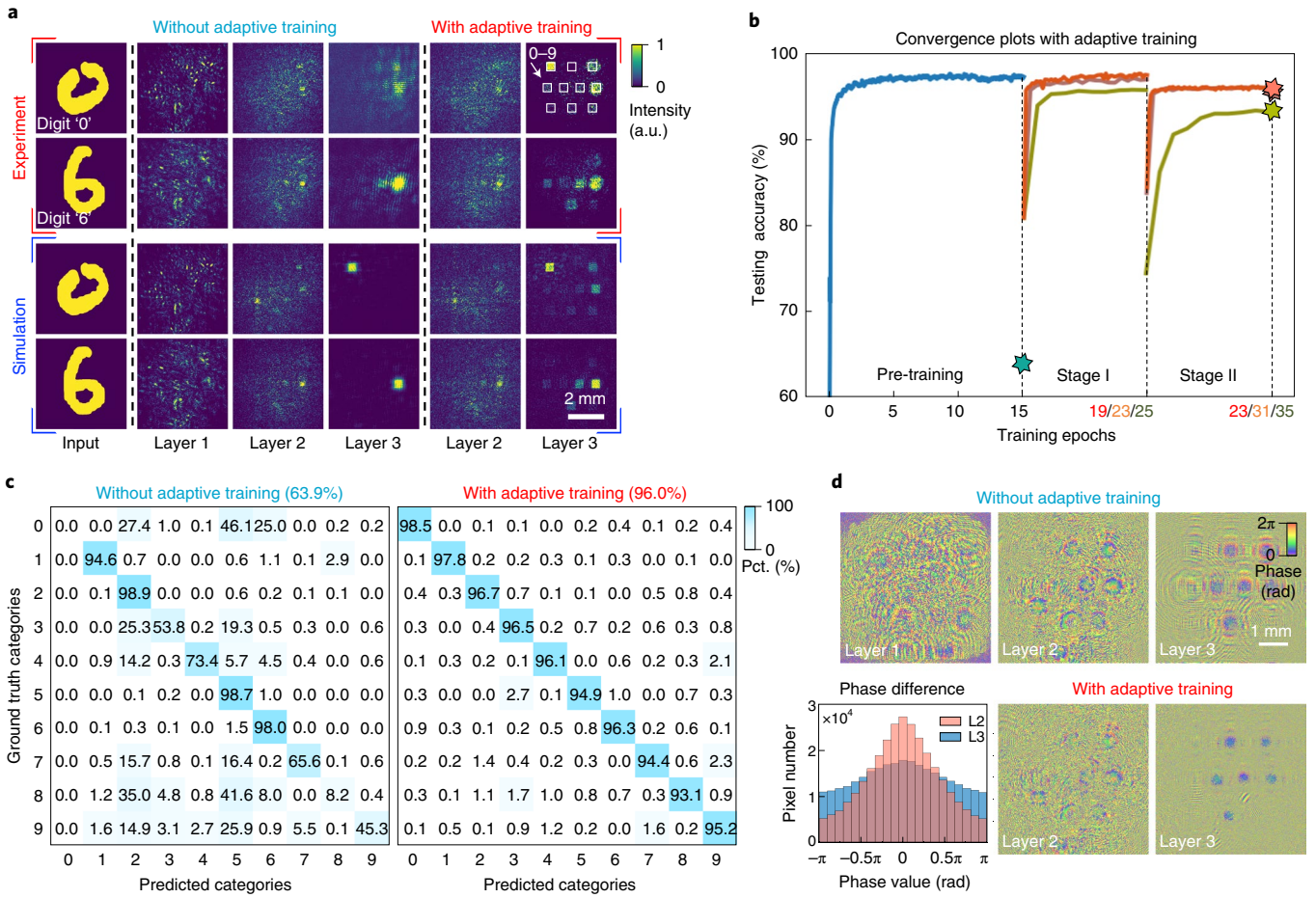

Fig. 2 | Adaptive training of the optoelectronic D²NN for handwritten digit classification. a, Experimental DPU outputs of a three-layer optoelectronic D²NN. The handwritten input digits (0–9) are classified by finding a maximum optical signal region among ten predefined regions at the output plane, each corresponding to one digit (from left to right, then top to bottom). **b**, Convergence plots of the optoelectronic D²NN evaluated on the blind-test dataset. The blue plot shows the pre-training process. Orange, brown and yellow plots represent the adaptive training with full, 20% and 2% training sets, respectively. **c**, Confusion matrices summarizing the experimental classification results of all instances in the test set. Pct., percentage. **d**, Final designs of three diffractive layers for high-accuracy handwritten digit classification before adaptive training (top) and after adaptive training (bottom middle and right) for layers 2 and 3, together with a histogram (bottom left) of the phase differences before and after adaptive training for layer 2 (L2) and layer 3 (L3).

a classification result was determined by finding the target region with the maximum optical signal (Fig. 2a). Without adaptive training, the direct transfer of the pre-trained model to our optoelectronic system dramatically decreased the recognition accuracy to 63.9% due to the accumulation of system errors layer by layer. As illustrated with the example digits '0' and '6' from a test set (Fig. 2a, left and middle), the layer error accumulation causes the intensity distribution of the DPU output at each layer to gradually deviate from that of the pre-trained model. This deviation reduces the percentage of energy focusing on the target region and thus may result in an incorrect recognition category, such as for the example digits '0'. The confusion matrix[27] (Fig. 2c, left) shows that transferring a direct model causes a large percentage of incorrect predictions and is particularly biased for digits such as '0', '3', '8' and '9', determined by the system imperfection state.

To circumvent the system error and improve the recognition performance, adaptive training of the constructed three-layer optoelectronic D²NN was implemented with two-stage fine tuning of the pre-trained model. Specifically, we had the option to trade-off between the experimental accuracy and training efficiency by using a full training set or a mini-training set (2% and 20% in this example, see Methods). With adaptive training, the intensity distributions of the DPU outputs between simulations and experiments are more matched, and both of the example testing digits '6' and '0' are correctly categorized during the experiments (Fig. 2a, right). The experimental demonstration can be found in Supplementary Video 1, where the handwritten digits were captured by a greyscale camera with their intensity inverted and thresholded to binary as the inputs to the DPU. More examples of correct D²NN testing results can be founded in Supplementary Video 2. The convergence plot in Fig. 2b shows that by refining the parameters of the pre-trained model with experimentally measured layer outputs, the first and second stages of adaptive training improve the system error-induced testing accuracy decrement with a full training set (orange plots), a 20% mini-training set (brown plots) or a 2% mini-training set (yellow plots). Despite the slightly lower accuracy, the mini-training set is more efficient with a shorter training time (see Methods). The experimental testing accuracy is improved from 63.9% (green star) to 96.0% (orange star) with the full training set, and to 95.4% (brown star) and 93.9% (yellow star) with the 20% and 2% mini-training sets, respectively. As shown by the confusion matrix in Fig. 2c, right, the correct prediction rates of the categories are improved

and are all larger than 93%. A histogram of the phase differences (Fig. 2d, bottom left) between the diffractive layers before (Fig. 2d, top) and after (Fig. 2d, bottom right) adaptive training reflects the fine-tuning process of the second and third diffractive layers for adapting the system error, where a large percentage of neurons have a small change in the phase modulation value.

In a CNN architecture (for example, LeNet[39]), segmenting the hidden layer into a set of feature maps with weight sharing is the critical mechanism that leads to high model performance. We demonstrate that the inference capability and model robustness of the optoelectronic D^{2}NN can be further enhanced by designing a multi-channel diffractive hidden layer as well as its external and internal interconnectivity structure (Figs. 1d and 3a). We call the constructed architecture a diffractive network in network, or D-NIN-1, as each of the diffractive feature maps is externally weighted and connected to all feature maps of the previous layer through the shared internal connectivities of the DPU layer (see Supplementary Note 4 for more analysis). Given the complex internal and external neuron connectivity structure used to compute more abstract features of each hidden layer, the network outputs of the D-NIN-1 are obtained after a DPU read-out layer for the final decision making.

We evaluate the performance of the D-NIN-1 by constructing a three-layer architecture, each layer with three diffractive feature maps, corresponding to 2.20 million optoelectronic neurons (see Methods), as shown in Fig. 3a. We demonstrate the superior model accuracy and robustness of the D-NIN-1 over the three-layer D^{2}NN (Fig. 2) on the task of MNIST classification. With the same DPU settings and in silico training procedures as with the D^{2}NN, the blind-testing model accuracy of the D-NIN-1 improves from 97.6 to 98.8%, surpassing the electronic CNN architecture LeNet-1. In addition, the model performance of the D-NIN-1 can be further boosted as D-NIN-1++, achieving an accuracy of 99.0%, which surpasses LeNet-4 (89.9%) and approaches LeNet-5 (99.2%) by integrating a low-complexity electronic fully connected layer at the end of the D-NIN-1 (see Methods). The learning capability of the D-NIN-1 was also evaluated with a more challenging database, that is, Fashion-MNIST[40], by encoding the greyscale unit input into the phase of input optical fields. Owing to the higher strength of non-linearity by converting from unit intensity measurement to the input phase modulation, the D-NIN-1 itself achieves a model accuracy of 90.2%, which surpasses LeNet-4 (98.9%) and is comparable to LeNet-5 (90.3%). The comparison of model convergence plots between the D-NIN-1 and LeNet on the MNIST and Fashion-MNIST databases are plotted in Fig. 3b and Supplementary Fig. 3, respectively.

With the programming of the optoelectronic DPU system to deploy the D-NIN-1 model, the experimental classification accuracy over the whole MNIST test dataset reaches a blind-testing accuracy of 96.6% after adaptive training. The confusion matrix in Fig. 3c (left) shows that the D-NIN-1 achieves the percentage of correct prediction larger than 94% for all of the categories. During the adaptive training, since the use of a DMD for amplitude encoding can provide a higher contrast and a more precise input optical field than the SLM for phase encoding, the Fashion-MNIST classification has a larger discrepancy between the model and experimental accuracy than the MNIST classification. The experimental blind-testing accuracy of the D-NIN-1 on the Fashion-MNIST is 84.6%, with the confusion matrix shown in Fig. 3c (right). For the constructed three-layer D-NIN-1 on the MNIST and Fashion-MNIST classifications, the energy distribution of ten predefined detection regions on the output layer of the inference results are calculated in Supplementary Fig. 4a,b, and the adaptive trained phase modulation coefficients at each DPU layer are visualized in Supplementary Fig. 4c,d. The corresponding phase-encoded unit inputs of the example product 'trousers' and output feature maps of example digit '3' (misclassified as digit '8' in the D^{2}NN) at each layer and their successful classification results are shown in the upper and lower rows of Fig. 3d, respectively. More examples of correct D-NIN-1 testing results on the MNIST database can be founded in Supplementary Video 3.

In addition to still images, the reconfigurability of the DPU allows us to construct a large-scale diffractive recurrent neural network, or D-RNN, to perform high-accuracy recognition tasks for video sequences. To demonstrate its functionality, we configure a standard RNN architecture based on the recurrent connections of DPU layers and apply it for the important task of video-based human action recognition[31]. The folded and unfolded representations of the proposed D-RNN are shown in Figs. 1e and 4a, respectively, and comprise the temporal sequential connections of the input, hidden and output layers with the shared diffractive parameters in time (see Methods for detailed formulation of the model). The memory of input sequences is formed by generating the diffractive hidden states at different time steps. The last hidden state of the D-RNN, summarizing the memory of input sequences, is extracted and read out with the DPU or electronic output layer to generate the categorical output distribution for determining the action categories.

The constructed D-RNN for the task of human action recognition was evaluated by using the metrics of sequence accuracy and action accuracy on two benchmark databases, that is, the Weizmann[41] and KTH[42] databases, with preprocessing to adapt to the network input (see Methods). After the ablation and performance analyses, the network sequence lengths were set to 3 and 5 for the Weizmann and KTH databases, respectively, with 1.47 million optoelectronic neurons (see Methods). The D-RNN architecture was first evaluated by configuring the DPU read-out layer, and then it was pre-trained in silico with the optimal fusing coefficient of 0.2 for both the Weizmann and KTH databases; it achieved a blind-testing sequence accuracy of 88.9 and 90.5% for the two databases, corresponding to an action accuracy of 100 and 96.3% for the two models, respectively. To implement the model experimentally, we performed adaptive training by fine tuning the modulation coefficients of only the read-out layer due to the recurrent connection inherence of the D-RNN. The designed modulation coefficients of the memory, read-in and read-out DPU layers after adaptive training are shown in Fig. 4b, where the upper and lower rows correspond to the models of the Weizmann and KTH databases, respectively. Compared with the experimental results without adaptive training, the adaptive training improved the experimental sequence accuracy from 51.0 to 82.9% and the experimental action accuracy from 56.7 to 96.7% for the Weizmann database. Similarly, after the adaptive training, the experimental sequence and action accuracies improved from 52.6 to 85.4% and from 53.7 to 94.4% for the KTH database, respectively.

We visualize the experimental testing results of all sub-sequences with the categorical voting matrix in Fig. 4c by calculating the percentage of votes for all categories in each testing video sequence, where the category with the maximum percentage of the vote represents the predicted category of a video sequence. The target testing video sequences were ranked in order with respect to the video categories so that the diagonal positions of the two categorical voting matrices represent the correct predictions. The experimental results show the miscategorization of one video sequence (that is, 25th) and three video sequences (that is, the 10th, 11th and 15th), marked with white arrows, for the Weizmann (Fig. 4c, left) and KTH (Fig. 4c, right) databases, respectively. Two example testing results are shown in Fig. 4e, including the correct categorized actions of 'jack' from the Weizmann database (Fig. 4e, left) and 'walking' from the KTH database (Fig. 4e, right). More examples of correct sub-sequence testing results from all actions of the Weizmann and KTH databases can be founded in Supplementary Videos 4 and 5, respectively.

The recognition accuracy and robustness of the D-RNN can be further enhanced, forming the D-RNN++ architecture, by transferring the trained D-RNN hidden layer and using a low-complexity

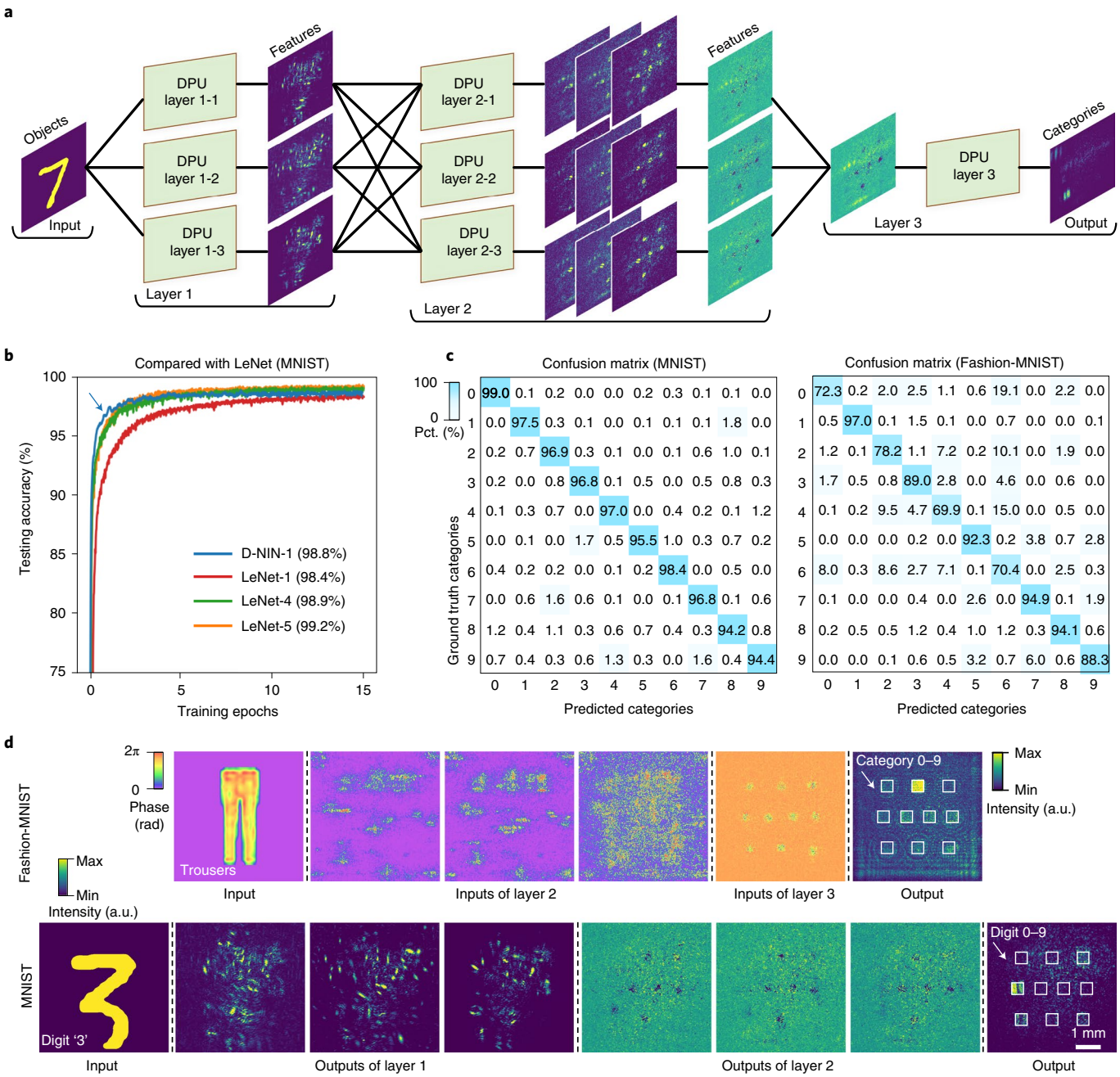

Fig. 3 | High-accuracy MNIST and Fashion-MNIST classification with the D-NIN-1. a, Configuring a three-layer D-NIN-1 with DPUs. The D-NIN-1 is a hierarchical interconnectivity architecture that comprises the internal and external connections of diffractive neurons. **b**, Convergence plots and comparisons of model accuracy between D-NIN-1 and LeNet[45]. The D-NIN-1 and its enhanced version (D-NIN-1++) achieve competitive classification accuracies, that is, 98.8 and 99.0%, respectively, with respect to the electronic CNNs on the MNIST dataset. **c**, Experimental confusion matrices of the MNIST and Fashion-MNIST classifications. The optoelectronic system configured with the D-NIN-1 model achieved a recognition accuracy of 96.6 and 84.6% for the MNIST and Fashion-MNIST test sets, respectively. For the Fashion-MNIST database, the category labels from 0 to 9 correspond to the fashion product types of t-shirt, trousers, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot, respectively. Pct., percentage. **d**, The layer inputs and outputs of example product 'trousers' and digit '3'.

electronic read-out layer to replace the DPU read-out layer (Fig. 4a, right; see Methods). We evaluate the experimental sequence and action accuracies of the D-RNN++ with respect to the number of read-out nodes, as shown in Fig. 4d. The results demonstrate that the D-RNN++ achieves an experimental action accuracy of 100 and 96.3% under the optimized number of electronic read-out nodes of 2,500 for the Weizmann and KTH (first scene) databases, respectively. Furthermore, the experimental action accuracy of the D-RNN++ for categorizing the Weizmann and KTH (first scene) databases achieves a comparable performance to and even outperforms the state-of-the-art electronic computing approaches[31,43,44] that have reported accuracies of 100 and 96.0%, respectively.

Discussion

The DPU system in this work was configured with an 8-bit phase modulation accuracy and a 16-bit accuracy of optical field

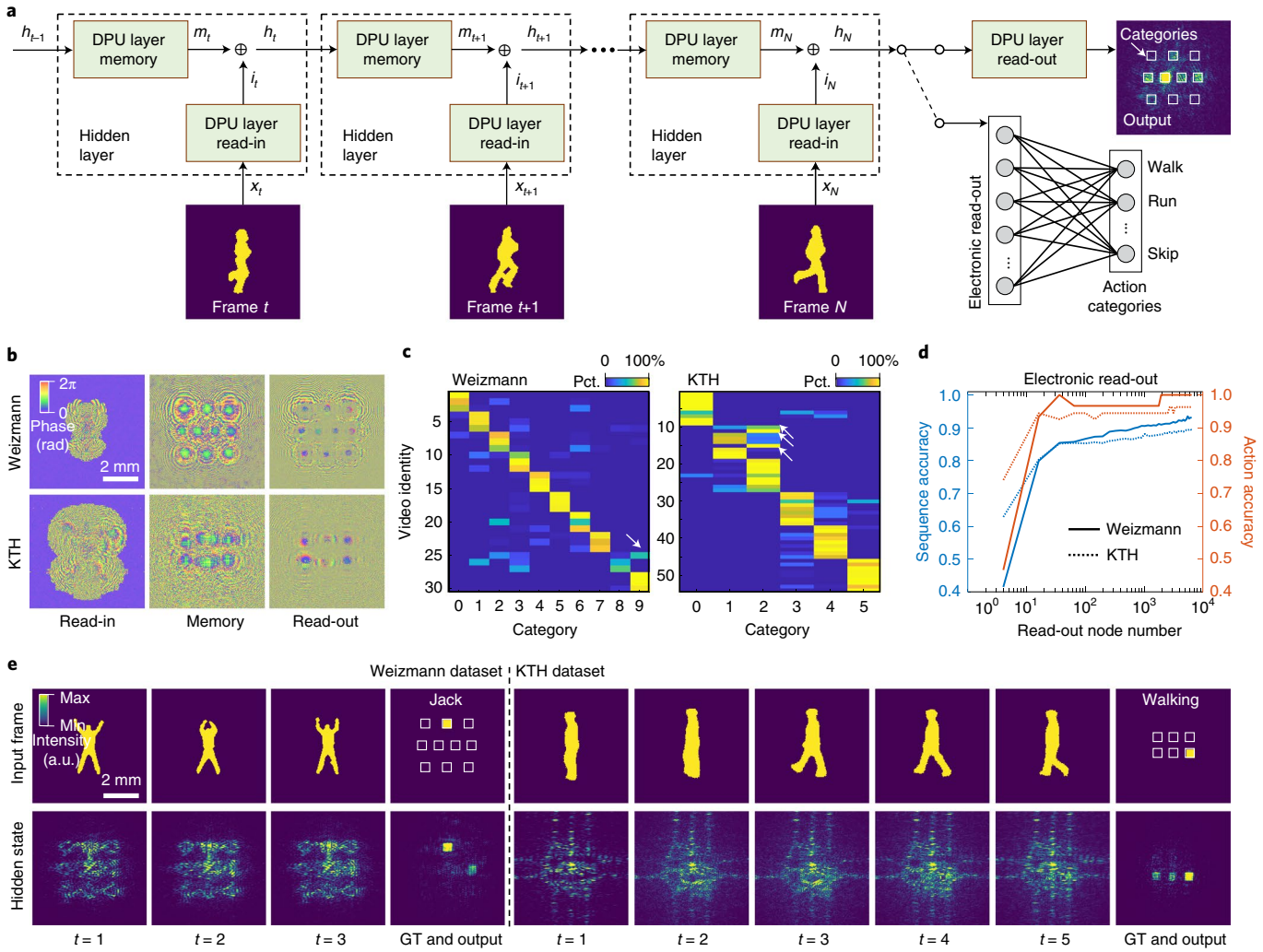

Fig. 4 | Human action recognition with the D-RNN. a, The architecture of the unfolded D-RNN. **b**, The designed modulation coefficients after adaptive training. **c**, The experimental inference results of all sub-sequences for the two databases. For the Weizmann database (left), the category labels from 0 to 9 correspond to the action types of bend, jack, jump, pjump, run, side, skip, walk, wave1 and wave2, respectively. For the KTH database (right), the category labels from 0 to 5 correspond to the action types of boxing, handclapping, handwaving, jogging, running and walking, respectively. Pct., percentage. **d**, Enhancing the performance of the D-RNN with an electronic read-out layer. **e**, The example input sub-sequences, hidden states, recognition results and ground truth (GT) category labels of two testing sub-sequences from two databases, respectively. The ground truth category of predefined regions is sequentially labelled from left to right, then top to bottom.

measurement. We adopt a DMD to provide the binary unit input and abstract the features for recognition tasks considering its properties of high optical contrast, high speed and easy calibration. We encode the 8-bit accuracy unit input into the phase of the input optical field with a phase SLM to achieve a higher input bit depth in this work. This can also be achieved by using an amplitude SLM with additional calibrations or high-speed changing of the binary pattern multiple times within a single cycle of the DPU working flow. We compare and demonstrate the advance of our architectures with respect to state-of-the-art optical neural networks in Supplementary Note 5 and Supplementary Table 3. The system can be made more compact and miniaturized by using a smaller modulation pixel pitch to reduce the propagation distance for diffractive connections and designing the application-specific integrated circuits for dataflow control and unit programmability. The scalability of the DPU's computing performance for both free-space implementation in this work and its future integration are detailed in Supplementary Note 6. Since the adopted programmable optoelectronic devices are widely used in many areas, such as in microscopic imaging[45] and free-space optical communication[46], the proposed architecture can be integrated into these applications with both real-space and Fourier-space implementations[28] as well as the spatial and spectral multiplexing techniques for the construction of more advanced ANNs.

In summary, we have experimentally demonstrated a reconfigurable optoelectronic computing processor, that is, a DPU, which can be programmed to adapt to different types of ANN for large-scale optical neural information processing with a high performance. By effectively designing an optoelectronic computing system to fuse the complementary advantages of optics and electronics, we anticipate that the proposed approach will accelerate the development of more powerful optical AI processors as critical support for modern computing and towards beginning a new era of AI.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of

author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41566-021-00796-w.

Received: 30 November 2020; Accepted: 5 March 2021;
Published online: 12 April 2021

References

1. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (eds Bartlett, P. et al.) 1097–1105 (Curran Associates, 2013).
3. Zhang, C. et al. Optimizing FPGA-based accelerator design for deep convolutional neural networks. In *Proc. 2015 ACM/SIGDA International Symposium on Field-programmable Gate Arrays* 161–170 (Association for Computing Machinery, 2015).
4. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
5. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
6. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
7. Graves, A., Mohamed, A.-r. & Hinton, G. Speech recognition with deep recurrent neural networks. In *Proc. International Conference on Acoustics, Speech and Signal Processing* 6645–6649 (IEEE, 2013).
8. Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
9. Lukoševičius, M. & Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* **3**, 127–149 (2009).
10. Waldrop, M. M. The chips are down for Moore's law. *Nature* **530**, 144–147 (2016).
11. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
12. Prucnal, P. R. & Shastri, B. J. *Neuromorphic Photonics* (CRC, 2017).
13. Zhang, Q., Yu, H., Barbiero, M., Wang, B. & Gu, M. Artificial neural networks enabled by nanophotonics. *Light Sci. Appl.* **8**, 42 (2019).
14. Ferrera, M. et al. On-chip CMOS-compatible all-optical integrator. *Nat. Commun.* **1**, 29 (2010).
15. Zhu, T. et al. Plasmonic computing of spatial differentiation. *Nat. Commun.* **8**, 15391 (2017).
16. Estakhri, N. M., Edwards, B. & Engheta, N. Inverse-designed metastructures that solve equations. *Science* **363**, 1333–1338 (2019).
17. Xu, X.-Y. et al. A scalable photonic computer solving the subset sum problem. *Sci. Adv.* **6**, eaay5853 (2020).
18. Liu, W. et al. A fully reconfigurable photonic integrated signal processor. *Nat. Photon.* **10**, 190–195 (2016).
19. Kwon, H., Sounas, D., Cordaro, A., Polman, A. & Alù, A. Nonlocal metasurfaces for optical signal processing. *Phys. Rev. Lett.* **121**, 173004 (2018).
20. Shainline, J. M., Buckley, S. M., Mirin, R. P. & Nam, S. W. Superconducting optoelectronic circuits for neuromorphic computing. *Phys. Rev. Appl.* **7**, 034013 (2016).
21. Van der Sande, G., Brunner, D. & Soriano, M. C. Advances in photonic reservoir computing. *Nanophotonics* **6**, 561–576 (2017).
22. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
23. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
24. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
25. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–447 (2017).
26. 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).
27. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
28. Tao, Y. et al. Fourier-space diffractive deep neural network. *Phys. Rev. Lett.* **123**, 023901 (2019).
29. Zhou, T. et al. In situ optical backpropagation training of diffractive optical neural networks. *Photonics Res.* **8**, 940–953 (2020).
30. Dou, H. et al. Residual D^2NN: training diffractive deep neural networks via learnable light shortcuts. *Opt. Lett.* **45**, 2688–2691 (2020).
31. Antonik, P., Marsal, N., Brunner, D. & Rontani, D. Human action recognition with a large-scale brain-inspired photonic computer. *Nat. Mach. Intell.* **1**, 530–537 (2019).
32. Rafayelyan, M., Dong, J., Tan, Y., Krzakala, F. & Gigan, S. Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction. *Phys. Rev. X* **10**, 041037 (2020).
33. Hughes, T. W., Williamson, I. A. D., Minkov, M. & Fan, S. Wave physics as an analog recurrent neural network. *Sci. Adv.* **5**, eaay6946 (2019).
34. Bueno, J. et al. Reinforcement learning in a large-scale photonic recurrent neural network. *Optica* **5**, 756–760 (2018).
35. Maktoobi, S. et al. Diffractive coupling for photonic networks: how big can we go? *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–8 (2019).
36. 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).
37. Luo, Y. et al. Design of task-specific optical systems using broadband diffractive neural networks. *Light: Sci. Appl.* **8**, 112 (2019).
38. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
39. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
40. Xiao, H., Rasul, K., & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. Preprint at https://arxiv.org/abs/1708.07747 (2017).
41. Blank, M., Gorelick, L., Shechtman, E., Irani, M. & Basri, R. Actions as space-time shapes. In *Proc. Tenth International Conference on Computer Vision* 1395–1402 (IEEE, 2005).
42. Schuldt, C., Laptev, I. & Caputo, B. Recognizing human actions: a local SVM approach. In *Proc. 17th International Conference on Pattern Recognition* 32–36 (IEEE, 2004).
43. Jhuang, H., Serre, T., Wolf, L. & Poggio, T. A biologically inspired system for action recognition. In *Proc. 11th International Conference on Computer Vision* 1–8 (IEEE, 2007).
44. Shu, N., Tang, Q. & Liu, H. A bio-inspired approach modeling spiking neural networks of visual cortex for human action recognition. In *2014 International Joint Conference on Neural Networks* 3450–3457 (IEEE, 2014).
45. Ji, N. Adaptive optical fluorescence microscopy. *Nat. Methods* **14**, 374–380 (2017).
46. Wang, J. et al. Terabit free-space data transmission employing orbital angular momentum multiplexing. *Nat. Photon.* **6**, 488–496 (2012).

Methods

Experimental system. Our experimental system adopted commercially available optoelectronic devices as the building blocks of a DPU (Supplementary Fig. 1). The energy source for optical computing is generated using a solid-state laser (CNI MRL-FN-698) at a working wavelength of 698 nm coupled with a single-mode fibre (P5-630A-PCAPC-1, Thorlabs). The coherent light wavefront is collimated using a relay lens and is reflected to the correct incident angle for a DMD (DLP9000X, Texas Instruments) using a total internal reflection prism. The DMD, mounted on a controller board (V4390, Vialux), consists of 2,560 × 1,600 micromirrors with a pitch of 7.6 μm that can modulate the incident light at a maximum speed of 14,989 Hz. The optical field from the input nodes is polarized using a linear polarizer (LPNIR100-MP2) and optically conjugated to a phase SLM (HSP1920-600-1300-HSP8, Meadowlark) with a 4*f* system comprising two relay lenses (AC508-100-A, Thorlabs). Another linear polarizer is adopted for tuning the incident light intensity. We further placed an optical iris at the Fourier plane of the 4*f* system to filter out high-order diffractions from the DMD and ensure the node-wise phase modulation of the input optical field with the SLM. The phase SLM is a reflective liquid crystal on silicon, which has a high zero-order diffraction efficiency of 88% with 1,920 × 1,152 modulation elements, each with a size of 9.2 μm and an 8-bit accuracy. It can operate at a maximum frame rate of 422.4 Hz with an update time of ~2.4 ms. To implement the diffractive connections, the zero-order diffracted optical field from the SLM is reflected by a 50:50 split ratio non-polarized beamsplitter (BS013, Thorlabs) to a sensor. We used a scientific CMOS (sCMOS; Andor Zyla 4.2) sensor to actualize the diffractive optoelectronic neurons, where the neuron functionalities are achieved with the photoelectric effect that naturally occurs at each pixel of the sensor. The high photosensitivity of the sCMOS pixels ensures a minimum requirement of the input energy. The maximum frame rate of the adopted sCMOS sensor can be customized by the predefined region of interest on the sensor, for example, 100 frames per second (f.p.s.) and 1,627 f.p.s. under pixel number settings of 2,048 × 2,048 and 128 × 128, respectively, where each pixel has a size of 6.5 μm and a 16-bit accuracy.

The frame updates of the DMD, SLM and sCMOS sensor were synchronized using software with external triggers (Supplementary Fig. 5) that control the massive dataflow of the unit for high-parallelism processing. To calibrate the system, we first aligned the device pixel and obtained the effective region correspondence between the DMD and SLM, as well as that to the sCMOS sensor, by projecting the chequerboard patterns and aligning them with optical translation stages. The designed layer distance between the SLM and the sCMOS sensor was accurately set by projecting arrays of Fresnel lens phase patterns on the SLM with focal lengths the same as the layer distance. The positions of the generated focus points on the sensor were used for aligning the two planes, that is, the SLM plane and the sensor plane. In addition, the SLM was calibrated using a look-up table-calibration method to generate a linear 0–2π phase response at a wavelength of 698 nm.

D-RNN model formulation and evaluation. For the D-RNN hidden layer at the time step of t, the hidden state h_t is a function of the hidden state h_{t-1} at the time step of $t-1$ and of the input sequence x_t at the time step of t. We adopt an affine combination to fuse the states from these two sources, that is, $h_t = \lambda f_1(h_{t-1}) + (1-\lambda) f_2(x_t)$, where $m_t = f_1(h_{t-1})$ denotes the memory state mapping from h_{t-1}; $i_t = f_2(x_t)$ denotes the input state mapping from x_t; and $\lambda \in (0,1)$ is the fusing coefficient that controls the strength of the memory state with respect to the input state. The complexity of the mapping functions $f_1(\cdot)$ and $f_2(\cdot)$ can be increased by using multiple DPU layers, forming a spatial deep hidden layer structure in addition to the temporal deep architecture of the D-RNN, for example, as shown in Fig. 1e. The recurrent connection of the hidden layer at different time steps allows the D-RNN to process a variable sequence length of inputs (Fig. 4a). Although a longer network sequence length (larger N) can incorporate more frames for the recognition decision, this causes difficulties for the network in training as well as the forgetting of long-term memory, that is, the vanishing of frame information at a time step that is far from the current time step. Therefore, for each of the video sequences in the database with a length of M, we set $N \ll M$ and divided the sequence into numbers of sub-sequences with the same length as N, with which the D-RNN was trained and tested. We quantitatively evaluated the model accuracy with two metrics, that is, the sequence accuracy and the action accuracy. The sequence accuracy was obtained by statistically summarizing the inference results of all sub-sequences in the test set. The action accuracy was calculated based on the predicted category of each video sequence in the test set and was derived by applying the winner-takes-all strategy[31] (the action category with the most votes) on the testing results of all sub-sequences in the video sequence.

Neural network settings. We built neural network architectures and conducted parameter optimization based on the forward model of the DPU (see Supplementary Note 1) and its optoelectronic implementation, where the hyperparameters were determined during in silico pre-training. To achieve a high recognition accuracy and for the sake of experimental robustness, we optimized and set the number of input nodes and optoelectronic neurons of the DPU as 560 × 560 for the D²NN and D-NIN-1 and 700 × 700 for the D-RNN (Supplementary Fig. 6), each with a size of 9.2 μm, during both the training and testing. Therefore, the network size at each layer is 5.152 × 5.152 mm for the D²NN and D-NIN-1 and 6.44 × 6.44 mm for the D-RNN. The total number of input nodes and optoelectronic neurons determined by the number of DPU layers in the architecture is 0.94 million, 2.20 million and 1.47 million for the D²NN, D-NIN-1 and D-RNN, respectively. We further included zero padding at the network periphery to guarantee the boundary condition of free-space diffraction during the numerical modelling using the angular spectrum method[34]. The layer distance for optical diffraction was optimized and set to 20 cm, which enabled a fully connected structure according to the Huygens–Fresnel principle for calculating the maximum diffractive angle of an input node[27]. To facilitate the fine tuning of the phase value during adaptive training and utilize the full-range phase modulation of the SLM, a sigmoid function was applied to constrain the phase modulation values of the input nodes to ~0–2π. The network hyperparameter α (see Supplementary Note 1) was initialized at a value of ten and gradually increased to 100 during the training, considering both the network trainability and the model matching quality. The training target was to maximize the output optical signal on the corresponding square region for each input object, and the loss function was defined (see Supplementary Note 7). The number of output regions was set to be the same as the number of categories, that is, ten regions for the MNIST, Fashion-MNIST and Weizmann databases, and six regions for the KTH database, each with a width of 0.46 mm. For all three architectures, the learning rate was set to 0.01, with a training batch size of ten for the D²NN and D-NIN-1 and 20 for the D-RNN.

After the ablation analysis, the optimal number of layers and the number of feature maps of the constructed feedforward DNNs, that is, the D²NN and D-NIN-1, for classifying the MNIST dataset were both found to be three (Supplementary Fig. 7a,b). The photoelectric non-linearity is sufficient to serve as the neuron activation function for the DPU-constructed neural networks (Supplementary Fig. 7c). Considering the recognition speed of our system and the complexity of our tasks, the mapping functions of the input sequence (x_t) to the input state (i_t) and the hidden state (h_{t-1}) to the memory state (m_t), at the time step of t, are both implemented with a single DPU layer (Fig. 4a) that operates the system at a read-in speed of ~63 f.p.s. (see Supplementary Note 8). By further considering the task complexity and training cost, the sequence lengths of the D-RNN for human action recognition using the Weizmann and KTH databases were optimized to 3 and 5, respectively, corresponding to an optimal fusing coefficient of 0.2 (Supplementary Fig. 8).

Adaptive training method. The entire training process for the experimental network design is illustrated in Supplementary Fig. 2, which includes the pre-training step followed by the iterative fine-tuning steps to adapt the model to the practical system. To implement, we first adopt an in silico electronic pre-training process[27] to simulate the network model and learn its parameters. In silico training has a lower training complexity than in situ training[25,26] by taking advantage of the existing high-performance physical models and system parameters. The pre-trained model is then transferred to the optoelectronic system by deploying the network structure and programming the SLM for individual DPU layers to write the network parameters. As the model transfer errors distort the wavefront connections of neurons in each layer, we derive the fundamental principle from adaptive optics[45] to sequentially compensate for the wavefront distortion in each layer and alleviate the accumulation of error. For correction of the current layer, we experimentally record its DPU outputs in situ by using the samples from the training set and adopt them as the inputs of the next layer for in silico re-training of the following diffractive layers. The error of the last layer is corrected by simply multiplying the energy distribution of the output categories by the decision coefficients, for example, ten coefficients for ten categories, and the optimization uses the same training schematic. The adaptive training process fine tunes the parameters of each diffractive layer to accommodate the model transfer errors and effectively recovers the high inference accuracy. As an example, two-stage adaptive training of the constructed D²NN is detailed in Supplementary Note 3.

Neural network training. We adopt a stochastic gradient descent algorithm, that is, the adaptive moment estimation (Adam)[47] optimizer, for optimizing the network parameters in silico and minimizing the loss function during both pre-training and adaptive training. Based on the pre-trained model, our adaptive training method uses the samples from the training set for generating the system outputs in situ at each layer, which are then applied to fine tune the network parameters in silico under the same network settings as those used in the pre-training. The in silico training and testing were implemented on a modern machine-learning platform using Python version 3.6.7 with the TensorFlow framework version 1.12.0 (Google) running on a desktop computer (Nvidia TITAN XP GPU, Intel Xeon Gold 6126 CPU with 64 cores, 128 GB of RAM and the Microsoft Windows 10 operating system). With the above-outlined network design and recognition tasks, the pre-training stage took 15 epochs to obtain the converged model, corresponding to the training times of ~3.8, ~8.4, ~1.1 and ~1.7 h for the constructed D²NN, D-NIN-1, D-RNN (Weizmann) and D-RNN (KTH), respectively. For the two-stage adaptive training of the D²NN, as each stage of adaptive fine tuning converges after four epochs using a full training set and eight and ten epochs using a 20% and 2%

mini-training set (Fig. 2), the corresponding total adaptive training times for the D²NN, for instance, were ~1.3 h, ~24 min and ~3 min, respectively.

System computing performance. The computing performance of the DPU-constructed DNN architectures, including the computing speed and energy efficiency, is benchmarked in Supplementary Tables 1 and 2, with the calculations detailed in Supplementary Note 2. The computing speed, measuring the total number of operations per second for the architectures, is 114.1 tera-operations per second (TOPS) for the D²NN and D-NIN-1 and 240.1 TOPS for the D-RNN. We adopt the direct energy efficiency to evaluate the potential upper limit of the operations per watt-second based on the energy consumption that is used directly for computations. With the ratio between optical and electronic computational operations at the order of 10^6, the direct energy efficiency is larger than 1.0 peta-operations per second per watt (POPS/W) for all architectures. We use the system energy efficiency for faithfully evaluating the computational energy efficiency of the prototype system by measuring the average power consumption of all optoelectronic devices, that is, the desktop computer, sCMOS sensor, DMD, SLM and solid-state laser (Supplementary Table 4). With the total power consumption of 159.4 W for the D²NN and D-NIN-1, and 152.1 W for the D-RNN, the corresponding system energy efficiency is 0.717 and 1.578 tera-operations per second per watt (TOPS/W), respectively. Therefore, our prototype system has achieved a greater than eight times higher computing speed and more than an order of magnitude better system energy efficiency than the Tesla V100 GPU[48]. Further, we evaluated the inference speed of our prototype system and its comparisons with the adopted electronic desktop computer, that is, the mainstream CPU+GPU processor, in Supplementary Note 8.

Data preprocessing. Both the MNIST handwritten digit database[39] and the Fashion-MNIST database[40] are composed of 60,000 training images and 10,000 testing images, each with a resolution of 28×28 belonging to one of ten categories. The MNIST database contains digit categories from zero to nine, while the Fashion-MNIST database contains product types of t-shirt, trousers, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot. To match the input size of the D²NN and D-NIN-1, all images of the MNIST database were binarized and up-sampled 20 times with nearest-neighbour interpolation to 560×560, and all images of the Fashion-MNIST database were up-sampled 15 times with bilinear interpolation and padded to 560×560.

The Weizmann human action database[41] contains 90 video sequences including nine subjects, with each subject performing ten natural actions, that is, bend, jumping-jack (jack), jump-forward-on-two-legs (jump), jump-in-place-on-two-legs (pjump), run, gallop-sideways (side), skip, walk, wave-one-hand (wave1) and wave-two-hands (wave2). Each video sequence has ~30–100 frames captured at a frame rate of 50 f.p.s. with a resolution of 180×144, which was cropped to 144×144 and up-sampled four times with zero padding to 700×700 to match the input size of the D-RNN. All frames are accompanied by silhouette binary masks of the subject that can represent different categories of human action and were used as the input to our system. We adopted the actions from six subjects (60 video sequences) as the training set, with the rest of the actions, that is, three subjects (30 video sequences), as the test set. Since the D-RNN sequence length was set to three, each video was divided into numbers of sub-sequences by sequentially extracting three frames for each sub-sequence with a frame interval of two. After the data augmentation (horizontally flip), the total numbers of sub-sequences in the training and test sets of the Weizmann database were 6,000 and 3,000, respectively.

The other benchmark database for human action recognition is the KTH database[42], which has 600 video sequences containing six different types of natural actions, that is, boxing, handclapping, handwaving, jogging, running and walking, performed by 25 subjects in four different scenes. Each of the video sequences was captured at a rate of 25 f.p.s. that includes ~350–600 frames with a spatial resolution of 160×120, which was also up-sampled four times with zero padding to 700×700 to match the input size of the D-RNN. The recognition performance of the D-RNN was evaluated and compared with that of existing approaches[31,43,44] by using the first scene (outdoors scenario 1, 150 video sequences) and 16:9 data splitting. To adapt the database to our system, the silhouette binary mask of the subject in each frame was extracted by using a pre-trained DNN model[49] to perform semantic segmentation. For the video sequences of the 12th and 13th subjects, we further adopted a user interface (MATLAB Toolstrip, Mathworks) to compensate for the failed segmentation regions due to high similarities between the clothes of the subject and the background scene. The subject was aligned in each video and the frames without subjects were discarded during preprocessing. Similar to that for the Weizmann database, with a D-RNN sequence length of five and data augmentation, the training set had 16 subjects, including 7,680 sub-sequences, and the test set had nine subjects, including 2,160 sub-sequences.

Integration with electronic output layers. The DPU-constructed DNNs are highly compatible with electronic neural networks. We demonstrated an improved model and experimental performance of the D-NIN-1 and the D-RNN by integrating a low-complexity electronic fully connected output layer. The boosted D-NIN-1 model, that is, D-NIN-1++, was obtained by average pooling the network output with a factor and fully connecting to ten output category nodes for the MNIST database. We adopted an end-to-end training strategy to learn both the diffractive and electronic weights of the D-NIN-1++, where the softmax cross-entropy[36] loss was used as a loss function during the training. With an optimized downsampling factor of 40, the size of the fully connected electronic output layer is 196×10, reaching a classification accuracy of 99.0% on the MNIST database (Supplementary Fig. 9).

Inspired by the memory read-out modality of reservoir computing[9,31,50], the D-RNN++ transferred the trained hidden layer of the D-RNN and replaced the DPU read-out layer with a linear electronic fully connected read-out layer. The fully connected weight between the read-out nodes and category nodes was efficiently trained with a ridge regression algorithm[31,50], where the ridge parameter λ determines the strength of a ridge penalty. The redundancy of memories in the hidden state enables us to reduce the complexity of the electronic read-out layer, that is, reducing the number of read-out nodes, by average pooling the last hidden state. For each of the databases, the ridge parameter was optimized by grid searching from 10^{-4} to 10^5 with a step size of an order of magnitude, and the minimal number of read-out nodes was determined for extracting the category information from the memory of input sequences at low complexity. The experimental results demonstrated that 2,500 and 2,704 read-out nodes were sufficient for robust recognition, and the action accuracy was improved from 96.7 to 100.0% for the Weizmann database and from 94.4 to 96.3% for the KTH database under the ridge parameters of one and ten, respectively (Supplementary Fig. 10).

Data availability

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

Code availability

All relevant code is available from the corresponding author upon reasonable request.

References

47. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
48. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
49. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018* (eds Ferrari, V. et al.) 833–851 (Springer, 2018).
50. Larger, L. et al. High-speed photonic reservoir computing using a time-delay-based architecture: million words per second classification. *Phys. Rev. X* **7**, 011015 (2017).

Acknowledgements

We thank T. Zhu and T. Yan for assistance with the software for the prototype system. This work is supported by the Beijing Municipal Science and Technology Commission (No. Z181100003118014), the National Key Research and Development Program of China (No. 2020AAA0130000), the National Natural Science Foundation of China (No. 62088102 and No. 61860206003) and the Tsinghua University Initiative Scientific Research Program.

Author contributions

Q.D., X.L. and L.F. initiated and supervised the project. X.L. and T.Z. conceived the research and method. X.L. designed the simulations and experiments. T.Z. conducted the experiments. T.Z. and Y.C. performed the simulations and processed the data. T.Z., J.W., Y.C. and H.X. built the experimental system. X.L., T.Z., J.W., H.W., H.X., J.F., Y.L. and L.F. analysed the results. X.L., T.Z., L.F. and Q.D. prepared the manuscript with input from all authors. All authors discussed the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41566-021-00796-w.

Correspondence and requests for materials should be addressed to X.L., L.F. or Q.D.

Peer review information *Nature Photonics* thanks Nathan Youngblood and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.